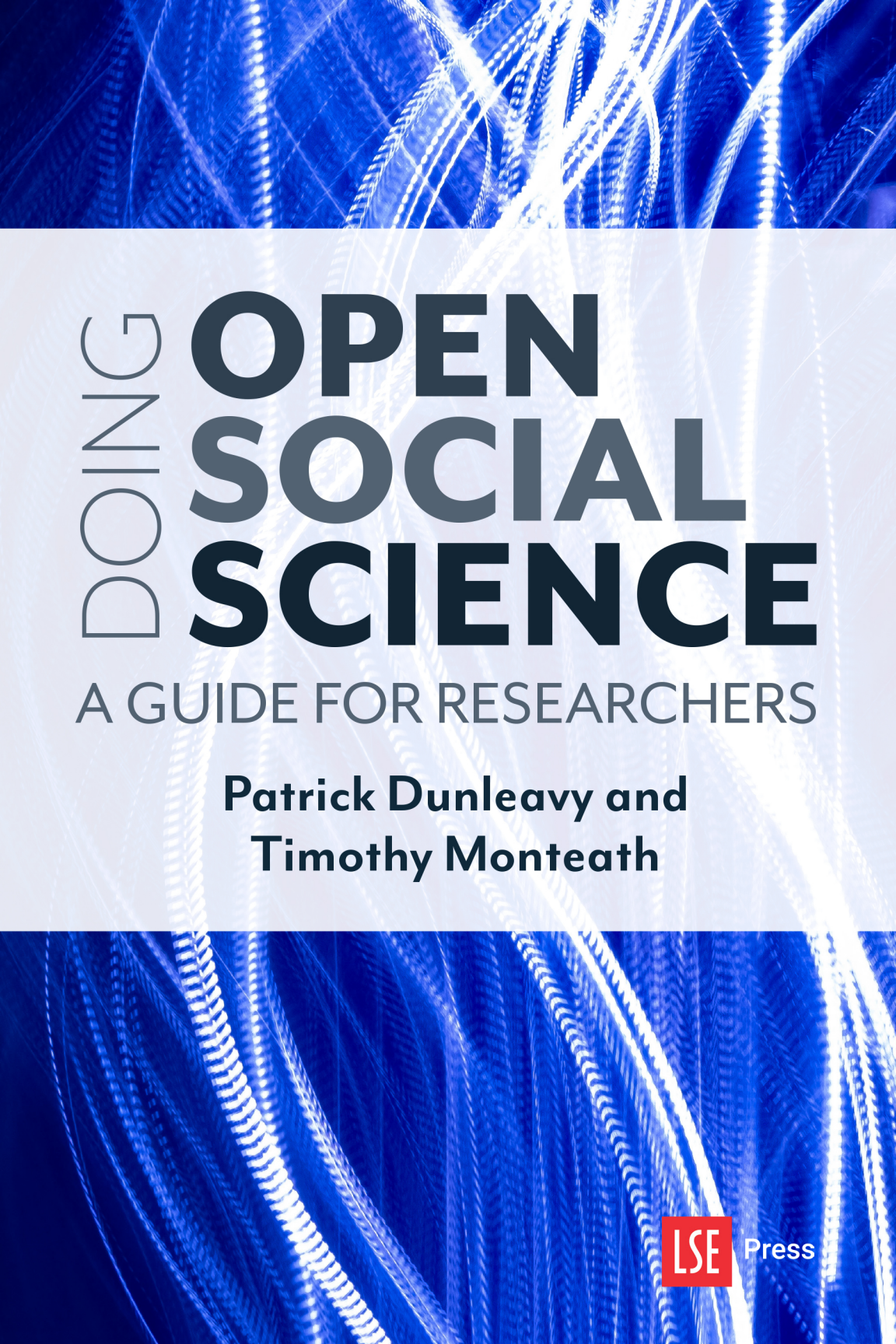

DOING
OPEN
SOCIAL
SCIENCE
A GUIDE FOR RESEARCHERS
Patrick Dunleavy and
Timothy Monteath
LSE Press

Doing Open Social Science

A Guide for Researchers

Patrick Dunleavy and Timothy Monteath
with
Hannah Boroudjou
Denisa Kostovicova

LSE Press

CIVICA
THE EUROPEAN UNIVERSITY OF SOCIAL SCIENCES

Published by
LSE Press
10 Portugal Street
London WC2A 2HD
press.lse.ac.uk

Text © 2025 Patrick Dunleavy and Timothy Monteath, except Chapter 7 © Hannah Boroudjou and Chapter 13 © Patrick Dunleavy, Timothy Monteath and Denisa Kostovicova.
Images and other exhibits © listed individually in source notes
First published 2026
Cover design by Diana Jarvis.
Print and digital versions typeset by Newgen Publishing UK.

ISBN (Paperback): 978-1-911712-59-6
ISBN (PDF): 978-1-911712-60-2
ISBN (EPUB): 978-1-911712-61-9
ISBN (Mobi): 978-1-911712-62-6

DOI: https://doi.org/10.31389/lsepress.dos

This work is licensed under the Creative Commons Attribution - NonCommercial 4.0 International (CC BY-NC-ND 4.0) License (unless stated otherwise within the content of the work). To view a copy of this license, visit https://creativecommons.org/licenses/by-nc-nd/4.0/ or send a letter to Creative Commons, 444 Castro Street, Suite 900, Mountain View, California, 94041, USA. This license allows for copying any part of the work for personal use, providing author attribution is clearly stated. You may not use the material for commercial purposes. If you remix, transform, or build upon the material, you may not distribute the modified material.

The full text of this book has been peer-reviewed to ensure high academic standards. For our full publishing ethics policies, see https://press.lse.ac.uk.

Suggested citation:
Dunleavy, Patrick and Monteath, Timothy (2026) *Doing Open Social Science: A Guide for Researchers,* London: LSE Press. https://doi.org/10.31389/lsepress.dos.
Licence: CC BY-NC-ND 4.0

To read the free, open access version of this book online, visit https://doi.org/10.31389/lsepress.dos or scan this QR code with your mobile device:

Contents

List of figures v
About the authors ix
Acknowledgements xi
Preface xv
Our references and citing this book xix
Glossary xxi

1. Introduction – using this book in research 1

PART I: OPENING UP THE SOCIAL SCIENCES 7

2. The 'open social science' agenda 9
3. Open science in STEMM disciplines – and lessons for open social science 39
4. Pre-registration for confirmatory and exploratory research 79

PART II: PRACTICAL APPROACHES TO MORE OPEN RESEARCH 105

5. Documenting research as you go – principles and tools 107
6. Reusing and mashing other people's data 129
7. Writing a data management plan
   Hannah Boroudjou 149

8. Open data for publications – replication archives and research methods annexes 177

PART III: ACHIEVING MORE OPEN QUALITATIVE RESEARCH 191

9. Interview-based research 195

10. Case studies and fieldwork 225

11. Systematic documentation analysis 255

12. Archival data – problems and early solutions 285

13. Doing 'citizen social science'
Denisa Kostovicova, Patrick Dunleavy and Timothy Monteath 299

PART IV: SHIFTING RESEARCH CULTURES 321

14. Presenting results well to reach a wide audience 323

15. Publishing and open access 343

16. Conclusions – achieving open social science 363

Index 389

List of figures

1.1 The main themes covered in this book (green and yellow rows) and less covered topics (white and grey rows) 4

2.1 The social sciences discipline group, mixed subjects and the social science 'zone of influence' – plus neighbouring discipline groups 11

2.2 Traditional false polarities between the 'natural' or 'physical' sciences and the social sciences 14

2.3 The place of the social sciences within the study of 'human-dominated systems' and 'human-influenced systems' 15

2.4 The years in which U.S. social scientists reported to Christenson et al. that they first adopted three kinds of open science practices 19

2.5 The varying use of 'open' actions across three types of researchers in the Christensen research (2017) 20

2.6 How the Turing Way project defines reproducible research 27

2.7 Pictograph of Turing Way's four types of 'open science' research 27

3.1 Ten key problems with conventional (non-open) science processes alleged by open science critics 46

3.2 Trends in the citing of data or methods URLs across computer science, physics and mathematics, 2012–2022 50

3.3 The contrasting approaches of two FOSS streams and two proprietary modes of developing software 54

3.4 How open source supports STEMM research 55

3.5 One of the first ever pictures of a black hole published in 2023 58

4.1 Some difficulties in specifying a precisely testable and theoretically coherent null hypothesis – the example of 'Duverger's law' of electoral competition 82

4.2 Two types of error in hypothesis testing 87

4.3 The distribution of z-statistics for different sub-samples of economics papers, by type of data 89

4.4 Comparing confirmatory research, exploratory data analysis and solely qualitative research 95

4.5 The relevance of pre-analysis plans and the role of pre-registration in CDA and EDA research 96

5.1 An example of how Timothy organizes his files 115

5.2 An example of how Patrick organizes his files 116

5.3 Poorly commented code may not elucidate what is being done – but makes it harder to understand 121

5.4 Well-commented-on code explains what is being done, why it is being done and the choices that were made at the time, matching the author's existing skill level 121

5.5 An example of Timothy Monteath's Zotero Library, sorted by date 124

6.1 Five key dimensions against which to assess reusing a dataset 135

6.2 Visualizing Anscombe's quartet – four different XY data relationships produce many of the same simple statistics 139

6.3 An example of a regular expression 142

6.4 The eight main systems of areas used in UK statistical geographies by the Office of National Statistics (ONS) 145

7A Data repositories to consider 172

8.1 The 16 files listed in the Richman and Roberts replication archive for their paper, 'Assessing spurious correlations in big search data' 179

8.2 The replication archive for Hoey, Peeters and Principe's paper, 'The transfer system in European football' 181

9.1 Main factors shaping interview outcomes 196

9.2 The three main types of research interviews in the social sciences 200

9.3 The major gains and limits of different recording modes 209

10.1 The three best known forms of case study 228

10.2 Selecting cases against a known population distribution – matched cases and limiting cases 232

10.3 Example – situating three case study countries against a wider set of comparable advanced liberal democracies 235

10.4 Three case approaches (4, 5 and 6) with quantitative elements (compared to classic forms) 236

10.5 An ambitious 'nested analysis' model, after Lieberman (2005) 237

11.1 Fifteen key problems in reading or qualitatively appreciating any documentation source 258

11.2 The extent to which an organization's claims to be undertaking well-regarded activities match with the profile visible in its public documentation 276

11.3 Example – results from a 2004 web census of UK government agencies' administrative forms scored on 40 quality indicators of care and attention for users 278

13.1 Older forms of citizen social science and newer forms falling within our broad definition 301

14.1 Summary of Midway's *Principles of Effective Data Visualization* 334

14.2 The geography of talk by telephone in Great Britain, in 2010 336

14.3 Hehman and Xie's examples of an 'Improved scatterplot visualizing the relationship between implicit and explicit anti-Black bias, including a 95% confidence band of the slope, with histograms of the variable on each axis in the opposite margins' 337

15.1 Users of academic work who may not be able to access most published journal articles or research books 345

15.2 The main types of Creative Commons (CC) licences 348

15.3 Pro-actively building knowledge of open access research publications 354

16.1 Fifteen key barriers for researchers in adopting open social science, and the main countervailing drivers for continuing innovation 365

16.2 Our estimates of the importance of different factors for academic career success, in article-dominated and mixed social science disciplines 370

16.3 Contrasting workflow processes used in manual and AI-driven systematic reviews in a 2025 study 375

16.4 An increasing proportion of citations were to non-elite journal articles from 1995 to 2013, across all disciplines 376

16.5 Older articles (more than five years old) were also cited more from 1995 to 2013 in most disciplines (except in engineering or chemistry) 377

16.6 The mean size of STEMM and social science research teams has been growing steadily, and work from larger teams is more cited than average 381

Authors

Patrick Dunleavy is Emeritus Professor of Political Science and Public Policy at the London School of Economics and Political Science (LSE). He worked in the Department of Government at LSE from 1979 to 2020, and before that at the Open University and Nuffield College, Oxford. He is also Emeritus Professor of Government at the University of Canberra, where he was Centenary Professor 2015 to 2021. A Fellow of the British Academy and the Academy of Social Sciences, he also served as Founding Editor in Chief at LSE Press from 2020 to 2023. He was Director of the UK Democratic Audit from 2013 to 2020 and co-led the Australian Democratic Audit from 2020 to 2024. His recent books include *Australia's Evolving Democracy: A New Democratic Audit* (open access from LSE Press, 2024, co-edited with Mark Evans and John Phillimore); *Maximizing the Impacts of Academic Research* (2021, Palgrave, now Bloomsbury Press, co-authored with Jane Tinkler); and *The UK's Changing Democracy: A New Democratic Audit* (open access from LSE Press, 2018, co-edited with Alice Park and Rosamund Taylor).

p.dunleavy@lse.ac.uk

Timothy Monteath has been Assistant Professor in Data Visualization in the Centre for Interdisciplinary Methods at the University of Warwick since 2023. Previously, he was Associate Lecturer in Human Geography and Data Visualization at University College, London; CIVICA Open Social Science Researcher at the LSE; and a post-doc Research Associate in Finance and Geography at the University of Oxford. His PhD in Sociology from the LSE focused on the information infrastructure of land registration in England. He is a co-author of *Atlas of Finance: Mapping the Global Story of Money* (2024, Yale University Press).

Timothy.monteath@warwick.ac.uk

Hannah Boroudjou is the Research Data Librarian at the LSE Library (the British Library of Political and Economic Science). She joined LSE in 2018 as part of the subscriptions team and previously worked in libraries at City University, the University of London and in public libraries.

datalibrary@lse.ac.uk

Denisa Kostovicova is Professor in Global Politics and Director of LSE Research on South Eastern Europe in the European Institute at the London School of Economics and Political Science (LSE). Previously she was a research fellow at Oxford University and at LSE. Her most recent books are *Reconciliation by Stealth: How People Talk about War Crimes* (2023, Cornell University Press); *Rethinking Reconciliation and Transitional Justice After Conflict* (2018, Routledge; co-edited with James Hughes); and *Persistent State Weakness in the Global Age* (2016, Routledge; co-edited with Vesna Bojicic-Dzelilovic). Her numerous research articles and chapters cover transitional justice, peace-building in former conflict zones, inter- and intra-ethnic dialogues on wartime violence, politics in the Balkans and south-east Europe, and the European Union's peace-building there.

d.kostovicova@lse.ac.uk

Acknowledgements

This work draws heavily upon many discussions and work done in the Civica I project (2021–2023) funded by the European Union's Horizon research programme. This large-scale collective endeavour involved eight of Europe's leading social science universities co-operating to explore a wide range of issues around developing open science in our disciplines. An extensive series of working parties, seminars and graduate student training sessions brought together PhDers, researchers and allied professional staffs in 2022–2023 from

- the London School of Economics and Political Science (and the British Library of Political and Economic Science);
- the European University Institute, Florence;
- Bocconi University, Milan;
- Sciences Po, Paris;
- the Hertie School, Berlin;
- the Central European University, Vienna;
- the Stockholm School of Economics; and
- SNSPA (the National University of Political Studies and Public Administration) in Bucharest.

After the end of our project, the second phase of Civica enlarged to also include Central European University and SGH Warsaw School of Economics.

Our first thanks then go to the hundreds of participants in the Civica I discussions and forums, who hugely extended our knowledge of the contrasting approaches of social scientists across European countries and across many different disciplines. We are especially grateful to those who attended the 11 online seminar sessions that formed key foundations for most of the book chapters below.

The lead coordinator of the Civica project's Workpackage 4 on 'Developing an Open Science Action Plan' was Helen Porter at LSE Library (now working in the non-profit sector). Our book was one of the leading LSE contributions to this four-pronged workstream, albeit a slightly semi-detached and autonomous one. We gained many insights from working closely with Helen, and from meetings of the Workpackage 4 committee that she chaired. Discussions with Helen and with Beth Clarke (Deputy Librarian, LSE Library) also greatly helped the project. Their generosity with their time and expertise, and their

encouragement for the authors throughout, puts us heavily in their debt. We are also grateful to Nicola Scally (LSE Librarian until 2023) for her unstinting help and wise advice. In addition, Frances Pinter (CEO of Central European University Press) was especially helpful in the early stages of the project, and specifically contributed many ideas to joint sessions that helped shape Chapter 15 on open access publishing.

We are additionally grateful to staff across all the other Civica universities who guided us in the maze of European Union (EU) project reporting requirements and in making contacts with Civica personnel and institutions. Key people at LSE were Susana Mourato (the External Pro Director at LSE), and Aygen Kurt-Dickson (LSE administration) who helped to get the 'Open Social Science Handbook' sub-project established and funded and kept us abreast of developments. Additionally, we thank Sophie Forcadell and Sarah Thompson at Sciences Po.

The incomparable Kundai Sithole (LSE Library, 2021–2023) timetabled and publicized all our Civica seminars for us, handled audience liaison before and after each session, and then helped create the lasting archive on Zenodo.eu for them that lies behind each chapter of the book (detailed in the first Note for most chapters below). The materials there include PowerPoints, audio files of sessions and YouTube videos and are all open access. Kundai was always tolerant of our often odd or rushed ways of getting stuff done – she never once complained and always delivered fully on her promises. We are deeply grateful.

Turning to the key academic influences on the book, our foremost debt is to our two co-authors who contributed wonderful chapters lying well beyond our own expertise. Hannah Boroudjou (LSE Library) wrote Chapter 7 on constructing a Data Management Plan, drawing on her work with many research teams of all types and sizes. Professor Denisa Kostovicova (LSE, European Institute) drew on her unrivalled record in undertaking citizen social science research to give the vital 'seed' seminar and start materials for Chapter 12, and co-authored the final text with us. In addition, Dr Anton Howes of the Royal Society of Arts (and author of the brilliant Age of Innovation blog) helped run our seminar for the Archives chapter, contributing experiences and ideas that he generously allowed us to freely reuse in Chapter 12. Professor Lou Saffra also ran an influential online seminar on pre-registration at Sciences Po and sent us her session script and PowerPoints – which greatly influenced how Chapter 4 discusses this topic. We are most grateful to all these colleagues for their generous help and participation.

For their generous permission to reuse their original materials in figures, we are deeply indebted to Cao Hancheng, Jesse Dodge, Kyle Lo, Daniel McFarland and Lucy Lu Wang (see Figure 3.2); Geoffrey Bower and the Event Horizon Telescope (EHT) Collaboration (see Figure 3.5); Abel Brodeur, Nikolai Cook and Anthony Heyes (see Figure 4.3); the UK National Audit Office

(see Figure 11.3); Stephen Midway (see Figure 14.1); and Santo Fortunato, Carl T Bergstrom, Katy Börner, James A. Evans, et al. (see Figure 16.6).

We are additionally grateful to the following for reusing open access data or materials – Garret Christensen, Zenan Wang, Elizabeth Levy Paluck, Nicholas Swanson, David Birke, Edward Miguel and Rebecca Littman for data (see Figures 2.4 and 2.5); Scriberia and Turing Way for an open access (OA) infographic (see Figure 2.7); Wikipedia user 'Schutz' for a picture (see Figure 6.2); the Organization for Economic Co-operation and Development (OECD) for data (see Figure 10.3); and Christian Cao, Rohit Arora, Paul Cento, Katherine Manta, Elina Farahani, et al. for an infographic reused (see Figure 16.3). We also thank Evan S. Lieberman for his agreement to redraw one of his diagrams (see Figure 10.5).

We are grateful to the following publishers who have kindly provided permission for figures to be reproduced. Figures 6.2 and 14.3 are reproduced from Hehman and Xie (2021), published by SAGE Publications Ltd. Figure 10.5 is adapted from Lieberman (2005), published by Cambridge University Press. Figure 16.6 is reproduced from Fortunato, Bergstrom, Börner et al. (2018), published by the American Association for the Advancement of Science.

In addition, the book has been shaped by many different insightful conversations and productive collaborations with academic colleagues and friends over longer time periods. Patrick Dunleavy records his deep gratitude to Simon Bastow (LSE) and Jane Tinkler (a long-time LSE colleague, but now at Nine Dots) and thanks many different talented staff from LSE Public Policy Group up to 2015. In addition, he thanks Helen Margetts (Oxford), Paul Kelly (LSE) and Daniell Fanelli (formerly LSE, now Heriot-Watt University) and staff at Canberra University from 2015 to 2020, especially David Marsh, Mick Chisnall, Selen Ercan and John S. Dryzek. We thank many other UK and Australian academics for their advice and assistance on countless issues and approaches. All of them helped shape some of the advice given here.

Timothy Monteath thanks Danielle Watts (LSE) for her always perceptive and insightful takes on OSS and for the many, many productive and generative conversations that have shaped his contributions to this book. He is deeply grateful for all he learnt from Dariusz Wójcik (Oxford and NUS) and the rest of the CityNet team (Oxford and now many other universities) before starting this book. In addition, he thanks Daniel McArthur (York) for many early conversations about pre-registration, amongst many other topics, and all his colleagues at the Centre for Interdisciplinary Methodologies (Warwick) from whom he has learnt so much, and still has so much more to learn.

At later stages of the book's development we are grateful to the LSE Press's four anonymous reviewers, and to an especially helpful final draft reviewer. We thank Sarah Worthington (LSE Law), Philippa Grand of LSE Press and Niamh Tumelty (Director of LSE Library) for valuable comments and advice. We thank LSE Press staff for their assistance in realizing the project.

Finally, Patrick Dunleavy thanks Sheila Dunleavy, without whose encouragement, ideas, advice, detailed comments and amazing forbearance the book would have been so much poorer. Timothy Monteath thanks Danielle Watts again for all her continued support and encouragement throughout the writing and editing of this book.

Preface

This book was undertaken as part of an EU Horizon-funded project called Civica I, which brought together LSE with seven other European social science universities in the period 2021–2023 – the European University Institute, Florence; Bocconi University, Milan; Sciences Po, Paris; the Hertie School, Berlin; Central European University, Vienna; the Stockholm School of Economics, Stockholm; and SNSPA, the National University of Political Studies and Public Administration, Bucharest. The Civica II project followed on in 2023 and expanded further to also include IE University, Madrid; and SGH Warsaw School of Economics, Warsaw (Civica 2025). Open Science was (and remains) a Civica research priority, and all partners sought to improve openness in the social sciences in a range of ways. (See also the Acknowledgements section).

Both Civica and this guidebook for researchers started from the promise, but also the novelty and the very uneven development, of the 'open science' (OS) agenda across the social sciences discipline group in Europe (and more widely). Workstream 4 of the project (guided by Helen Porter of LSE Library and Patrick Dunleavy) undertook many co-operative activities with Civica I faculty and allied research professionals. It focused on helping new and established researchers to understand what OS entails and how universities can enhance the transparency, reproducibility and openness in the work that their academics produce. As part of LSE's contribution, this book stood somewhat at a distance. It was envisaged as a freestanding handbook or advice text – one that could reach out to audiences both within the Civica-participating institutions themselves and beyond to the wider European and global social science communities. LSE Press's involvement should also mean that the book can endure well beyond the lifetime of the project itself. We were lucky to attract Timothy Monteath to join Patrick Dunleavy in the core book team as lead researcher and co-author, extending our existing expertise.

The project began as a 'handbook', seeking to collate together existing literature and good practice, and summarize it in a value-adding way. However, our initial hope of drawing down well-worked-out ideas across an extensive extant literature quickly proved unworkable. The literature on open social science is still being written in large part, and will be for perhaps a decade more ahead of us. So instead, we moved from a handbook mode to a more proactive 'guide for researchers' mode – driven initially by running a series of 11 seminars for Civica academics, early career researchers and PhDers across

many disciplines. Here we sought to lead off discussion with stimulating ideas for feasible OS changes and then to listen carefully to reactions and gauge some of the different situations and impulses of our audience members. We are very grateful to all those who attended the sessions and those who commented especially, and to many other academics who answered our questions and extended our knowledge. Subsequently we widened and deepened our research efforts in each topic, and wrote up what we hope are coherent chapters for them that reflect both the current 'state of the art' and potential for changes.

Hence this is an integrative but intermediate exercise – an advice book for researchers. Giving helpful advice on any generic topic in academic life is a difficult thing to do. We want to underscore that the topics covered in our text form no more than a guide or set of suggestions, akin to a restaurant menu or a foreign language phrase book. Our volume tracks current developments and is quite unlike conventional textbooks summarizing well-defined research methods or statistics. Apart from our three years of work in this area, and our experiences as 'early adopter' advocates for open science approaches, we claim no special position or status for the advice here. And we are certain that researchers across a wide range of disciplines will have additional factors that they need to consider. We inevitably will have drawn a lot on our own research and authoring knowledge here, as well as relevant literatures.

So the last thing we want to suggest is that anyone should try to implement every piece of guidance in the pages here. Just as no one sets out to eat the whole menu of a restaurant in a single sitting, so we hope that readers will engage with those parts of the text that seem most relevant for their research and appropriate in the light of their experiences and norms in their sub-field, picking up any suggestions that work for them. And just as foreign phrase books (or even Google Translate) are very hard to use effectively in any actual fast-moving conversation, nothing in this book should be taken to suggest that the transition to open ways of working can be accomplished easily or without a lot of hard work, practice and dedicated effort.

We begin in Chapter 1 by briefly sketching what open social science is all about, and the role that we hope this book can play in encouraging researchers to further advance its development.

Part I next considers the overall theoretical and methodological aspects of changing long-lived 'closed' approaches to research so as to show greater transparency and openness. Chapter 2 considers some 'bricolage' ways in which some rather random open science themes have already reached the social sciences, and some risks that OS might be used as just another disguise for advancing the hegemony of quantitative studies in the discipline group. We reject both prospects, instead defining an ambitious and inclusive open social science agenda drawing on themes developed in the Turing Way project. Chapter 3 fills in some key backstory, looking at how OS started first in science, technology, engineering, mathematics and medicine (STEMM) (i.e. the 'physical' as opposed to 'social' sciences), and showing how much of this

progress can fit within our conception of open social science (OSS). A central theme of OS approaches is registering hypotheses and analysis approaches in advance before undertaking confirmatory research, but Chapter 4 argues that OSS is fully compatible with the greater importance of exploratory research in studying human societies.

Part II reviews the practical issues in improving the reproducibility of research. Chapter 5 shows that fully documenting research decision-making as projects go along is the key OSS development, and considers a few tools that may help. Reusing other people' data and mashing it with your own always raises some tricky issues, as discussed in Chapter 6, which also reviews some useful tools. Next, Chapter 7 examines in detail how to construct a data management plan (DMP) of the kind needed for any substantial research project, and the stages leading up to depositing full datasets for permanent storage in institutional repositories and reuse by other researchers. Rounding out this part, Chapter 8 considers two shorter additional data annexes, the 'replication archives' now needed for most empirical journal articles and the research methods annexes housing backup data and methods information for research books and PhDs.

Part III focuses on the qualitative social sciences, where many OSS goals pose much greater difficulties for researchers in adapting the main empirical methods. Chapter 9 begins by looking at interview-based research and Chapter 10 considers case studies and fieldwork. In both these areas establishing the provenance of qualitative research is especially important and difficult, and there may be constraints on openness and on what researchers can say publicly about sources. Yet mitigating solutions also exist, explored in depth here. Similar problems in going public with evidence occur in documentary work using contemporary sources, but again there are creative approaches that can mitigate them, as Chapter 11 shows. Historical studies and archival research are vital approaches in qualitative social sciences as well as the humanities, yet Chapter 12 outlines how they have been something of a laggard area for OSS. However, it also identifies some early ways forward and grounds for optimism. Chapter 13 looks at 'citizen open social science' where researchers use participatory research, co-production or deliberative democracy approaches. Here impulses to involve research participants in the work are inherent in much of this expanding sphere of qualitative work, and they fit well with making appropriately careful steps to more openness and reproducibility.

Part IV wraps up the book by examining two areas where major steps forward to openness have already been made in recent years, but where the opportunities for further progress remain considerable. Chapter 14 looks at presenting research results in fully clear and genuinely accessible ways, to maximize the audience for research. Chapter 15 covers open access (OA) publishing, now widely accomplished for journal articles, and growing also for books. OA has extensively changed the optimal publishing strategies for researchers seeking to maximize the impact and recognition of their work.

Finally, Chapter 16 sums up the huge cultural shift that OSS entails, looking at the key drivers and barriers to changing researchers' behaviour.

Looking across subjects in a wide discipline group means that there will surely be issues, arguments and contexts that we have not fully appreciated, and some gaps or mistakes. Any comments, corrections, extensions, criticisms, or suggestions that you may have can really help us. So please email or message us using any of the addresses provided below. And if you have enjoyed reading the book and want to help advance OSS as well, why not give us a positive mention on social media – such as, a retweet on BlueSky or X, a 'like' on Facebook or a post on LinkedIn, etc.? We very much hope that the book proves helpful.

Patrick Dunleavy (London School of Economics)
email p.dunleavy@lse.ac.uk
X (Twitter) @PJDunleavy
Blue Sky @PatrickDunleavy.bsky.social
LinkedIn https://www.linkedin.com/in/patrick-dunleavy-46b87b40/

Timothy Monteath (Warwick University)
email Timothy.monteath@warwick.ac.uk
LinkedIn www.linkedin.com/in/timothy-monteath/

Reference

Civica II (2025) 'CIVICA as a ten-member alliance', CIVICA II website, https://perma.cc/93ZP-S2EE

Our references, reusing materials and citing this book

As committed open access (OA) authors, our book is permanently downloadable free from LSE Press either in the three formats available for the whole book (PDF, ebook and Kindle reader) or in the PDF format available for each individual chapter. The whole book is also available in print as a paperback. So we anticipate that the vast majority of readers will use digital formats. To add to the benefits of doing so, the vast majority of references here are also to other OA and the DOI webpage number or permanent URL links for all open access sources are included in the end of chapter references list.

A few of our older references unfortunately still lie behind paywalls, and here in the end of chapter references we indicate paywall sources using the **$** sign and again provide a link to the DOI webpages or an equivalent stable URL wherever possible. Hopefully readers with university or expert library access may be able to access paywall journal articles. Unfortunately, for paywall or paper-only books our DOI or URL links mostly still only lead to the publisher's webpage. But even this possibility, or looking up the volume on Google Books, can normally help readers to see a bit more information, such as chapter headings or some 'snippet' content. That can help readers decide if it will be worth trying to find the source and read more.

For any potentially evanescent content from our references – such as government, public policy or parliamentary reports, consultants reports, think tank publications and all other 'grey literature' sources, as well as blogs and newspaper or magazine articles – we use the PermaLink system from Harvard Law Library (see https://perma.cc/2GTS-R6LU) to make permanent records of each webpage cited. Previously such sources were vulnerable to becoming inaccessible over time, creating 'link rot' problems. We hope that this now cannot occur with this book and we record our appreciation to Harvard Law School Library for this free system.

Reusing materials from this book

As part of our commitment to open science, this book has a Creative Commons licence – of the type Non Commercial No Derivatives (CC BY-NC-ND) (see https://perma.cc/YL3V-RKXK), which means that 'reusers can copy and distribute the material in any medium or format in unadapted form only, for non-commercial purposes only, and only so long as attribution is

given to the creator'. Academics, students and ordinary citizens are completely free to reuse all our materials in any non-commercial publications of their own, and in teaching, blogging, etc., without seeking further permission from us, *so long as* each use has a readable acknowledgement of this text as the source, *and* our text or figures have not been altered in any way. This licence covers all the charts, tables and diagrams that we have produced for the book, as well as text. Unfortunately, it does not cover some small instances of copyright graphics for which we sought permission in a Figure, noted as such in the Acknowledgements.

For permission for any commercial reuse of any part of the book (going beyond using short extracts for academic discussion or review), please contact LSE Press at the London School of Economics at: lse.publishing@lse.ac.uk.

Citing the book

Please cite the whole book as follows:

> Dunleavy, Patrick and Monteath, Timothy (2026) *Doing Open Social Science: A Guide for Researchers*, London: LSE Press. https://doi.org/10.31389/lsepress.dos

To cite an individual chapter of the book, the general format (using Chapter 2 as an example) is:

> Dunleavy, Patrick and Monteath, Timothy (2026) 'How OSS builds on and differs from STEMM open science', Ch. 3 (pp. 39–78) in Dunleavy, Patrick and Monteath, Timothy (2026) *Doing Open Social Science: A Guide for Researchers*, London: LSE Press. https://doi.org/10.31389/lsepress.dos.c

Footnotes on the opening page of each chapter also show how to cite it specifically, including the single-chapter DOI. Note that the author of Chapter 7 is Hannah Boroudjou alone; and that the authors of Chapter 13 are Denisa Kostovicova plus Dunleavy and Monteath.

Page number references will work for readers using PDF or paperback versions of the whole book, and for all PDF chapters, because pagination is stable. However, readers using ePub or Kindle whole-book downloads can change typesizes and hence pagination, so that page number references cannot work for these users. To be maximally helpful for all readers, we recommend including a distinctive micro-quotation of five or six words from our book within your own text, and then linking to the DOI web address file for its relevant book or chapter occurrence. Readers can then search using control+F (or command+F on Macs; or equivalent key in other software) to find the exact quoted words and to place the relevant text fully in its context. Incidentally, this is also a good referencing practice for any other OA works.

Glossary

AI artificial intelligence, see also *DSAI* below

APC article processing charge, a fee charged to the authors by publishers to finance their article being available open access

API application programming interface, software that connects datasets to applications. Public APIs are accessible by third parties, and internal APIs are restricted to the organization owning the dataset

ASI as submitted initially – some publishers who allow *green open access* require that authors can only make freely downloadable the version of a journal article as it was first sent to the journal, excluding any changes or *value added* made by reviewers, editors or the journal's production processes

BPC book processing charge, a fee charged to authors by publishers to finance their book being available open access

CAD creative arts and design

CDA confirmatory data analysis, i.e. conventional statistical approaches focused on tests to support or falsify a hypothesis

CC Creative Commons licences, allowing for text and materials to be reused freely, albeit sometimes with conditions (see Chapter 15, Figure 15.x)

COS Center for Open Science (Charlottesville)

CSS citizen social science

data dredging multiply analysing and reanalysing a dataset with successive analysis methods so as to surface apparently 'significant' relationships that are then interpreted post hoc as if they were initial hypotheses

DD deliberative democracy, an approach to liberal democratic practice that stresses reflective and balanced discussion as key foundations for enhancing the legitimacy of decision-making and seeking to co-produce agreement or policy choices via deliberation with citizens themselves

diamond OA an open access research output that as well as being free to read is also published at no charge to the authors

DMP data management plan – a structured document where researchers outline broadly what data they will use in their research and how they plan to manage, store, secure and share it in the long term (see Chapter 7)

DPIA data protection impact assessment – a tool for avoiding risks when collecting sensitive personal data where the research topic or methodology are likely to result in higher risk to individual research subjects or communities (see Chapter 7)

DSAI data science/artificial intelligence – modern techniques for using big data to create software that can learn from analysis in real time and use algorithms, machine learning and sophisticated maths or statistics-based extraction techniques to analyse massive datasets. DSAI also sustains automated machines and robotics, *generative AI* and large language models (*LLMs*)

ECRs early career researchers

EDA exploratory data analysis, forms of search across data undertaken to inductively generate new understandings or theory, uncovering possible relationships and surfacing potential new variables and hypotheses. EDA is an especially apt approach with new data or social phenomena, and in searching for possible innovations – e.g. in very large drugs/pharma trial databases

FAIR a set of principles for good data management that specify that data from research projects should be findable, accessible, interoperable and reusable

FOSS free and open source software, the movement in STEMM advocating software that is free to use, modify and redistribute

FPTP first past the post – see *plurality voting*

FOI freedom of information provisions (also FOIA) – rules allowing citizens to request all official information or documentation from their governments – in theory, but often heavily constrained in practice by wide exemptions and long response delays

fuzzy set approaches an approach to qualitative analysis applying fuzzy set theory to case-orientated data – see *QCA qualitative comparative analysis* below, and Chapter 4

GDP gross domestic product, the total value of goods and services produced in an economy in each period

GDPR the General Data Protection Regulation in the European Union (EU) which protects the privacy of personal data within the EU and the wider European Economic Area (EEA) by specifying how researchers must safeguard their data and research participants' anonymity

Generative AI a type of DSAI (or AI) software that can reuse data already inputted or learnt from by the system in order to answer queries and to generate new text, images, music, videos, or other novel content

GFC global financial crisis – the financial and austerity crisis in many developed countries from 2007 to 2011

github a web-based platform that allows developers of software and datasets to store, create and collaborate on code and data, track changes to code, manage projects, and make these widely available

green OA a severely limited form of *open access* where the final form (and most heavily publicized version) of an output remains completely paywalled, either permanently or for an embargo period. However, a submitted manuscript or pre-publisher version of the text can nonetheless

be downloaded free from the authors' personal web site or institutional repository

gold OA an output is immediately available open access in final form for individual readers and from the original publisher

Harking hypothesizing after results are known. Analysing the complete dataset when available to look for any unanticipated and apparently significant relationships; and then reconstructing a plausible explanation that is presented as an initial hypothesis 'tested' by the data. See also *data dredging* above.

HD hypothetico-deductive, a scientific research method stressing the prior production of specific and potentially falsifiable hypotheses that are tested empirically

HDS human-dominated systems, those parts of global civilization, societies and environments that are decisively constituted or shaped by human interventions and technologies (see Chapter 2)

HIS human-influenced systems, those parts of global physical environments where 'natural' or autonomous physical systems (still) operate strongly, but there are also major influences from human interventions and technologies (see Chapter 2)

IP intellectual property – inventions, artistic or literary works, images, and symbols or trademarks – especially those legally protected by copyrights, patents and registration of designs

IT information technology

LGBT+ Lesbian, gay, bi-sexual, trans-sexual+. This is the most commonly used acronym internationally

LGBTIQ+ Lesbian, gay, bi-sexual, trans-sexual, intersex, or queer+. This is the recommended official acronym in some countries, such as Australia

LLM large language model, one of the most widely used forms of generative AI. Accurately prompted LLMs can help produce research-level academic texts, e.g. ChatGPT

log file a file automatically generated by a PC or device recording all operations carried out by the user and the outputs, activity or changes that resulted

NDA non-disclosure agreement, a legally binding contract limiting or prohibiting the publication of some specific kinds of information known to researchers

NGO a non-governmental organization such as a charity, community organization, advocacy or research organization, or other civil society group

Non-reactive data or methods objective or behavioural approaches where information about research subjects is collected from sources without their being asked about it – e.g. analysing social media or transactional data, or documentary evidence

null hypothesis significance testing (NHST) an approach that uses statistical inference to determine whether a hypothesis that variable *A* has no

effect on variable *B* cannot be falsified (see Figure 4.1), or whether that hypothesis can be rejected

OECD Organization for Economic Co-operation and Development

OA open access, article or book outputs that are freely downloadable by any reader

OAPEN a leading open access books site in Europe and globally (www.oapen.org)

open data data made available in a machine-readable and downloadable format so that anyone (with a PC or other device) can access it, use it and share it with others

open notebooks computerized notebooks for recording all research data operations, data modifications and results

open source a form of software where users can freely access the code and reuse and modify it themselves – see also *FOSS*

OS open science – the movement pushing for greater accessibility, reproducibility, transparency and reliability in academic work, especially in STEMM (see Chapter 3)

OSF Open Science Foundation (https://osf.io/) – a leading advocacy foundation for open science and with a large pre-publication repository

OSS open social science – a version of OS adapted to and appropriate for the social sciences

PAP pre-analysis plan, registered before research is undertaken, filed on an open science repository, setting out the hypotheses to be tested in research and the analysis methods to be used

PAR participatory action research, an approach that seeks to involve research subjects (e.g. individuals, groups, communities or 'stakeholders') in helping to define (aspects of) the research process itself, and in evaluating or using the results found

***p*-hacking** where a relation between variables is initially found to be 'non-significant' by conventional tests for 'significant' *p* value results, *p*-hacking entails editing or adjusting the relevant data or analysis undertaken so that the relation in the now-modified dataset can (marginally) pass these thresholds

***p* value** the statistical probability that an apparent relationship between two variables in a dataset has arisen by chance, given the size of the dataset. Results like $p <= 0.01$ or $p <= 0.05$ have conventionally been judged as 'significant'

PhDers graduate students completing a PhD degree

platforms proprietary digital sites/systems offering basic data storage alongside many other functionalities, including tools for collaboration, dissemination and analysis

plurality voting an electoral system where the party or candidate in an election district with the largest vote wins, irrespective of whether they have an overall majority of votes or not. Also called 'first past the post' (FPTP)

PR proportional representation, any system of voting where seats in a democratic legislature are distributed in proportion to the votes won by parties or candidates

Pre-registration registering pre-analysis plans (PAPs) on an open science repository, setting out the hypotheses to be tested in a planned piece of research and the analysis methods to be used

RA replication archive, a data annex file accompanying a journal article and at a minimum giving full details of how each table and chart included was created, and the variable definitions and analysis methods used, etc. An RA may also include relevant parts of the original dataset underlying each exhibit (but usually not the whole dataset)

reactive data or methods approaches that rely on asking research subjects to respond to questions, prompts, or stimuli, e.g. surveys, interviews, etc.

QCA qualitative comparative analysis – a case-based research approach (most commonly employed for 10–50 cases) where set theory, Boolean algebra and minimization algorithms are used to explain why effects are seen or not seen in varying combinations of case conditions (see Chapter 4)

RCT randomized control trial – a study that randomly assigns research subjects to one or more treatment groups and a no-treatment control group so as to determine the effectiveness of the treatments. Normally RCTs are conducted 'double blind' so that neither the participants nor researchers know which group is which, and the results are assessed across time

REF Research Excellence Framework, the government-run evaluation of the research outputs of universities in England, Scotland, Wales and Northern Ireland, operating from 2008 to now, with a next evaluation date of 2029 (the 2004 Research Assessment Exercise was similar)

RMA research methods annex, a data annex file for a book, research project or PhD thesis, at a minimum giving full details of how each table and chart included was created, how the variables shown were designed and what analysis methods were used. RMAs may also include parts of the original dataset used (or the whole for small *N* datasets)

repository a digital facility designed for the long-term storage and dissemination of datasets and information from research projects, usually with a developed (web) user interface

reproducibility researcher B can analyse the same dataset as A in the same way and come up with the same results as A

replication researcher B analysing a similar but different dataset from A comes up the same, similar or consistent results as those found by A

scratch pad paper, or simple tablet or PC software used to quickly make small notes of results or records of actions done, for temporary retention

social sciences the discipline group studying human and social behaviour in all its aspects, and comprising the subjects shown in Figure 2.1 (see Chapter 2)

STEMM the discipline group including all the sciences, technology, engineering, mathematics and medicine

STS science and technology studies, the social science, philosophical and interdisciplinary analysis of how STEMM science research creates and develops new knowledge, set in its wider historic and societal context

VPN virtual private network, a digital point-to-point link from a user A to a proprietary remote server that serves to disguise A's online address

Zenodo free online research data and information repository run by the EU (https://zenodo.org/)

1. Introduction – using this book in research

Open social science (OSS) is the development and application of an 'open science' approach across the whole social sciences discipline group. This is simple to say but a complex task to get done. In our view OSS approaches represent the most important effort in recent times*

- to strengthen the reliability and usefulness of the social sciences, including many shared fields between our discipline group and the humanities, and overlap areas with science, technology, engineering, mathematics and medicine (STEMM) (see Chapter 2);
- to improve the 'provenance', reproducibility, reliability and generalizability of social science research;
- to encourage the re-analysis, iteration and cumulation of research over time and across researchers;
- to connect this research more effectively to 'ordinary knowledge' and developed bodies of expertise outside academic work (e.g. in public policies and organizing businesses and markets effectively);
- to democratize access to social science research for citizens and other potential users in civil society; and
- to build public trust in the value of the discipline group's scientific research.

Enhancing the openness of academic work requires an important change of working culture (or organizational cultures) amongst researchers themselves, one that only they can accomplish. Funders can encourage open science approaches. And academic-related professional support staffs in many parts of universities (like libraries, research offices, and PhD schools and academies) can help to disseminate (general) knowledge about how to make

How to cite this book chapter:

Dunleavy, Patrick and Monteath, Timothy (2026) 'Introduction – using this book in research', in Dunleavy, Patrick and Monteath, Timothy (2026) *Doing Open Social Science: A Guide for Researchers.* London: LSE Press, pp. 1–6. https://doi.org/10.31389/lsepress.dos.a

research more open and transparent. But if the change is to be sustainable, it is researchers themselves who must adopt diverse innovative initiatives and approaches, meshing them together with their existing research methods and practices. Social scientists will need to autonomously adopt 'open' practices for their own, good reasons – because open works better to get research done, explained, accepted, reused and acted on. But what open means for social science research (and cognate areas of the humanities) has yet to be fully settled. So it is up to researchers to figure out and create OSS across many diverse subfields, and to evaluate how it can better further their research.

The OSS agenda seeks to establish more clearly than ever before the provenance of research – in a kind of academic 'show and tell' that makes visible not just the final writing up of results and accompanying exhibits but also the underlying evidence and data in ways that can be easily checked, reused and cumulated in the digital era. Being open about evidence, plus (hopefully) achieving open access across all research publications, can enable iterative improvements in knowledge via the ability to track changes and previous analysis through well-documented data sources and transparent publications.

The main changes we discuss are relatively new in the social sciences, but they have been developing for a longer time elsewhere in the physical sciences, the STEMM disciplines, in universities, and in some industrial research firms. Improvements have been made in many diverse fields towards greater openness and transparency in how research is designed, undertaken, reported and published, with an accelerating tempo in last 10 years (Ramella and Manzo 2021). The huge promise (and also the complexity) of the open science movement in this century was well captured by Vicente-Saez et al. (2020), who wrote:

> We specify openness as a multidimensional variable that can be measured and formulated by means of the proposed levels of transparency of science outputs, accessibility to science outputs, authorization in science production, and participation in science production. This new policy is a tool for building local, national, regional, and global knowledge communities and raising the welfare level of each. We are at the dawn of an open exploration era. (p. 8).

The same open science movement has had a big impact in some crossover areas between social science and the STEMM disciplines, such as social psychology and health studies, and in science and technology studies (STS). But elsewhere not much discussion has yet covered what a push for greater openness means in the many different forms of social research. Yet we show in Chapter 3 that many useful lessons appropriate for the social sciences can be drawn.

This book reviews how open science themes matter now to academics and researchers across the full range of social science core disciplines and diverse

other fields now shared with or influenced by social research (see Chapter 2). It seeks to suggest some useful guidance, distilling out the available experience of both the benefits and the difficulties of doing OSS. We have written this book for researchers in the hope of reaching a wider audience across our discipline group – not just in quantitative research where the past awareness and transfer of open science ideas has been greatest, but also in qualitative fields where devising more open research methods has so far been less developed.

Embracing OSS is not just about learning how to do specific things differently or acquiring a small number of discrete new skills or technology competencies. It is instead about making a far-reaching change in academic culture, one that is distinctively modern (see Chapter 16). For anyone aged under 40 (whether they are now doctoral students, early career researchers or younger full-time faculty) achieving greater openness in their research is likely to be a career priority. Finding workable paths to doing more open research will surely play a major role in their daily work life and help to shape their careers' direction of travel over the next 20 or 30 years (at least). More experienced academics, with well-established positions, are also often the strongest advocates for and practitioners of more openness in research.

There is perhaps a risk that some established faculty (especially at elite universities) might feel that this new agenda has less relevance for their future work or career paths, given that most big or 'top' journals have already transitioned to open access publishing, and major institutions (at least) have found routinized ways of paying for that. However, open goes far beyond simply publishing open access, as we show below. Even publishing journal articles increasingly means 'publishing with supporting appendices, a replication archive and a full dataset deposited in a major repository'. In addition, all senior faculty must give up-to-date and sound counsel and advice to younger colleagues, and that is already impossible to do without being alert to OSS trends. And, of course, principal researchers (running projects with teams) both bear this responsibility in an acute way and must respond to pressures from funders and their universities for greater openness. So we hope that this book is for everyone doing research, not just younger cohorts, and the whole text is written with this strongly in mind.

Our focus here is always on doing more open research better from the outset, strengthening approaches that genuinely have value for researchers and that stem organically from the nature of the work being undertaken. Always framing things from a researcher perspective is integral to the book, and most chapters concern practical steps needed to embed an open approach into different kinds of research in the most effective and productive ways, as the green and yellow rows in Figure 1.1 show.

As a corollary, the grey and white rows in Figure 1.1 show that we mention but pay less attention to some other 'official' or policy aspects of OSS – including the rationale for governments and funders to promote open science; the different open science policies that have developed across countries (e.g. the USA and UK) and international bodies (e.g. the EU); and the details of how

Figure 1.1: The main themes covered in this book (green and yellow rows) and less covered topics (white and grey rows)

Open science main themes	Sub-themes	Coverage in this book
Defining open social science (OSS)	The main elements and ambitions of OSS	Chapter 2
	STEMM approaches to open science	Chapter 3
Open reproducible research	Pre-registration and exploratory approaches	Chapter 4 and Chapters 9–13
	Documenting research and meeting open data standards	Chapter 5 and Chapters 9–13
	Reusing existing data and fostering greater cumulation of knowledge	Chapters 6, 11 and 12
Open data	Data management plans	Chapter 7 and some comments in Chapter 13
	Using repositories	
Open science tools	Open lab notebooks	Some comments in Chapter 5
	Replication archives and research methods annexes	Chapter 8
Open access publishing	Communicating research results well	Chapter 14
	Publishing strategies	Chapter 15
'Openness' as a cultural change in doing research	Key drivers, barriers and mitigations for fostering OSS innovations	Chapter 16
Open science policies and funding requirements		Some comments in Chapter 7
History of open science's development		Many aspects touched on in Chapter 3
Open science approaches to research evaluation, metrics, etc.		Some comments in Chapters 15 and 16

Source: This table draws on but extensively re-presents ideas included in the FOSTER 'open science taxonomy' discussed in Tennant et al. (2020).

researchers need to comply with externally imposed ('bureaucratic') hurdles or requirements. Similarly, we have not given a complete historical account of how OSS has evolved from earlier open science trends in STEMM, although some key synoptic aspects are covered in Chapter 3. We have referenced but not tried to duplicate the existing literature on the historical development of open science but focused on the current picture. Nor have we covered work in educational research and philosophy that concerns the genealogy and early motivations for open science advances. However, our final chapter (Chapter 16) briefly covers the most general aspects of the cultural changes currently needed to adapt to the digital OSS environment in future.

By framing our text in this way, focusing on practical issues and ways forward for practising researchers, we nonetheless also hope that two other potential sets of readers may gain valuable insights:

- the many accomplished research support professionals and training staff in libraries, graduate schools and other locations, who currently provide researchers with most advice and help on open science issues; and
- the senior decision-makers on university boards and committees considering policy choices about encouraging OSS.

We hope that readers from both groups will better understand how closely choices around openness interlock with other aspects of social sciences research, and how these interactions may be most productively fostered and developed in our discipline group, the digital humanities, and perhaps also neighbouring areas of the humanities. The text may also help teachers advising multidisciplinary groups of graduate students to understand the mass of expertise about OSS that exists in different subjects across the discipline group.

However, to avoid any misconceptions perhaps we can end by reiterating a point made in our preface, one that applies to this book or indeed any 'advice' text in academia. This book is a menu of suggestions only, not any form of specific blueprint for change. We are far from envisaging phalanxes of social scientists changing their research practices in exactly the ways recommended or laid out here. Readers instead need to pick and choose within this text only those lessons or ideas of interest, that they find useful, paying attention only to advice that fits with their own situation, experiences, needs and orientations, and adapting it to *their* research.

Note

* We are grateful to Sarah Worthington (Law LSE), Philippa Grand (LSE Press) and an anonymous reviewer for advice and comments on this chapter.

References

Ramella, Francesco and Manzo, Cecilia (2021) *The Economy of Collaboration: The New Digital Platforms of Production and Consumption*, London: Routledge. https://perma.cc/4V8W-26Q9

Tennant, Jonathan P; Agrawal, Ritwik; Baždarić, Ksenija; Brassard, David; et al. (2020) *A Tale of Two 'opens': Intersections Between Free and Open Source Software and Open Scholarship*, Charlottesville, VA: Center for Open Science. https://doi.org//10.31235/osf.io/2kxq8

Vicente-Saez, Ruben; Gustafsson, Robin; and Van den Brande, Lieve (2020) 'The dawn of an open exploration era: emergent principles and practices of open science and innovation of university research teams in a digital world', *Technological Forecasting and Social Change*, vol. 156, no. 120037. https://doi.org/10.1016/j.techfore.2020.120037

PART I
Opening up the social sciences

The open social science (OSS) programme of cultural change seeks to maximize the accessibility and reproducibility of research by changing the research process itself to make it more reliable, establishing sound data management policies and practices, making the cumulative body of research work easily accessible, and using open access publication and improved communication of research findings to reach the widest feasible audiences. The binding element in this agenda, that makes it cohere as a movement and a project, is increasing *openness*, implemented in ways that help maximize the access of other disciplines and academics, together with citizens, enterprises and civil society at large, not just to the conclusions of studies, but also to their evidence, methods of analysis, limitations and essential contexts.

Part I tackles three key more 'theoretical' issues involved in the social sciences not only 'catching up' with open science (OS) progress already made in science, technology, engineering, medicine and mathematics (STEMM) disciplines, but also defining their own distinctive version of what 'openness' means (Chapter 2). This ambition has a special significance in the social sciences because in all of them research and 'professional social inquiry' inherently seek to produce 'useable knowledge' that carries through from theory, via empirical research and into applied work. Arguably too, all social scientific work relies heavily on well-founded 'ordinary knowledge' in order to connect up small, 'pinprick' islands of formal research scattered across a broad societal landscape; and to continuously analyse human societies that constantly adapt and evolve in response to rapid technological, social, economic, political and cultural changes (Lindblom and Cohen 1979). The social studies of science (STS) sub-field has also greatly expanded academic appreciation of the many forms of non-science expertise (Collins and Evans 2009).

We next look a bit more deeply at what OS has meant across the full range of STEMM science disciplines (Chapter 3), encompassing both the most 'advanced' OS sectors in computational subjects or physics, and other STEMM disciplines (like ecology and medicine) that offer salient lessons for how OSS can proceed. We show how some fundamental differences are likely to characterize OSS when compared to OS. To round off Part I, Chapter 4 explores a single conspicuous area of apparent tension between the emphasis of open approaches on pre-registering expectations about how research will pan out in advance of conducting research and the emphasis in qualitative social sciences on undertaking 'exploratory' research prior to and in addition to confirmatory testing work.

2. The 'open social science' agenda

Modern social science developed largely in the shadow of the science, technology, engineering, medicine and mathematics (STEMM) discipline group. These sciences have had unique prestige, coherence and recognition within all advanced industrial societies – for instance, in the USA they massively dominate federal research grants (80% in 2021), states' funding and almost all corporate funding for research and higher education. By contrast, the social sciences as an interconnected discipline group have struggled along behind, with much less (uncontested) state recognition, much smaller budgets, far fewer full-time researchers and (often) lower salaries (Dunleavy, Bastow and Tinkler 2014; Bastow, Dunleavy and Tinkler 2014). With some exceptions (such as economists), many social scientists have tended to be very aware of their 'second-class' status with grant funders (and the media and general public) compared with 'pure' and applied science colleagues (Montuschi 2008, 2003).*

So, as the open science (OS) movement has developed in recent decades, there has been a temptation to simply pick up and copy across a variety of STEMM initiatives in unadapted forms. This approach creates a risk of partial or ill-judged OS 'false starts', potentially creating misunderstandings about what the agenda is for, and considerable resistance amongst researchers and controversy in consequence. Successfully unifying an agenda across the 40 plus interconnected but diverse social science disciplines also raises difficulties that can easily seem intractable (Caws 2005). We argue here that they can be overcome.

We begin by briefly mapping the disciplines and subjects that constitute the social science discipline group, including some joint/overlap subjects with STEMM and humanities disciplines. Section 2.1 also reflects a little on how social science now operates in contemporary conditions. We draw a contrast between its modern, more interdisciplinary operations and past siloed modes of working, often still over-coloured by a deference to the earliest

How to cite this book chapter:

Dunleavy, Patrick and Monteath, Timothy (2026) 'The 'open social science' agenda', in Dunleavy, Patrick and Monteath, Timothy (2026) *Doing Open Social Science: A Guide for Researchers.* London: LSE Press, pp. 9–38. https://doi.org/10.31389/lsepress.dos.b

'founding authors' of disciplines, and perhaps too by an 'inferiority complex' versus STEMM.

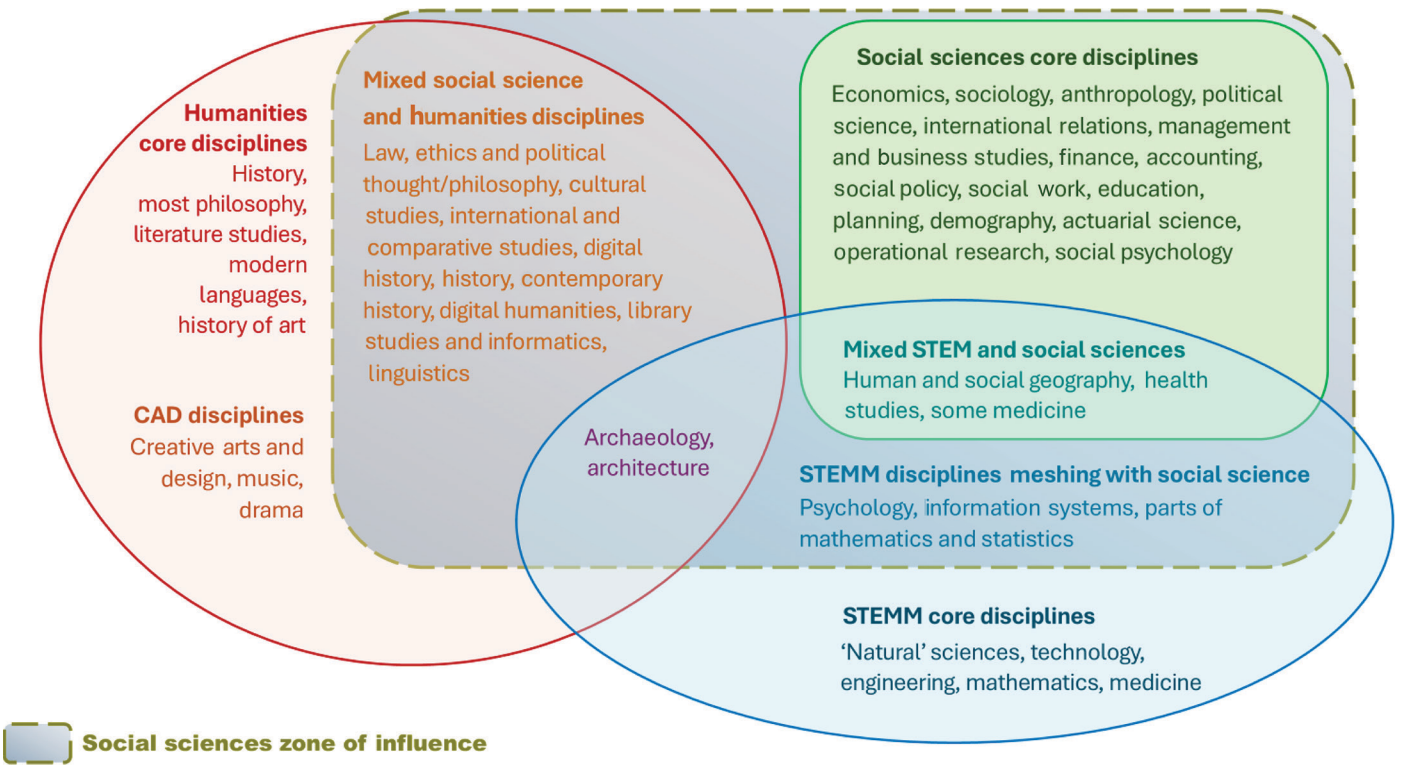
Just as social scientists moved to adopt peer review as standard in the period from the 1950s to the 1970s, so in the last decade some have wanted to quickly follow along with the conventional OS agenda (Christensen, Freese and Miguel 2019) – especially in terms of strengthening impartial knowledge development against abuses arising from the modern rise of 'paywall' journal science (Ionnidis 2005; Pennington 2023). Yet some early efforts developed to 'transfer' lessons or practices from OS to the social sciences have been framed in rather premature and ill-considered ways (Section 2.2).

By contrast, we seek to define OSS in a better, more inclusive way – covering the discipline group as a whole without excluding or marginalizing any disciplines or sub-fields; and avoiding the unfortunate contrasts between quantitative and qualitative research modes that plagued earlier approaches (Section 2.3). The transition to open is about making all types of social science research more reproducible, replicable, robust and generalizable (the Turing way's definition of 'open', covered below). We believe this agenda is one that every responsible social scientist can support and contribute to, whatever type of research they do.

2.1 The social sciences discipline group

In the past all academic disciplines were famously siloed, singular and rather 'tribal'. They often seemed primarily concerned to establish their own in-group against outsiders, carving out a unique turf where the discipline's values, priorities and standard operating practices could exist substantially unchallenged. Until the 21st century, the social sciences perhaps over-conformed to this model. The discipline group came to be composed of around 40 different discipline and sub-disciplines, with the core subjects shown in Figure 2.1. Social science overlapped with a number of STEMM sciences in major ways in geography, archaeology, psychology, health studies, and information sciences and IT – sometimes creating separate 'social' sub-disciplines most akin to other social sciences, such as human geography and social psychology. Lesser overlaps have occurred with engineering and parts of mathematics and statistics. An extensive overlap area of mixed social sciences and humanities subjects has also been created, including key disciplines like law and cultural/media studies, and substantial parts of history and philosophy where social science perspectives and methods are increasingly constitutive of modern research. Thus, the social sciences now have a substantial 'zone of influence' where their distinctive theories, methods and data extensively shape how research is undertaken in mixed or allied subjects within the STEMM disciplines and the humanities. This zone extends far beyond the core social science discipline group and constitutes a large part of the modern academic world, covering all aspects of human and societal behaviours and conditions (shown in Figure 2.1).

Figure 2.1: The social sciences discipline group, mixed subjects and the social science 'zone of influence' – plus neighbouring discipline groups



Source: Dunleavy 2014, Slide 11.

What makes these numerous, diverse and allied approaches into a coherent and interconnected discipline group, rather than just a listing of separate academic efforts? One recent influential account argued that:

> The key common features are:
>
> - They focus on the study of contemporary human societies, economies, organizations and cultures, and their development.
> - The intellectual spine of all these subjects is provided by formally set out theories, normally developing logically consistent 'models', often utilizing mathematical notation, but always with distinct rules and logics of theory development.
> - They focus a great deal on systematically collecting data and information using well-worked out and rigorously tested methods, with most branches making significant use of both quantitative and qualitative data. All social sciences look for 'laws' of social development, for patterns of association and causation that make sense theoretically and can be evaluated by careful empirical investigation [*using both quantitative and qualitative research methods*].
> - Finally, the social sciences strongly share or seek to emulate standards of good science and of effective scholarship as developed in the physical sciences, stressing the importance of using carefully checked data, analysing data rigorously, replication of information, critical testing of evidence and critical engagement with theories and models, and a conditional acceptance of 'knowledge' only to the extent that it survives falsification' (Bastow et al. 2014, p. 4). (*Our additions in italics.*)

The unity of the discipline group has also been transformed by contemporary trends. We live in a fully globalized world now, on a single planet that has been extensively terra-formed by our overwhelmingly dominant species. Human artefacts have spiralled in significance, for instance, when calibrated simply in terms of their physical weight:

> Humanity has become a dominant force in shaping the face of Earth. An emerging question is how the overall material output of human activities compares to the overall natural biomass. We [can] quantify the human-made mass, referred to as 'anthropogenic mass', and compare it to the overall living biomass on Earth, which currently equals approximately 2.1 teratonnes. We find that Earth is exactly at the crossover point; in the year 2020 (±6), the anthropogenic mass, which has recently doubled roughly every 20 years, will surpass all global living biomass (Elhacham et al. 2020, p. 442).

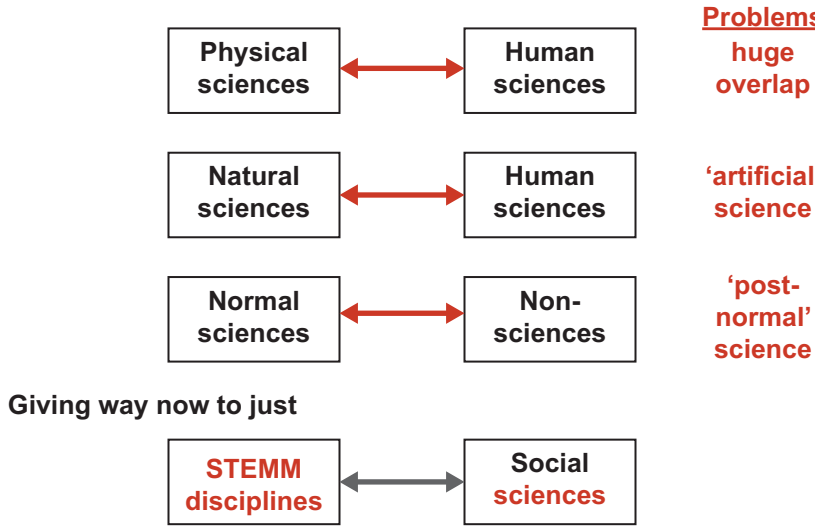
The biomass of all the Earth's animals is 4 gigatonnes, half that of the plastics in the world. And the biomass of all the trees and plants is 900 gigatonnes, now comfortably exceeded by that of all the buildings and infrastructure at 1,100 gigatonnes (Elhacham et al. 2020, Figure 3). While only around 1% of all habitable land is taken up by urban areas, a further 46% of that total is used for farming (Ritchie 2019). And the designation of the Anthropocene as a distinct epoch in geological time (or alternatively as an 'Event') reflects the fact that human activities have visibly begun to affect even the Earth's climate systems and the long-run physical evolution of all its habitats and landscapes (Gibbard, Walker, Bauer et al. 2022).

These changes stress the increasingly vital importance of integrated or trans-disciplinary work, especially spanning between the social sciences and STEMM in studying two main types of integrated systems operating in the physical and social worlds (Dunleavy, Bastow and Tinkler 2014):

- **Human dominated systems** (HDS) are phenomena like cities, markets, economies, institutional structures, health systems, public policies, cultures and technologies. This category includes a large and rapidly expanding area that Herbert Simon long ago termed 'the sciences of the artificial' – spanning computer science and IT, most engineering, mathematics, medicines and drugs, and other physical technologies (Simon 2019, originally 1969). But the HDS concept also now includes many physical systems, terra-formed landscapes and areas of technological activity where human behaviour has become the dominant shaping force of what happens in the 'physical' world.

  In all these fields traditional concepts of the STEMM disciplines as studying autonomously operating 'physical' or 'natural' phenomena have little applicability. Of course, across all human-dominated systems STEMM disciplines and knowledge continue to play the key part in understanding how they operate. But there is no autonomous 'natural' world or set of systems that exist unshaped by human and behavioural uses. And traditional polarities sustaining STEMM disciplines' intellectual dominance for the last two centuries have begun to be refounded (Figure 2.2). HDS and artefacts have created huge overlaps between 'physical' and 'human' systems (see below). The 'natural' and 'human' sciences interpenetrate across all HDS, while Simon's 'sciences of the artificial' operate in a no-man's land that is fully neither of these. The conception of 'normal' science as being STEMM disciplines alone (despite being loosely derived from Popper), with all other formal knowledge condemned to a 'non-science' category, has long since collapsed – and such arguments now read as little more than propaganda devices untethered from the realities of research. Both the social sciences and major parts of the STEMM discipline group have moved in different ways to reasonably convergent, 'post-normal science' modes of operating.

Figure 2.2: Traditional false polarities between the 'natural' or 'physical' sciences and the social sciences



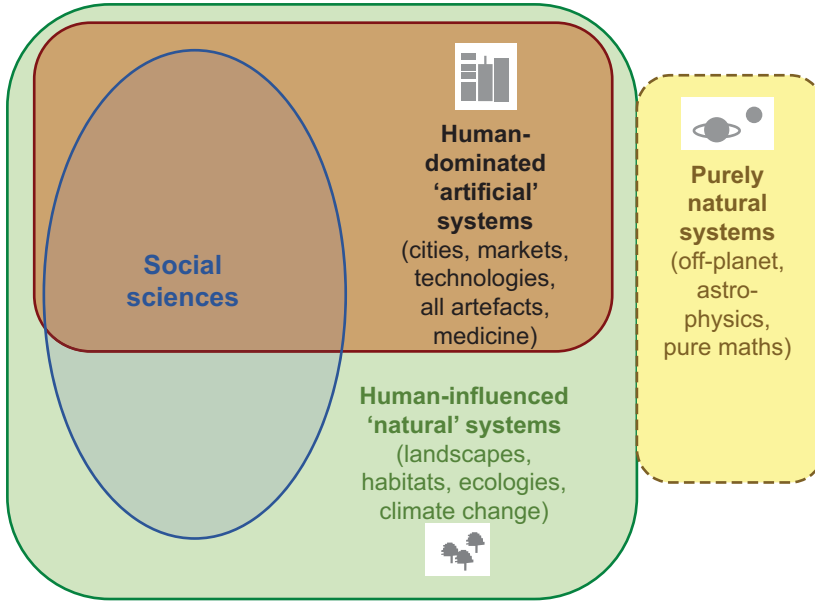
Source: Dunleavy, Bastow and Tinkler (2014).

To fully understand or effectively manage what is going on in HDS entails far better-integrated knowledge development, spanning across both the STEMM and social sciences in mixed modes. A convergence of the two discipline groups is simply unavoidable – and all single-silo approaches have strictly limited applicability (as Figure 2.3 shows).

- **Human influenced systems** (HIS) are those where partly autonomous 'natural' or physical processes still predominate over human interventions in separable ways that can be studied in STEMM-only ways, but where human interventions have nonetheless restructured their operation in a significant manner. HIS dominate most of what happens on the world's surface beyond the HDS sphere. There are now few if any fully natural processes left on Earth, beyond the geo-physical processes shaping continental drift, volcanoes and earthquakes. And even here many of our research concerns focus on trying to predict or manage the impacts of these processes on human settlements. The global change from natural systems to human-influenced ones has opened up new fields for investigating societal influences on physical systems, including global warming and climate change and their potential mitigation at the macro scale.

However, social scientists have been reluctant both to recognize the increased integration of both HDS and HIS, or the need for interdisciplinary efforts with STEMM scientists. The modern importance of their own research efforts meshing with STEMM knowledge for preservation and advance of human civilization

Figure 2.3: The place of the social sciences within the study of 'human-dominated systems' and 'human-influenced systems'



Source: Adapted from Bastow, Dunleavy and Tinkler (2014).

that Figure 2.2 suggests is also underplayed. One factor behind this academic conservatism has been a long-running problem of disciplinary over-siloing on historic lines. Christakis (2013) argues that key subjects have remained in thrall to the discipline-shaping influence of their founding fathers in the 'establishment' periods from the 1880s to the 1920s (dominated by figures like Marx, Weber, Durkheim, Marshall, Freud, Jung, Mead, etc). And while new sub-disciplines have been added, none of the 'old' topics have been set aside or left behind (even if they are now 'solved' or well-understood). Unlike the STEMM sciences:

> The social sciences have stagnated. They offer essentially the same set of academic departments and disciplines that they have had for nearly 100 years: sociology, economics, anthropology, psychology and political science. This is not only boring but also counterproductive, constraining engagement with the scientific cutting edge and stifling the creation of new and useful knowledge. (Christakis 2013)

Yet, Goodall and Oswald (2014) ask:

> Is that a problem? Descriptors do not matter much in themselves. The truly serious issue is whether immovable department names are a worrying signal of immovability of thought. Sadly, we suspect that, in the UK at least, they are.

By contrast, Davies (2014) dismisses Christakis' exaggerated deference to STEMM patterns as resting on a false premiss.

The second source of constraints on change in the social sciences has been the lesser prestige and influence of the discipline group's research methods compared to those in STEMM disciplines. Randall Collins (1994) locates this higher status in physical science value systems grounded on a belief that Nature (with a capital N) is autonomous, unchanging and constant. In principle this means that (genuine) new discoveries at the research frontiers in the 'natural sciences' endure for all time. Hence, they deserve super-high prestige and rewards – kudos that applies even if longer run changes in scientific theories and advances in measuring capacities in practice often imply continuous reinterpretation of the 'facts' at the research frontier. By contrast, STEMM disciplines assign much less significance or academic kudos to scientists who only rework or integrate already established knowledge, around which a single, shared or dominant view is most commonly maintained. This 'rapid advance, high consensus' model triumphed in STEMM in the late nineteenth century and has continued to develop strongly since.

By contrast, Collins argued that the social sciences neither achieved rapid advances in generating (agreed) new enduring knowledge, nor proved able to sustain strong professional consensus on already-studied areas. Instead, new advances were often disputed and contested by (two to six) rival theories, rooted in divergent philosophies and methods, and assigning meaning to social developments in different ways. (In analysing the patterns prevailing in philosophy Collins (2000, Chapter 15) posited that there are always two to six approaches in contention – two being the logical minimum, and six the empirical maximum number.) Constant societal change also produced a churn of theory perspectives and some non-cumulation of empirical work. Consequently, professional consensus in fields like sociology, political science or economics has proved elusive.

Over time, many social scientists have sought to advance their own discipline by adopting research approaches and methods that have worked well in particular, high-prestige STEMM disciplines and moving them across to apply in their own field. Some important and sustained areas of major advances have resulted, including:

- a universal shift towards full peer review in (reputable) social science journals;
- far more rigorous confirmatory testing of hypotheses, propositions and wider 'perspectives';
- the useful 'mathematization' and formalization of parts of key disciplines, following the longer-lived pattern of change in biology and medicine (Valentinuzzi and Kohen 2013; Hartmann and Sprenger 2010) – although not all mathematization has proved 'useful';
- the adoption of improved statistical and data-summary-and-display methods; and

- the faster deployment of successive new and more advanced IT systems and computational methods in social research.

Each of these changes has contributed to the rapid evolution of the modern discipline group, but not without strong controversies. Some critics have also complained of 'physics envy' attitudes creating dead-end fields where models of disciplinary endeavour imported wholesale from chemistry or physics can only be superficially or inappropriately applied (Primo and Clarke 2012; Nelson 2015). Where the over-development of formal methods has exceeded the capacity for meaningful empirical testing or application, formalizing but not testing strategies may obscure underlying situations. A much-cited example concerns economics and financial theory, where professionals' faith in mathematized 'perfect markets' models from the 1980s up to 2007 contributed strongly to disguising the cumulating onset of a disastrous, 'black swan' global financial crisis (Taleb 2007; Besley and Hennessy 2009).

2.2 Five 'false starts' for 'open' in the social sciences

Open social science is just starting out, but for it to take root its concept of openness must be adapted or specific to the social sciences as a discipline group. It must be capable of addressing the wide range of concerns addressed by the 40 or more main fields noted above, and the sharply varying conditions in which social research must be carried out. Some premature or partial efforts already made to 'graft' OS components from STEMM disciplines onto existing methods have created three false starts:

- some rather random direct transfers of STEMM practices;
- pitching 'open' as a change exclusively driven by poor research integrity issues; and
- presenting open working practices in ways that are only practicable for quantitative disciplines.

Each of these orientations can work (or already has) to obstruct the widest development of open social science, as we show next.

Transferring components directly across from STEMM (or bricolage)

This approach sees OSS as just about picking up some of the best-known or most discussed OS components in the STEMM disciplines – focusing on specific innovations such as open access publishing, the FAIR (findable, accessible, interoperable, reusable) principles of good data management for making

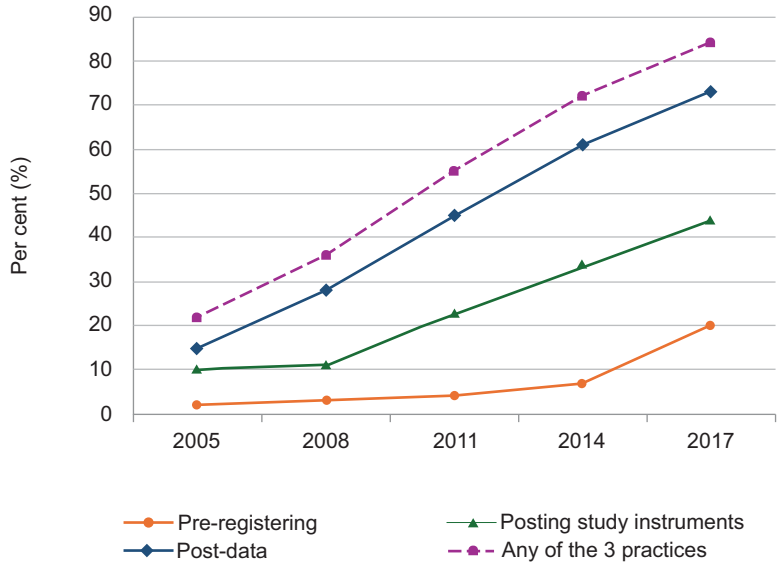
datasets publicly available (see Chapter 7), undertaking replication studies, or the pre-registration of hypotheses and research strategies. Important though they are, these components are only symptoms and researchers using them in isolated ways may not speak to any deep conversion to more innovative working.

Nonetheless, at least one empirical study sheds some helpful light on how social scientists are moving towards open by asking them how far they have ever used key selected practices. Survey research by Christensen et al. (2020) on U.S. social scientists who are published authors asked them to say when they first did something 'open' at least once. (Where that occurred in earlier years, before (say) 2019, then the responding authors' replies were obviously recalled year-dates.) Researchers were not asked how often or extensively they engaged in these OS practices, or even if they were still doing so. Thus, the responses may considerably overstate the extent of adoption within the discipline group at any one time. Nonetheless, the authors found some encouraging trends, with four fifths of U.S. social scientists having done something open by 2017. In particular, publishing at least something open access was near universal by 2018. Other, more demanding types of activity showed some marked differences, however (see Figure 2.4). Posting data online was the most prevalent, followed by publishing survey instruments (i.e. questionnaires, coding processes and sampling frames). Only a fifth of social scientists had pre-registered hypotheses in 2017 (Christensen et al. 2020).

Using the same criteria, Christensen et al. also found some noteworthy differences between three types of researchers – those using experimental methods; those using other quantitative techniques; and solely qualitative researchers (see Figure 2.5). Depositing data at least once was common in experimental fields, and a bit less so in quantitative work. There were sharper differences in posting the research instruments from studies online. And pre-registering hypotheses at that time (2017) had yet to take off much in non-experimental work (Christensen et al. 2020). The results also unequivocally showed substantial differences between authors in experimental fields (who came out top in their use of open practices), those using non-experimental quantitative work (who were in the middle) and those in theoretical or qualitative fields (whose use of open approaches was lagging).

In addition, Christensen et al. asked responding authors to give their perceptions of how much progress open approaches were making in their field, versus how much they had used open working themselves. (This approach generates interesting results in many social science fields. Exponents argue that it gives a better picture of what respondents feel. For example, asking people how they think voters in general will cast their ballots has often been more accurate than polls based on how people themselves say they will vote.) Academics' estimates of progress in their field as a whole were always considerably more pessimistic than what the same authors said about their own practices – especially about posting data or code online in accessible ways. Apparently, it was other academics who were more conservative (= hesitant

Figure 2.4: The years in which U.S. social scientists reported to Christenson et al. that they first adopted three kinds of open science practices

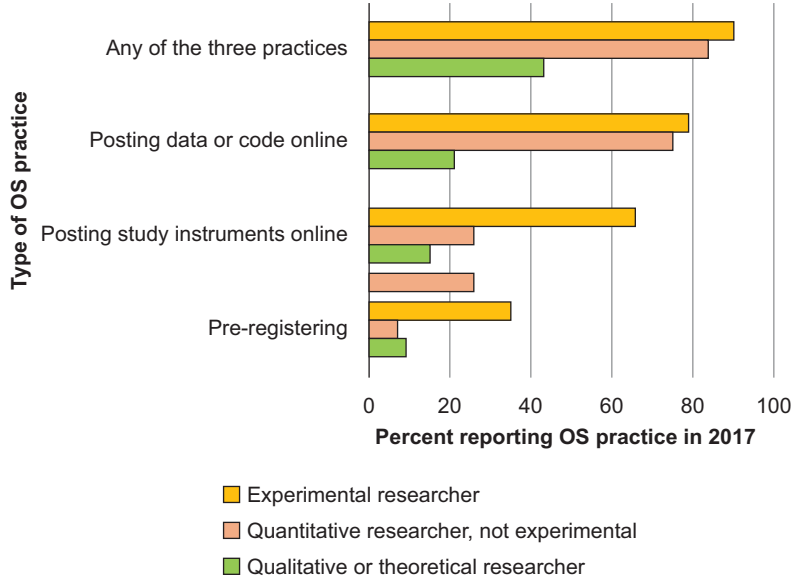


Note: The researchers asked respondents:
'Approximately when was the first time you publicly posted data or code online?'
'Approximately when was the first time you publicly posted study instruments online?' and
'Approximately when was the first time you pre-registered hypotheses or analyses in advance of a study?'
The estimated adoption rate for the entire sample contacted, including non-respondents, was somewhat lower than the lines here for respondents only, calculated using methods explained in the article.
The sample was restricted to published authors in American social science (defined in the paper) who had completed their PhDs by 2009 (*N* = 637).
Source: Computed from data in Christensen et al. (2020).

or 'laggard') than respondents themselves. The respondents' field perceptions also seemed to be more pessimistic than the actual behaviours in their fields estimated by Christensen et al., using data garnered from the sample as a whole.

Of course, it is useful to get social scientists to consider adopting concrete practices like those that STEMM academics have already found useful and developed protocols for. In later chapters (Parts II) we discuss a wide range of approaches and initiatives already trialled and their application across the social sciences. Yet it is also important to situate specific methods steps within an overall approach to the discipline group. We need to recognize that the OSS agenda is a far wider and deeper one than

Figure 2.5: The varying use of 'open' actions across three types of researchers in the Christensen research (2017)



Note: See Figure 2.4.
Source: Compiled from data in Christensen et al. (2020).

the surface indicators alone might suggest. Changing researchers' culture towards embracing openness is the key first step. Only once this is accomplished will further developments emerge at the research coalface, and suggest the most relevant practices to consider in each field (some of which we cannot yet discern).

Associating 'open' solely with integrity issues and surveillance

A second false start, and a follow-on implication of focusing down too early on this or that isolated practice, has been to misrepresent the cultural shift to open as being about yet more external surveillance of what researchers are doing. All universities, most academic disciplines and all grant funders already have very well-developed requirements governing research integrity and research ethics. Yet, if enunciated by these bodies in a surveillance and regulatory mode, open ways of working can easily come across as just a new set of burdensome and extraneous bureaucratic hoops that researchers must jump through in order to secure funding, or pass muster in their PhD oral exam, or ace the assessment of their post-doc research. For instance, a 'research briefing' by well-intentioned library staff or research grant advisors can all too easily end up as a lengthy set of PowerPoint slides, each detailing the varying open requirements of this or

that funding body, or this or that set of data management principles. Over a third (35%) of STEMM academics are full-time researchers with no teaching, and thus may be able to specialize in grant-seeking and meeting onerous funding or regulatory requirements. But only 11% (one in nine) social scientists are research-only (Bastow, Dunleavy and Tinkler 2014, p. 8). Since busy researchers are always multitasking already, and most social scientists in particular have heavy teaching and administration loads to cope with, more time-consuming bureaucracy is doubly unwelcome.

Many university staff have also opposed or critiqued contemporary trends towards the greater use of metrics in official research assessments (like the UK's crude and time-consuming Research Excellence Framework (or REF system). Critics have been understandably vocal about further increases in (allegedly 'neo-liberal') surveillance techniques that have the effect of indexing academic 'productivity' or 'impacts' in quantitative terms. Some of this commentary has arguably been overstated and many critiques have simply ignored both the numerous autonomous sources of the impact agenda within academia itself, and the advances in knowledge and scientific practices that a better understanding of impact makes feasible (Dunleavy and Tinkler 2020).

Yet the reaction against excessive surveillance or the bureaucratization of research is nonetheless widespread and deeply felt in large parts of the social sciences. Even in STEMM some radical critics have extended the critique of metricization and 'surveillance' to apply also to OS. For instance, Philip Mirowksi (2018, p. 171) argues that: 'The [OS] agenda is effectively to re-engineer science along the lines of platform capitalism, under the misleading banner of opening up science to the masses'. These problems are made worse by justifications of OSS that ground it primarily or even solely in terms of counteracting a wide range of abuses or dubious practices designed to push papers into high status journals – such as:

- *p*-hacking (massaging model regression analyses to get them over particular 'statistical significance' levels);
- 'harking' (presenting a post hoc hypothesis derived only *after* data analysis as if it was an a priori one that has been subjected to 'falsification' testing);
- journals' repeated emphasis on publishing only positive results, while negative results (that might be countervailing) go unpublished; and
- non-reproducibility – 'discoveries' reported by one researcher that cannot subsequently be replicated in other studies.

See Chapter 3 for a detailed discussion.

All these problems are certainly live issues across all the social sciences, although they have been discussed most extensively in psychology and economics as a 'replication crisis'. OSS approaches can mitigate many such issues and help develop higher quality research, as Section 2.3 argues. Yet adding

to the safeguards for research integrity is not all or even primarily what OSS is about.

Nor is enhancing research integrity necessarily even an essential component of the OSS approach, since open practices can themselves be 'played' by unethical or 'black hat' authors. And unlike for many of the STEMM disciplines, where hypothetico-deductive testing is normally appropriate, social scientists cannot assume that they are studying an invariant entity ('nature') with unchanging, law-like (if very complex) mechanisms at work. Instead, they must deal with a constantly and inherently changing social world. All human actors have reflexive individual and societal capabilities to absorb new knowledge played back to them, and then to adapt how they do things as a result (see Chapter 3).

OSS is a push to resocialize researchers. It aims to address integrity problems more successfully and reliably than ever before, but by changing researchers' norms, culture and scholarly/scientific processes in fundamental ways – not by playing on or increasing their fears of discovery or punishment for research misbehaviours. Instead, the central OSS effort is to achieve better research from the earliest beginnings of research ideas and projects through to their final publication and discussion by both professional and wider audiences.

Privileging quantitative work in ways that may reawaken methodological sectarianism

Perhaps the greatest danger of lifting OS approaches unthinkingly from STEMM practices is that it could reopen some of the acute 'methodology wars' that previously occurred in many social science disciplines, and that are still smouldering to some extent. On one side of these conflicts were the exponents of formal (mathematicized) theories and using quantitative and experimental research approaches. In the late 1990s they often framed ambitious agendas for formal approaches to completely dominate disciplines – apparently seeking for social science research to become as technical and inaccessible for outsiders as the knowledge in many STEMM disciplines. In epistemological terms they often also seemed to assert the straightforward primacy (perhaps even the universal desirability) of hypothetico-deductive methods in social science (see Chapter 4). On the other side were social scientists using more qualitative methods, focusing more on social meanings and the close interpretation of social discourses and behaviours, and others seeking to broaden the application, impacts of and access to their work so as to include citizens, enterprises, policy-makers and others (Dunleavy and Tinkler 2021). They additionally adopted epistemological positions denying 'scientistic evidentialism' in their approach to research and emphasizing a range of alternative exploratory or interpretivist positions (Kvale 2008). In addition, across all the social sciences a great deal

of vitally important scholarship is devoted to 'sense making', that is, creating descriptive or 'first draft of history' accounts of new and often hard to understand societal and economic changes. Such efforts necessarily use less formal analysis, more inductive methods and perhaps more intuitively theorized approaches.

In recent years the methods wars conflicts have mostly declined from their peak intensity. For instance, in political science the editors of a top journal recently welcomed robust qualitative work (APSR Editors 2022), signalling a change from the journal's hardline stance at the time of the 'perestroika' conflicts at APSA's 2001 conference in San Francisco (Laitin 2003). However, pushing a premature, STEMM-copying notion of open social science could all too easily reopen such past wounds. It might well reawaken alarm amongst qualitative scholars that their less easily 'opened up' approaches could be squeezed out in funding or prestige/reputational terms. This disruptive potential was especially true where open changes were pushed by open advocates as solving an alleged replication crisis in psychology (Pennington 2023; Fanelli 2018) or other disciplines; or where open became associated with attempts to rewrite research integrity rules in crude ways that are in fact only practicable in (some kinds of) quantitative work; or where (as in health studies and medicine) researchers handle sensitive and private data linked to specific people that it would be unethical or impossible to make reaccessible (Lamb et al. 2024).

In their book, *Transparent and Reproducible Social Science Research: How to Do Open Science*, three authors from the Berkeley Initiative for Transparency in the Social Sciences (BITSS), Garret Christensen, Jeremy Freese and Edward Miguel (2019), unfortunately seem to fall into this trap. The book claims to be:

> The first book to summarize and synthesize new approaches to combat false positives and non-reproducible findings in social science research, document the underlying problems in research practices, and teach a new generation of students and scholars how to overcome them.

Yet in practice this volume only covers the disciplines and areas of social science with the most developed quantitative methods, and it only stresses the replication/research integrity arguments linked to quantitative methods. No effort is made anywhere in the volume to consider how qualitative methods and approaches might be involved in movements towards open science, or how open ways of working might be developed in non-quantitative social science work. Indeed, remarkably enough, the book's nearly 200 pages manage never to mention the word 'qualitative' at all, effectively blanking claims to scientific status for qualitative methods or research areas completely. The authors' implied position almost seems to be that qualitative research has no problems with non-reproducibility or false positives because these approaches generate

no valid or worthwhile scientific 'findings' in the first place. Searching the institute's website also revealed (in early 2023) only a few scattered examples of an effort to address qualitative research as OSS.

Other U.S.-based promoters of OS, like the Center for Open Science (COS), have also covered mainly quantitative social sciences disciplines (like psychology and health studies), although the Board of COS has included a lot of social scientists amongst its members. The organization's somewhat disturbing 'stages' strategy for accelerating cultural change towards open practices starts with 'Make it [openness] Easy', goes through 'Make it Normative [*sic*]' and then 'Make it Rewarding', but ends up with 'Make it [openness] Required' (Center for Open Science 2019). No apparent explanation is given of how that last, mandatory stage would work outside of a few fields (i.e. laboratory or experimental work, randomized control trials, or purely computational research).

So, without a strong rationale adapted to the circumstances of social sciences across the discipline group as a whole, importing OS components from STEMM disciplines in a random 'add-on' way is risky. It can easily appear as just more external 'bureaucratic' demands that serve only to generate increased workloads for researchers. Alternatively, OS innovations may appear as just protections against (relatively rare) professional malpractices, already covered by extensive guidelines, rules and norms on research ethics and professional methodology. Sadly, too, simply 'bigging up' OS requirements that cannot in practice be met in many social science studies can seem to be part of a drive to marginalize qualitative research, or even to drive it out of a discipline altogether, in favour of quantitative studies.

Even after only looking within the STEMM disciplines themselves, Leonelli (2023, p. 5) argued that:

> Many of the more institutionalised OS initiatives tend to privilege a homogenous, universally applicable understanding of the scientific method over a pluralistic and situated one. It is much easier to set up OS guidelines when assuming that science consists of a coherent body of knowledge and procedures that can and should conform to common norms – an assumption that flies in the face of the rampant plurality of research approaches used across domains, locations and contexts, and the significance of such plurality in delivering a robust, comprehensive and reliable understanding of the world.

Little wonder then that up to now many social science researchers appear to have resisted demands for OS changes, with OSS progress consequently lagging when compared with STEMM disciplines. This is especially notable for open access publishing models, where authors or universities pay an article processing charge (APC) rather than relying on subscriptions. Compared with STEMM disciplines, their implementation in the social sciences and humanities took many more years because most of their learned societies were heavily dependent for funding on earning high subscription income from

their own journals. And with more 'lone scholar' or unfunded research, these professional bodies long doubted if they could replicate their past incomes, or even cover their costs, if all their income came from APCs (see Chapter 15). Accordingly, they lobbied governments and funders hard for the longest possible transition timetables in their fields (Tennant et al. 2016). In mostly the same disciplines, open access arrangements for publishing books also lagged a decade or more behind those for articles. Some heavily book-dependent social science and humanities disciplines (like mainstream history) also held out much longer for part-funding research scholarship via commercial sales of books. Despite these difficulties, and across all disciplines, wherever open access publishing made progress – in leaps and bounds by the 2020s – it proved an important entry point for far more diverse OS ideas to penetrate and challenge 'legacy' closed academic approaches.

Some valuable insights about different OS approaches can be gleaned from the STEMM disciplines, as we show in Chapter 3. Yet an even broader and more inclusive approach is needed if openness is to develop fully across the social sciences. This process will also hopefully reshape research cultures in crossover disciplines with the humanities (like law and the digital humanities), and spill over into a wider range of humanities subjects (especially mainstream history, philosophy and literature studies).

OSS is not about science envy or putting in place more institutional surveillance of research within academia. Nor can it ever risk becoming a new episode in destructive and misguided methodology wars, still less being weaponized as a way of driving out more qualitative and theoretical scholars from the discipline group. As Lamb et al. (2024, p. 662) argued for health studies:

> Rigid requirements for qualitative researchers to make full data sets publicly available are inappropriate, but … there are some convincing reasons to consider whether qualitative data could be made available for secondary analysis (given appropriate consent and vetting procedures).

So OSS must instead be a positive agenda about moving forward all social science research (however it is conducted) towards being more soundly based, more coherent and more open for other scholars and citizens to access and understand, in ways spelt out in the next section.

2.3 Defining 'open social science'

Yet broadening a push for openness beyond its STEMM origins should not mean diluting the term 'open' to mean 'almost anything somewhat more transparent'. As OS has been taken up by the governments and philanthropic funding organizations supporting academic research, so some efforts to make

the term more widely applicable have only blurred the concepts involved. For instance, the Wikipedia entry for 'open science' adopts an over-inclusive approach, listing dozens of different levels or areas of application of open concepts, many of them very specific and devoid of any strong intellectual rationale (Wikipedia 2023). With such an eclectic ('all things to everyone') approach to open, it is difficult to identify any central agenda concepts, still less priorities for change.

Recently a more 'official' elite consensus has emerged in some countries that maximizing 'reproducible' research is the central thrust of OS. Yet this has tended to go along with a blurring or wearing down of what that term even means, as demonstrated in a 2023 report by a UK House of Commons committee. After endorsing a focus on the reproducibility of research their report went on to quote approvingly from evidence submitted to them by Kings College London, who argued that:

> For many subjects, outside of STEM, particularly within Arts & Humanities [*sic*], the concept of reproducibility is not applicable to the research methods employed and terms such as *transparency* are more appropriate to discuss the rigour of the research findings. If we approach reproducibility through broader definitions, we can engage with a broader range of researchers to better understand the scale of the issue in wider disciplinary areas (Kings College 2022). [*Our italics*.]

The committee of MPs then concluded that:

> Neat definitions should not be a focus of our Report … In line with the evidence we heard, we use the term reproducibility to broadly refer to the transparency and quality of research—i.e. have the researchers been sufficiently open to allow for robust assessment of their work's data, methodology, and conclusions, and, through such an assessment, is their research proven to be thorough and accurate? (SITC 2023, p. 8).

Encouraging though the effort to take a broader perspective may seem here, it still over-stressed a research integrity perspective. And the Kings College line of diluting reproducibility to just mean any and all 'transparency' efforts risks running into the sands of extreme vagueness.

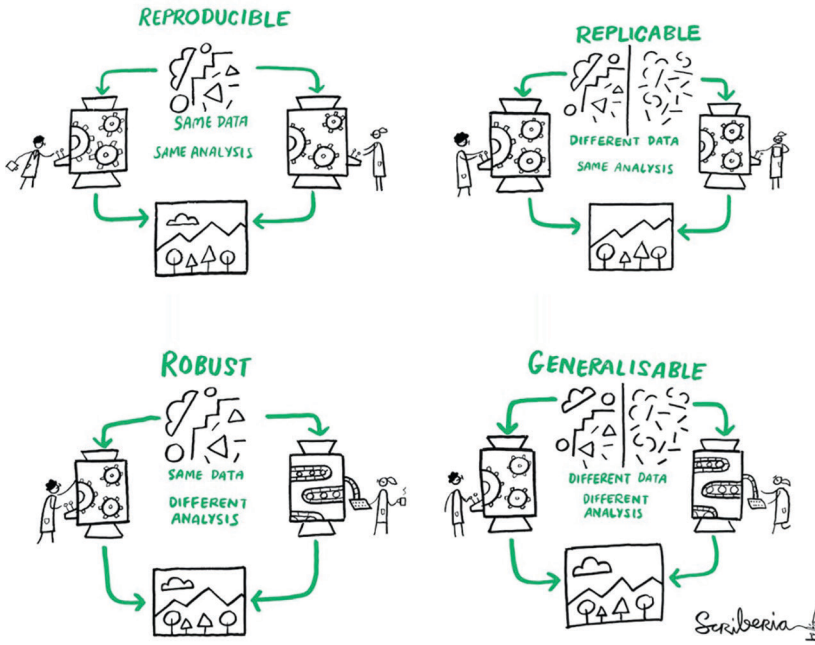
In addition, much of the past debate around the push to open was undertaken with rather confused vocabularies. A great deal of existing discussion does not distinguish well between making scientific research 'replicable' (a pretty tough criterion to meet, as we discuss below) and making it 'reproducible' (a more modest and achievable aim). To make matters worse, careful tracking of the two concepts suggests around 50 different meanings for the two 'r' labels across different concepts (Plesser 2018).

More recently, however, the influential Turing Way project (originating in computer-based research and coding) has defined OS in a more inspirational and precise way, covering a broad range of measures to boost the reproducibility and integration of scientific work. They define 'reproducible research as work that can be independently recreated from the same data and the same code that the original team used'. Being reproducible is distinct from being replicable, robust or generalizable in ways that are set out in Figures 2.6 and 2.7.

Figure 2.6: How the Turing Way project defines reproducible research

		Data used in the two studies is	
		the same	**different**
Analysis used in two studies is	**the same**	the same results = reproducible	the same results = replicable
	different	the same results = robust	the same results = generalizable

Figure 2.7: Pictograph of Turing Way's four types of 'open science' research



Source: Illustration by Scriberia at https://doi.org/10.5281/zenodo.3332807. Used under a CC-BY 4.0 licence. The component pictographs are slightly respaced here.

The different dimensions of reproducible research described in the matrix above have the following definitions:

- '*Reproducible:* A result is reproducible when the *same* [*research*] analysis steps performed on the *same* dataset consistently produces the *same* answer.
- *Replicable:* A result is replicable when the *same* analysis performed on *different* datasets produces quantitatively or qualitatively similar answers.
- *Robustness*: A result is robust when the *same* dataset is subjected to *different* analysis workflows to answer the same research question (for example, using a slightly different set of statistical tests or technical implementation) and a qualitatively similar or identical answer is produced. Robust results show that the results from the work are not dependent on the exact specificities of a statistical test [*or more broadly, of the analysis method*] used.
- *Generalizable:* Combining replicable and robust findings allows us to form generalisable results. Note that running an analysis on a different software implementation and with a different dataset does not provide *generalized* results. There will be many more steps to know how well the work applies to all the different aspects of the research question. Generalization is an important step towards understanding that the result is not dependent on a particular dataset nor a particular version of the analysis pipeline' (Turing Way project 2023). [*Our inserts*]

We set out a viable set of OSS objectives following the central impetus of the Turing Way's approach and that can apply across the discipline group as a whole. Operationalizing this vision is the central purpose of this volume and resists easy summary at this initial stage. However, we have derived a framework of eight substantive and hopefully not too vague ways in which this approach can be advanced.

1. *Ensuring that research results and analyses are published and communicated in open access ways*. Full text analyses should be maximally accessible to other researchers, and able to be communicated freely to other potential audiences using modern, digital and accessible means (e.g. via blogs, social media, podcasts, video talks). Wherever feasible, research results and analyses should also be published openly, making datasets, codings and other information and evidence maximally accessible to other researchers, including those in nearby disciplines. In addition, it is helpful if research results (and not just published works) are also communicated to other potential audiences using modern, digital and accessible social media. We noted above that

all modern social science work deals with 'human-dominated systems' or 'human-influenced systems' that are fast-changing, complex and also involve STEMM and other disciplines. In this respect, the 'ordinary knowledge' of the people who are research subjects, participants and practitioners, potentially engaged through open publication, has considerable value for contextualizing and improving scholarship, as Lindblom and Cohen (1979) famously argued.

Effective dissemination also ensures the widest possible academic scrutiny of new work, thus improving fact checking, problem-solving, research accuracy and the emergence of diverse scholarly voices. The growth and improvement of 'citizen social science' also rely heavily on maximizing open access to scholarship (Tauginienė et al. 2020 and Chapter 13). Finally, researchers' experiences of using open access publishing and impact-conscious dissemination has often been their first taste of how open approaches work (see Figures 2.4 and 2.5), and can encourage them to move further into using other OS approaches.

2. *Ensuring that quantitative research data and analyses are accessible and* ***reproducible*** – which means applying the FAIR principles and other approaches to data stewardship (see Chapter 6). All the data supporting published quantitative research articles and books, plus details of their definitions, coding and analyses, should be straightforward to obtain and be formatted accessibly – so that any of them can be easily checked if questions arise. In addition, researchers should take steps to encourage the wide reuse of their data by other analysts, including those conducting different kinds of studies. Most public and philanthropic grant funders now require research data to be deposited in open archives (with well-explained documentation and data in accessible formats) once projects are completed. However, many corporate- or university-funded research datasets are not yet made available at all to outsiders. The growing importance of using proprietary social media data in modern research has further exacerbated this problem.

   In other areas, the trend towards multinational and multi-team research efforts in social science has fostered the development of collective, reusable and comparable datasets covering multiple countries, or datasets that include pre-defined common questions that can be 'pooled' into more inclusive or authoritative resources (such as the World Values Survey 2024). Achieving more improvements may entail new research taking better account of the best/strongest available research in a field (e.g. by using systematic reviews). Developing a culture of open data along these lines would both promote and enable less silo-bound disciplinary approaches to defining questions and variables in quantitative social science.

3. *Facilitating the **reuse** of quantitative data by other analysts* and the development of reusable datasets relevant to multiple studies, or datasets that can be pooled into more inclusive or authoritative resources. The research results achieved by one team should be reproducible by other researchers using the same datasets and analysis methods (see Figure 2.7). The incorporation of data management standards into universities' own research regulation has extended this approach to cover smaller and self-funded studies. Encouraging reuse of data is a wide imperative, one that goes far beyond the alleged reproducibility crisis – too often misleadingly termed this – in some disciplines, such as psychology (Fanelli 2018). Many quantitative journals also now require a replication archive (RA) as a condition for publishing quantitative articles. Academic journals that have appointed specific Data Editors to check the quality, scope and useability of RAs have especially seemed to advance a culture of openness.

   Yet improvements in reproducing results are hard to achieve, even in the experimental STEMM disciplines. Transferring methods between an original team and other colleagues checking for reproducibility entails overcoming multiple tacit or implicit knowledge constraints at many stages of a project (Collins 2010). At a micro-level, progress may rely on expanding the recording and documentation of information, data and analyses undertaken across multiple experiments, models, data runs, fieldwork expeditions or archive visits. More systematic use of journaling, lab notebooks, and documentation of searches, coding analyses, etc., are also potentially important here (see Chapter 7).
4. *Developing the **reproducibility** of **qualitative** evidence, information and data,* for instance, data drawn from interviews, archives, documentation, text sources, images and so on, is in its early days – as is the reuse of qualitative research materials. Non-quantitative researchers in many social science disciplines have undeniably made less progress in creatively adopting OS practices. There are more logistical issues in making more widely available for scrutiny or reuse knowledge, information and data drawn from methods that we discuss in Part III – especially interviews, archives, documentation, text sources, images and so on (Jamieson and Lewthwaite 2019). Common barriers include information being made available only 'off the record' or 'non-attributably'; difficulties in preserving the anonymity of research interviewees, patients, or the groups affected by the problems being analysed; restrictions imposed by owners or museums on the reuse of images or scans of documents; and extensive documentary sources not being available in digital forms (Jamieson and Lewthwaite 2019). So, we are still in the foothills of exploring what OS may mean in most qualitative research areas, and the scope of feasible changes there remains contested. Yet the challenge is to find

new ways of undertaking research, and new steps that can mitigate the impact of current barriers (Tarrant and Hughes 2020; and see Chapters 9–12 below).

5. *Enhancing (wherever feasible) the **replicability** of social science research*, so that results claimed can be reproduced by other researchers using the same or similar analysis methods on different datasets or bodies of evidence. At a macro-level this entails new research taking better account of the best/strongest available research in a field (e.g. by using systematic reviews) and not being silo-bound by a discipline or theory/ideological viewpoint. At a micro-level it often relies on improving the level of documentation of information, data and analyses undertaken across multiple experiments, models, data runs, fieldwork expeditions or archive visits – for instance, via more systematic use of journaling, lab notebooks and documentation of searches, analyses, etc.

   As the Turing Way diagram (Figure 2.7) illustrates graphically, it is also vitally important for the integrity and prestige of all the social sciences that research analyses and results found in one dataset or body of evidence by one team can then be replicated by other researchers using the same or similar treatment or methods on *different* data sources. Although no invariant 'natural world' lies behind the complexities of human and social behaviours, similar situations and datasets should not produce radically divergent findings when analysed using the same methods – for example, just because the range of countries, dates, or data included have changed. Improvements in replicability are especially difficult, both in STEMM discipline research fields like ecology and in the non-experimental social sciences. In both contexts, complex background conditions may be shaped by hundreds of causal influences, all operating at once and changing constantly.

6. *Seeking to enhance the **robustness** of social science research*. When different teams of social scientists analyse the same datasets or bodies of evidence, using different analysis methods, hypotheses and theories, they should nonetheless generate consistent results (see the bottom left-hand quadrant in Figures 2.6 and 2.7). Greater 'triangulation' of different research approaches focusing on the same data and phenomena can generate stronger levels of assurance that the results in any given study are not artefactual.

   In some areas of the discipline group methods for pooling and sifting published evidence have been studied intensively (as with central bank economists seeking to incorporate the widest range of trends data into their modelling of the economy as a whole). Yet divergent overall judgements often still result – testifying to the difficulty of future predictions in both sense-making and causal analysis when studying human behaviour. Improving the robustness of economic and social research is key in combatting the widespread media

assumption that STEMM researchers 'find' results and 'discover' evidence, while social scientists can only 'argue for' their 'interpretation' of results (Evans 1995, p. 173; Bastow, Dunleavy and Tinkler 2014, Chapter 7).

7. *Enhancing (wherever feasible) the **generalizability** of social science research results*, so that their feasible application across social situations can be broadened from a given place, time, or social setting. Ideally, the research findings gathered by different teams using different analyses, datasets or types of evidence should add up to a consistent and coherent picture of what is going on. Cumulating social science knowledge more, and in more widely agreed ways, and comparing the results and core implications of different perspectives, should make social science theories and applications more applicable across a greater range of social settings and time periods. It would also allow research findings to be calibrated and adapted with some confidence to better predict future outcomes, adapting knowledge effectively (as engineers must do when building bridges in different settings). Achieving steps forward here could boost the confidence of both researchers and practitioners/citizens that the theories being applied to situations have been correctly framed; that research evidence provides reliable bases for action, organizational decisions, or policy-making; and that some expert consensus exists.
8. *Moving towards more citizen social science (CSS)* is a final, tricky step. The core principle should be that research results about given groups, individuals, organizations and communities, and the evidence collected from these actors, should be communicated back to them wherever possible (see Chapter 13). Any push for openness implies making this happen far more often than has previously occurred, even in the recent past. The reactions and understandings of people being researched should be sought, assessed and incorporated into research processes and the development of knowledge, wherever it is feasible to do so and as early as possible. Taking citizen, user or practitioner feedback seriously can also be more generally practised in many less formalized ways that nonetheless can refine and strengthen researchers' understandings, especially in applied sub-fields.

More ambitious concepts move towards 'co-producing research' with its subjects – creating opportunities to keep participants engaged throughout the research process, rather than results only emerging with formal publication, years after fieldwork or data collection were conducted. The fullest co-production is especially developed in formal processes, where:

- research informs participatory processes and public decisions (as with deliberative democracy);

- citizens and networks undertake some tasks themselves that feed into research; or
- researchers and communities co-develop and co-steer how outputs are written up and reported.

These eight suggestions are developed further below.

Conclusions

The STEMM disciplines' greater prestige (reflected in massively more funding and research effort) than the social sciences has occurred partly because the 'natural' sciences developed a 'high consensus/rapid discovery' approach to research. Professional prestige could only be gained by making new findings and inside the research frontier high ('normal science') levels of consensus were maintained on the prevalent scientific paradigm. By contrast, social science research has in the past tended to have marked theoretical/ideological dissensus on what counts as fundamentals, and a relatively high turnover of evanescent findings or 'fashions' of analysis that lowered its prestige (Coleman 1990, pp. 230–237).

The push for OSS re-emphasizes that most prestige should be assigned to *reliable* advances in forefront knowledge about human and societal behaviours. And as research methods and data access both improve, we may plausibly expect to see a somewhat greater (but still likely moderate) level of social scientific consensus on fundamentals and the understanding of core processes. In addition, the kind of interdisciplinary or trans-disciplinary work implied by the frameworks of HDS and HIS, discussed above, is far more likely to develop effectively if both STEMM and the social sciences converge on the same or similar OS/OSS modes of working.

So far, despite some optimistic recent findings, well-founded OS practices have made a relatively modest impression in many social science disciplines. Perhaps this reflects OS's overly 'physical' or STEMM roots, where a lot of discussion of OS has seemed to focus only on quantitative lab, experimental and computational work. Apparently excluded or 'on sufferance' (as we show in the next chapter) have even been other STEMM fields like ecology, where precise controls of background conditions for fieldwork are unfeasible. This pattern repeats in the social sciences, where some early handbooks and pioneer units in U.S. universities promoting OSS ideas have focused solely on quantitative data analysis research, construed in narrow ways.

Rebooting OSS entails explaining openness in ways that apply across the whole discipline group, that encourage innovative agendas and that appeal to researchers themselves (not just oversight officials or funders) – as this box shows.

The eight components of open social science

1. Ensuring that research results and analyses are published and communicated in open access ways.
2. Ensuring that *quantitative* research data and analyses are accessible and reproducible.
3. Facilitating the reuse of quantitative data by other analysts.
4. Developing ways to enhance the reproducibility and reusability of *qualitative* evidence.
5. Enhancing (wherever feasible) the replicability of social science research.
6. Seeking to enhance the robustness of social science research.
7. Enhancing (wherever feasible) the generalizability of social science research results.
8. Moving towards more citizen social science.

As we envisage it, OSS is a broad cultural movement that is neither sectarian on the one hand, nor diluted by using vague or vacuous language on the other. Instead, following a specific framework derived from the Turing project, the box above sketches an eight-part programme of specific changes – that are more fully developed in Parts II and III, which make up the practical heart of this book.

Some deeper and more general theoretical or 'philosophical' issues around academic and scientific research have become bound up with the open agenda, and so it will be helpful to cover them in the rest of Part I. In the next chapter we consider in more detail the diversity of OS practices across different STEMM disciplines, showing that some not-so-visible debates around the concept have occurred there that hold some key lessons for open social science.

Note

* This chapter draws on research and seminars forming part of the CIVICA Research programme, which was funded in 2022–24 by the European Union's research and innovation programme Horizon 2020, under the chapter Science in Society.

References

$ denotes a paywall publication. All other sources are open access.

APSR Editors (2022) 'Publishing your qualitative manuscript in the APSR', American Political Science, review webpage, 3 March. https://perma.cc/4B8Z-6C9C

Bastow, Simon; Dunleavy, Patrick; and Tinkler, Jane (2014) *The Impact of the Social Sciences: How Academics and their Research Make a Difference*. London: Sage. $ https://doi.org/10.4135/9781473921511

Besley, Tim and Hennessy, Peter (2009) 'The global financial crisis – why didn't anybody notice?' *British Academy Review*, no. 14. https://perma.cc/YDW9-R9JX

Caws, Peter (2005) 'First and second order unification in the social and human sciences', *Graduate Journal of Social Science*, vol. 2, no. 2, pp. 151–169. https://perma.cc/8FUL-8UT4

Center for Open Science (2019) 'Strategy for culture change'. https://perma.cc/8W62-52E6

Christakis, Nicholas A. (2013) 'Let's shake up the social sciences', *New York Times*, 19 July. $ https://www.nytimes.com/2013/07/21/opinion/sunday/lets-shake-up-the-social-sciences.html

Christensen, Garret; Freese, Jeremy; and Miguel, Edward (2019) *Transparent and Reproducible Social Science Research: How to Do Open Science*, San Francisco: University of California Press. $ https://perma.cc/WY8C-UZDZ

Christensen, Garret; Wang, Zenan; Paluck, Elizabeth Levy; Swanson, Nicholas; Birke, David et al. (2020) 'Open Science Practices Are on the Rise: The State of Social Science (3S) Survey', UC Berkeley: CEGA Working Papers. 13 January. https://perma.cc/SFM4-6SEU

Clarke, Kevin A and Primo, David M. (2012) 'Overcoming "physics envy"', *New York Times*, 30 March. $ https://www.nytimes.com/2012/04/01/opinion/sunday/the-social-sciences-physics-envy.html

Coleman, James (1990) *Foundations of Social Theory*, Cambridge, MA: Harvard/Belknap Press. $ https://scholarship.law.upenn.edu/faculty_scholarship/1734

Collins, Harry. (2010) *Tacit and Explicit Knowledge*, Chicago, IL: Chicago University Press. $ https://press.uchicago.edu/ucp/books/book/chicago/T/bo8461024.html

Collins, Harry and Evans, Robert (2009) *Rethinking Expertise*, Chicago, IL: University of Chicago Press. $ https://perma.cc/A9DE-9QGR

Collins, Randall (1994) 'Why the Social Sciences Won't Become High-Consensus, Rapid-Discovery Science', *Sociological Forum*, vol. 9, no. 2, pp. 155–177. $ http://www.jstor.org/stable/685040

Collins, Randall. (2000) *The Sociology of Philosophies*, Boston: Harvard University Press/Belknap Press. $ https://perma.cc/7R6Q-JSQX

Davies, Will (2014) 'Call to arms for shaking up social sciences relies on false premise that science alone can solve all social problems', *LSE Impact blog*, 14 October. https://perma.cc/2RTR-EU6Y

Dunleavy, Patrick (2014). 'The social science of human-dominated and human influenced systems', Academy of Social Sciences Annual Lecture, London, 3 July. Slide 11. http://dx.doi.org/10.13140/RG.2.2.20523.66084

Dunleavy, Patrick and Tinkler, Jane (2021) *Maximizing the Impacts of Academic Research: How to Grow the Recognition, Influence, Practical Application and Public Understanding of Science and Scholarship*, London: Palgrave (now Bloomsbury Press). **$** https://www.bloomsbury.com/uk/maximizing-the-impacts-of-academic-research-9781352010992/

Dunleavy, Patrick (2014) 'The social science of human-dominated and human-influenced systems', Academy of the Social Sciences Annual Lecture, London, 3 July. Slide 11. https://www.researchgate.net/publication/383422890_The_social_science_of_human-dominated_and_human-influenced_systems Also available on YouTube: https://www.youtube.com/watch?v=nY4ra5QWJeE

Dunleavy, Patrick and Tinkler, Jane (2020) 'The impacts agenda is an autonomous push for opening up and democratizing academia, not part of a neo-liberal hegemony', *LSE Impacts blog*, 20 November. https://perma.cc/W5QM-A38R

Dunleavy, Patrick; Bastow, Simon; and Tinkler, Jane (2014) 'The contemporary social sciences are now converging strongly with STEM disciplines in the study of "human-dominated systems" and "human-influenced systems" ', *LSE Impacts blog*, 20 January. https://perma.cc/Q78Z-C2XS. Also on *Sage Social Science Space* blog, 22 January. https://perma.cc/MNH5-3ZSM

Elhacham, Emily; Ben-Uri, Liad; Grozovski, Jonathan; Bar-On, Yinon M; and Milo, Ron (2020) 'Global human-made mass exceeds all living biomass', *Nature*, vol. 588, p. 444. **$** https://doi.org/10.1038/s41586-020-3010-5

Evans, William (1995) 'The mundane and the arcane: prestige media coverage of social and natural science', *Journalism & Mass Communication Quarterly*, vol. 72, no. 1, pp. 168–177. **$** https://doi.org/10.1177/107769909507200114

Fanelli, Danielle (2018) 'Is science really facing a reproducibility crisis, and do we need it to?', *PNAS*, vol. 115, no. 11, pp. 2628–2632. https://doi.org/10.1073/pnas.1708272114

Gibbard, Philip; Walker, Michael; Bauer, Andrew; Edgeworth, Matthew; et al. (2022) 'The Anthropocene as an Event, not an Epoch', *Journal of Quaternary Science*, vol. 37, no. 3, pp. 395–399. **$** https://doi.org/10.1002/jqs.3416 But OA at https://perma.cc/3PSL-T3BP

Goodall, Amanda and Oswald, Andrew (2014) 'Do the social sciences need a shake-up?', *Times Higher Education*, 9 October. $ https://www.timeshighereducation.com/features/do-the-social-sciences-need-a-shake-up/2016165.article

Hartmann, Stephan and Sprenger, Jan (2010) 'Mathematics and statistics in the social sciences', in Jarvie, Ian and Bonilla, Jesus Zamora (eds) *The SAGE Handbook of the Philosophy of Science*, London: Sage. OA at: https://perma.cc/9GAH-4YDB

Ionnidis, John (2005) 'Why most published research findings are false', *PLOS Medicine*, vol. 2, no. 8, article: e124. OA at: https://perma.cc/4LT6-GWMW

Jamieson, Lyn and Lewthwaite, Sarah (2019) 'Big Qual – Why we should be thinking big about qualitative data for research, teaching and policy', *LSE Impacts* blog. https://perma.cc/GM2P-59EK

Kings College (2022) 'Written Evidence Submitted by King's College, London' (RRE0091), Evidence submission to the House of Commons. https://committees.parliament.uk/writtenevidence/39936/html/

Kvale, Steinar (2008) 'Qualitative Inquiry between scientistic evidentialism, ethical subjectivism and the free market', *International Review of Qualitative Research*, vol. Kva1, no. 1, pp. 5–18. https://doi.org/10.1525/irqr.2008.1.1.5

Laitin, David D. (2003) 'The Perestroikan Challenge to Social Science', *Politics & Society*, vol. 31, no. 1. https://doi.org/10.1177/0032329202250167 and OA at https://perma.cc/72ME-UFUF

Lamb, Danielle; Russell, Amy; Morant, Nicola; and Stevenson, Fiona (2024) 'The challenges of open data sharing for qualitative researchers', *Journal of Health Psychology*, vol. 29, no. 7, pp. 659–664. https://doi.org/10.1177/13591053241237620

Leonelli, Sabina (2023) *Philosophy of Open Science*, Cambridge: Cambridge University Press. OA at https://perma.cc/P5M2-7XYG

Lindblom, Charles E and Cohen, David K. (1979) *Useable Knowledge: Social Science and Social Problem Solving*, New Haven: Yale University Press. $ https://perma.cc/4YWQ-F8UE

Mirowski, Philip. (2018) 'The future(s) of open science', *Social Studies of Science*, vol. 43, no. 2, pp. 171–203. https://doi.org/10.1177/0306312718772086

Montuschi, Eleanora (2003) *The Objects of Social Science*, London: Continuum. $ https://www.bloomsbury.com/uk/objects-of-social-science-9781847141101/

Montuschi, Eleonora (2008) 'Should we still compare the social sciences to the natural sciences?', *Sociologica*, no. 3.0, 10 pp. https://doi.org/10.2383/28770 or https://www.rivisteweb.it/doi/10.2383/28770

Nelson, Richard M. (2015) 'Physics envy: get over it', *Issues in Science and Technology*, vol. 31, no. 3, pp. 71–78. **$** https://www.jstor.org/stable/43314857

Pennington, Charlotte (2023) *A Student's Guide to Open Science: Using the Replication Crisis to Reform Psychology*, London: Open University Press/McGraw Hill. **$** https://perma.cc/5FQM-NZN9

Plesser, Hans E. (2018) 'Reproducibility vs replicability: a brief history of a confused terminology', *Frontiers in Neuroinformation*, vol. 11, no. 6 https://doi.org/10.3389/fninf.2017.00076

Ritchie, Hannah and Roser, Max (2019) 'Half of the world's habitable land is used for agriculture', 'Our World in Data', *Financial Times*, 16 February. https://perma.cc/9XNW-669N

Royal Society (2013) *Sciences as an Open Enterprise*, London: Royal Society. https://perma.cc/R8GB-U6QX

Simon, Herbert A. (2019) (1st edn 1969) *The Sciences of the Artificial*, Boston: MIT Press. **$** https://doi.org/10.7551/mitpress/12107.002.0001

SITC – Science, Innovation and Technology Committee (2023) *Reproducibility and Research Integrity* – Sixth Report of Session 2022–23, HC 102, 10 May. https://perma.cc/YT9C-75W5

Taleb, Nassim Nicholas (2007) *The Black Swan: The Impact of the Highly Improbable*, London: Penguin. $ https://perma.cc/6KRY-DMWC

Tarrant, Anna and Hughes, Kathryn. (2020) 'The re-use of qualitative data is an under-appreciated field for innovation and the creation of new knowledge in the social sciences', *LSE Impacts* blog, 8 June. https://perma.cc/P87U-FTN4

Tauginienė, Loreta; Butkevičienė, Eglė; Vohland, Katrin; Heinisch, Barbara; Daskolia, Maria; et al. (2020) 'Citizen science in the social sciences and humanities: the power of interdisciplinarity', *Palgrave Communications*, vol. 6, no. 89. https://doi.org/10.1057/s41599-020-0471-y

Tennant, Jonathan P; Waldner, François; Jacques, Damien C.; Masuzzo, Paolo et al. (2016) 'The academic, economic and societal impacts of open access: an evidence-based review', *F1000Research*, vol. 5, article 632. **$** https://doi.org/10.12688/f1000research.8460.3 OA at: https://pmc.ncbi.nlm.nih.gov/articles/PMC4837983/

Valentinuzzi, Max and Kohen, Alberto (2013) 'The mathematization of biology and medicine who, when, how?', *IEEE Pulse*, vol. 4, no. 1, pp. 50–56. **$** https://doi.org/10.1109/MPUL.2012.2228591 OA at https://perma.cc/QTJ2-SA6F

Wikipedia (2023) 'Open science'. https://perma.cc/ZRR5-QZSR

3. Open science in STEMM disciplines – and lessons for open social science

The overwhelmingly impressive fabric of modern science and technology disciplines centres on three foundations – open academic debate amongst knowledgeable colleagues (since the 1890s, within recognized disciplines); the transparent publication of results and investigation methods; and submission to peer review and outside scrutiny. This traditional form of academic-only openness and associated knowledge institutions – like journals, laboratories and field research (McNeely and Wolverton 2009) – have always been key to separating science from non-science (Wooton 2015). Yet in recent history there have been major shifts in the *practical* openness of science and technology scholarship to other researchers, and to other organizations or civil society actors.*

We begin with a brief, synoptic picture of how the regressive, 'closed science' trends in the 1945–1990 period shut off academic knowledge behind paywall barriers and boosted secrecy about innovations more than before (Section 3.1). Open science's (OS's) diverse practices evolved in overall opposition to this pattern, but across four main sets of disciplines with somewhat different OS approaches:

- computational/mathematical subjects;
- 'hard' laboratory or experimental sciences;
- open source technology trends; and
- 'field sciences' (Section 3.2).

Despite this diversity, by the mid-2020s the main features of a more integrated OS model for the science, technology, engineering, mathematics and medicine (STEMM) disciplines emerged, and we briefly note how each component also affected the emergence of open social science (OSS) (Section 3.3).

How to cite this book chapter:

Dunleavy, Patrick and Monteath, Timothy (2026) 'Open science in STEMM disciplines – and lessons for open social science', in Dunleavy, Patrick and Monteath, Timothy (2026) *Doing Open Social Science: A Guide for Researchers.* London: LSE Press, pp. 39–78. https://doi.org/10.31389/lsepress.dos.c

3.1 The postwar 'paywall science' period

What is science and why does it offer us a different, special kind of knowledge to which we should pay special attention? The modern field of science and technology studies has demonstrated that short or snappy definitions are often inadequate or misleading (Soneryd and Sundqvist 2023). We argue that at least these seven features characterize 'physical science' endeavours:

(i) *Focus*. STEMM disciplines study some specific complex fields – *either* the structure and operations of the natural world of physical materials and entities (including aspects of human behaviour); *or* the foundational operations of human-invented physical technologies; *or* the logical, mathematical or other computational operation of human-discovered or invented algorithms or systems (see Figure 2.3). So, the focus is on either an unchanging Nature or physical technologies founded on 'natural' science or on equally unchanging logical/mathematical relations.

(ii) *Regularities*. All STEMM disciplines look for *either* natural laws concerning how the physical world operates *or* permanently enduring probabilistic patterns of association and causation in the physical world; *or* deterministic logical/mathematical/computational laws within which 'artificial' technologies or systems operate. Scientific theory values parsimony, the ability to efficiently explain general results from the basis of the most feasible set of simpler premises.

(iii) *Tested models*. STEMM disciplines study important topics using formally set-out theories and logically consistent 'models' (often using mathematical notation) that can also be evaluated empirically by careful empirical investigation or shown to work by testing (including being reproducible by other researchers). Well-specified and empirically refutable hypothetico-deductive (HD) models of causal processes, with satisfactory *ceteris paribus* (all things being equal) controls, meet this criterion most directly. However, many inductively developed models backed by strongly founded empirical knowledge also qualify (Leonelli 2023).

(iv) *Cumulation*. If conditionally established results survive falsification efforts, and the models involved work in a way that is consistent with other known phenomena, the research frontier can move on to investigate new topics in a cumulative way, since underlying nature or rational logic both endure unchanged. Thus, STEMM disciplines always build on generations of previous work (Collins 1994).

(v) *Handling 'puzzles'*. If any disconfirming evidence for a proposed hypothesis or explanation surfaces, or if research results key for the hypothesis cannot be reproduced by other researchers, these setbacks cast a shadow on the propositions concerned, showing the need for more research or a theoretical rethink. They may not

immediately lead to abandoning a hypothesis or explanation, but they further conditionalize its acceptance by the scientific community, adding potential red flag warnings. A discipline's professional confidence in the propositions involved will also take account of further factors, including their fit with the currently accepted or dominant integrative theory 'paradigm' explaining the topics concerned, and the availability of equivalently plausible and powerful alternative paradigms (Kuhn 2012).

(vi) *Quantitative methods*. STEMM disciplines seek to systematically collect data and information using well-worked-out, publicly transparent and rigorously tested methods, especially those generating quantitative data outputs. Where probabilistic explanations are necessary the scope of available data and the adequacy of statistical analysis methods used are factors of the first importance – with the ideal in many ways being 'big data' (Kitchin 2014) allied with aptly employed methods.

(vii) *Scrutiny*. Finally, the apparatus of peer review is critical for policing the boundaries of the STEMM disciplines, involving:

- the submission of research to relevant professional journals run independently of the authors or their universities and organizations;
- assessment by multiple, well-qualified and professionally responsible reviewers not linked to the authors, following either closed or open peer review models, implemented in well-recognized ways and with views condensed by the journal editor(s) (Mayden 2012; Wikipedia 2024);
- the possibility for critical challenge of results and outcomes by other scholars and scientists, with open, public debate about the status of particular bits of 'knowledge'. Controversies should be accessible, managed in open ways and adjudicated by professional bodies assigning primacy to research integrity – such as university departments and review committees, chief editors and the editorial boards of journals, professional associations and research grant funders drawing on wider professional advice. Where research integrity fails, and empirical errors or methodological mistakes are clearly established, the articles involved must be retracted or modified (Royal Society 2018). Clear evidence of abuses by researchers should lead to them entailing significant costs – ranging from reputational damage through to their exiting academia (and perhaps even the wider profession altogether).

In modern science this complete set of features must be present simultaneously to separate it out from other kinds of knowledge. This completeness was only really first achieved in the German 'research university' model that emerged by the very end of the nineteenth century. Often mistakenly seen as

driven by STEMM innovations, in fact the academic side of the model was developed first and fastest in humanities fields like philology (the study of languages), archaeology and law (Watson 2011). This model then spread rapidly to STEMM disciplines. Meanwhile, from the 1870s steam technologies and related engineering, chemistry and medicine developments, plus early factory semi-automation, produced a strong expansion of industrial laboratories that changed and helped systematize university science also (Morris 2021). By the early twentieth century 'science' was recognized as key to almost all innovation. Rapidly copied in the US, the research university model transformed American academia in short order in the early twentieth century (Geiger 2017). However, it took decades longer to fully penetrate the UK and French university systems. STEMM models of strong independent peer review spread into medicine and from there into all other science disciplines by the 1940s. From there the sequence of PhD then peer-reviewed publications progressively dominated the social sciences by the 1970s and strongly influenced humanities disciplines too.

By the 1980s a strongly developed academic research system operated, partly funded by the expansion of mass university education and partly by heavy state and corporate funding of STEMM research, creating many research-only staff (35% in the UK). Academic career paths standardized across all subjects within universities – centring on the large-scale production of PhDs, and academics being continuously enmeshed in a 'publish or perish' system producing peer-reviewed publications to secure jobs and promotion (Baker and Powell 2024). The numbers of journals mushroomed, creating an expensive peer review economy financed by high journal subscription prices. This increasingly limited any near-complete access to all major journals to a small elite of globally significant universities.

Some powerful additional exclusionary effects also followed from the mathematization of the STEMM disciplines, one after another, apparently following Black's (1962, p. 242) dictum that 'every science must start with metaphor and end with algebra'. Previously text-heavy or more qualitative or judgemental sciences (like biology or medicine) increasingly moved to using mathematical expressions as the basis for expositions, with formal models and data-heavy analyses comprising the largest parts of 'original research' (Valentinuzzi and Kohen 2013). Intuitive and descriptive elements, especially word-heavy text expositions, and even descriptive statistics, were stripped back to meet journal compression demands, making articles increasingly technical and 'hard-boiled' in character (Murphy 2022). Additionally, many academic authors clearly felt that research professionalism demanded a new kind of 'alchemical' communication ethos, speaking only to the provenly 'elect' in their sub-field in equations or jargon-laden prose. Journal articles became the overwhelmingly important means of professional assessment. And to maximize metrics scores for 'outputs', research project findings were increasingly 'salami-sliced' across as many different articles as possible (*Nature Materials*, 2005). Publication hierarchies based on journal reputations rather than the inherent value of each article

became increasingly set in stone, with the top 10 journals in most STEMM disciplines dominating references and citations, even in the vastly expanded journals universe by the 1990s (Acharya et al. 2014). These trends quickly extended into the quantitative social sciences (albeit with some lag) in a complete manner, and they became increasingly influential even in more qualitative fields and in the humanities.

Finally, the commercialization of science and technology innovation and learning greatly expanded the modern scope of industrial- and government-funded STEMM disciplines within universities, and grew the size and relative intellectual influence of external corporation or government laboratories. Some of the most consequential and intellectually influential research (e.g. in IT and the internet, drugs development, genomics) was now applied commercial or governmental science, conducted in labs with superior salaries, funding, equipment and access to data. Both academic and industrial applied research hunted for commercially exploitable comparative advantage features in ferocious secrecy – the diametric opposite of previous norms of fast communication and open critique in science (Resnik 2006). University STEMM researchers were left increasingly just reacting to or evaluating innovations made elsewhere, for example helping governments regulate industrial science's drugs, materials and IT-related changes (Baker and Powell 2024). Companies and governments established their own scientific review processes for internal research, partly because academic peer review was too erratic in its value, and too time-consuming in a swift-moving environment. (Even now, after a strong shift towards OS practices, recent studies suggest that around 40% of research studies held by U.S. life sciences and government researchers have never been published (Bowers et al. 2023)). In the 'neo-liberal' era from the 1980s onwards, universities were also pushed by donors and governments to do 'entrepreneurial' science or tech themselves via joint ventures with firms, hiving off research or establishing 'starburst' companies – a trend that extended into many social sciences via consultancy work. So each of these regressive trends seemed to push the STEMM disciplines ever closer towards becoming a closed academic scientific world, cut off from wider readers and influences.

3.2 The push for open science – key proximate sources

A set of precipitating influences for change brought long-run discontents and minority STEMM concerns (like 'democratizing science') from their previous positions on the fringes of major disciplines into much more central (if still 'insurgent') positions (Bartling and Friesike 2014; Hangchen et al. 2023).

First, the previously 'privileged' status of STEMM was challenged far more from the 1990s onwards. A whole series of scandals involving the physical sciences and medicine emerged, denting confidence in STEMM research.

Some followed professional 'experts' and science-informed regulators giving a clean bill of health to substances or products that proved damaging. Examples include asbestos, tobacco, multiple drugs that went to market with side effects discovered only after widespread use, contaminated blood factor products and levels of 'safe' exposure to air, water and contact pollution and nuclear risks that were only ever revised downwards over time. Allied with some scandals' close association with 'industrial' science, these changes have somewhat damaged western publics' confidence in STEMM researchers' abilities to independently judge the human impacts of many developments they originated or helped to promote (Cologna et al. 2025).

Public trust in scientists also declined as successive revelations emerged of their research being funded or linked via jobs, conferences and data-accesses to important corporations with rather clear vested interests in the research. The stringency of interest declarations increased strongly, especially in medicine, where successive major figures were found to have signed off as authors on research in which they had barely participated – thanks to the 'big pharma' apparatus of 'medical communication' writing mills for drug company-sponsored research. Nor was the problem confined to one field. In different liberal democracies many government expert committees and advisory bodies turned out to be dominated by scientists with potentially conflicting corporate interests or links (Ruff 2015). Yet their expertise was crucial for governments seeking to set standards for the safe use or consumption of almost all products, such as drugs, food safety standards, safe alcohol limits, tobacco smoke exposures and cladding for high-rise buildings (Demortain 2011; Ball 2021). Some critics also argued that STEMM sciences showed gender bias towards male topics, with under-citation of female scientists' work (Murphy et al. 2020). Enhancing openness to remedy all these manifold transparency defects became an increasingly important part of STEMM researchers' professional ethics. As a result, the norms around disclosure of conflicting interests tightened up a great deal.

Second, broader concerns also emerged about apparent over-confidence amongst scientists and technologists that lab or tech innovations which had passed peer review, and even survived randomized control trials (RCTs), could then be directly applied to 'solve' problems in causal processes or social situations, without first researching or appreciating how these 'solutions' were experienced by people (e.g. McClaren and Markussen 2020). Internal STEMM critics stressed the need for separate 'translation studies' to examine how new technologies or solutions could be applied in 'real life' situations massively more diverse and complex than the variables tested in labs, trials, beta versions or even RCTs. Especially in medicine and health studies, qualitative research into people's empirical experiences of illness and disability made an important comeback, generating new insights from listening to patients and research subjects (Wethington and Dunifon 2012). Similar lessons gradually extended across many other fields – for example, in studies of how to mitigate global warming and climate change an emphasis on engineering or tech-only

solutions repeatedly proved fruitless in getting firms or citizens to change their actual behaviours (McLaren and Markusson 2020).

Within key science professions doubts were also increasingly voiced about the conventional HD model, used by scientists in very simple-form ways to defend their practices, arguing that 'non-falsified' hypotheses sanctified by acceptance during peer review had to be accepted. Amongst the earliest blows to the orthodoxy was Kuhn's (2012) analysis of scientific conventional wisdoms in fact tolerating numerous apparently disconfirming results as 'puzzles' unless an alternative paradigm explanation could explain them better. It was succeeded by a flood of other work stressing science as a social process. In the emerging sub-field of social studies of science and technology (STS) some observers even argued:

> Since the early 1970s … the 'logic of scientific discovery' has been displaced by detailed examinations of science in practice; this has eroded the cultural position of scientific expertise. Furthermore, the 'crown jewels' of science, Newtonian physics and the like, are no longer accepted [by many populists] as justifying science's contribution to citizens' more diffuse technical concerns. Scientific expertise now seems more fallible, less removed from ordinary decision-making and less insulated from political and social forces. Populist leaders, who attack scientific expertise because it limits their power, can draw on these ideas (Collins et al. 2020).

Awareness also grew of how more pluralistic and more inductive scientific models operated across different STEMM disciplines and sub-fields, with the effect of 'dethroning' physics models and processes as the dominant form of scientific truth-finding (Leonelli 2023, 2010, 2012, 2020). A more sociologically realistic picture of scientific processes emerged, along with a recognition of pluralistic models of scientific enquiry. This process also tended to discredit the naïve argument that simply copying STEMM methods (in whatever variant) would be enough to understand 'human behaviour' in law-like ways.

The third key precipitating influence that transformed the feasibility of the whole OS agenda was a growing critique of dominant, 'publish or perish' science processes as liable to produce too many poor or false results. Critics identified a host of emergent adverse effects (10 of them shown in summary form in Figure 3.1). They argued that paywalled science publication efforts by academics in the mass education era differed a great deal from ideal procedures:

> The careful scepticism required to avoid massaging data or skewing analysis is pitted against the drive to identify eye-catching outcomes. Unbiased, negative and complicated findings lose out to cherry-picked highlights that can bring prominent articles, grant funding, promotion and esteem.

Figure 3.1: Ten key problems with conventional (non-open) science processes alleged by open science critics

Problem	What is meant
Peer review failings	Repeated instances of retracted articles and withdrawn work (especially afflicting the highest-prestige journals) show that poor science claiming 'original' findings can often pass peer review, which is only ever as good as the reviewers and editors involved. Strong sub-field specialization and complex mathematical techniques make expert close assessment hard to secure. Equally, conservative peer reviewers may turn down too much innovative work that seems unexpected to them
p-hacking	Massaging data or variable specifications in ways that allow them to pass the conventionally used 5% or 1% 'significance' test limits in statistical package software, where unmassaged variables would not. A serious problem is indicated by unusually high numbers of cases occurring at or just above significance levels in many fields
Harking	'Hypothesizing after results are known' – leaving the propositions supposed to be tested by the research vague or unstated at the outset, and then firming them up only after analysis outcomes are known
Data-dredging	Generating a dataset with every conceivable variable included, or with a variable list conditioned by the existing availability of data rather than theoretical guides; and then systematically looking within the data for apparent effects that 'pass' statistical significance, followed by constructing post hoc rationalizations around them
Omitted variables bias	Failure to control for theoretically salient explanatory variables that may well be empirically important or condition the results found for tested and 'significant' variables
Over-publication of positive results	Numerous studies claiming positive associations are printed in journals, but disconfirming, negative or null hypothesis confirmed studies are far less likely to be accepted – especially in high-prestige journals
Multiple publication tries	Redoing rejected articles to meet referee criticisms and submitting them to another journal, lower down the hierarchy of journals, until the paper is accepted. The process is industrialized in some medical and pharmaceutical fields where professional writing agencies organize blanket submissions to many outlets on behalf of the researchers named as authors

Figure 3.1: (continued)

Problem	What is meant
Publication glut effects	Very high levels of outputs swamp real professional attention. This may be especially true of imitative articles following exactly 'fashionable' or influential work. With high levels of professional 'noise' work with only flaky peer reviews can enter 'the scholarly canon' with little outside critical assessment
Metricization feeds over-competition	Career advancement depends on maximizing research output numbers, assessed using either journal 'prestige' measures, or citations or other usage indicators (e.g. Altmetrics scores). Scientists face pressure to 'salami slice' connected research (gaining more outputs), and to over-find positive results and over-claim for originality (to reach higher ranked journals or attract more cites)
Non-reproducible research findings	Large proportions of accepted articles in some disciplines show results that cannot be found in replication studies, or effects that are greatly over-stated – producing a 'replication crisis' in some disciplines (like psychology)

Source: Authors.

> The 'results paradox' is a chief cause of unreliable science. Negative, or null, results go unpublished, leading other researchers into unwittingly redundant studies. Ambiguous or otherwise 'unattractive' results are airbrushed (consciously or not) into publishable false positives, spurring follow-up research and theories that are bound to collapse (Chambers 2019).

Of course, the points in Figure 3.1 did not emerge fully formed all at once, but rather were built up over time in quite a piecemeal fashion.

Psychology was the mixed STEMM and social science area worst hit by claims of a 'replication crisis' (Pennington 2023, Figure 3.2). We summarize Pennington's valuable timeline in this paragraph, starting with a now famous paper by John Ionnidis (2005), 'Why most published research findings are false'. He argued that most psychology research outputs had too few participants or respondents (small *N*s) and were 'underpowered statistically' to show the findings claimed. A band of 'replication studies' authors in psychology subsequently generated more examples of dubious propositions passing peer review to get published. For example, in 2011 Bern et al. used nine experiments to cast severe doubt on an always implausible article by Stapel that claimed to have experimental evidence of pre-cognition in lab subjects. Ionnides and colleagues widened their focus to 'questionable research practices' widespread in the discipline. In 2014 the first results of a large-scale replication study suggested that many apparently important results in social psychology could not

be stood up. In 2015 the Open Science Collaboration found that only 36 out of 100 randomly selected results could be reproduced. In response to the wave for change, some journals embraced OS methods and requirements, especially pre-registering studies, explaining methods more fully and depositing data for reusers to analyse. And by the 2020s, journals also began opening up to reporting some null results or the non-confirmations of hypotheses (Santoro 2022; Mullen 2024).

A broadly similar trajectory occurred in medicine and health studies. In 2012, Begley and Ellis (2012) successfully reproduced only 3 out of 53 pre-clinical cancer drug studies. In a 2016 *Nature* survey, 67% of medical researchers in an online survey said that they had failed to reproduce a published result when trying to do so (Stoddart 2016). Later on, a major study of cancer treatment studies found that only 46% were replicable (Mullard 2021). Other STEMM disciplines resembled medicine in the 2016 *Nature* survey, but some reported much higher non-replication levels, notably in chemistry and biology (87% and 77%, respectively) (Stoddart 2016). Some scientists also reported editors and reviewers asking them to tone down adverse comparisons between their different results and original studies.

As with other science controversies, a pushback soon got under way against some more dramatic or exaggerated estimates of research misdoings. For instance, only small proportions of medical studies showed evidence of questionable practices (Fanelli 2017). Achieving an effective reproduction of original research carried out in a different lab is actually a very complex task (Brandt, Ijzerman and Dijksterhuis 2014) – so that non-reproducibility may have multiple causes (Fanelli 2022). Nonetheless recognition crystallized that STEMM disciplines needed to establish the reproducibility of important results.

The possibility of perverse evolutionary trends in science also could not be ruled out (Smaldino and McElreath 2016). Karl Popper's falsificationist optimism about science relied on critical evolution pressures to winnow out poor studies, so strengthening sound science over time (Popper 1979). Yet even if there was no replication crisis in most disciplines, STEMM disciplines in general recognized that the problems in Figure 3.1 might indeed lead to the perverse selection claimed by critics. In effect, the influence of poor studies 'forced' through journals acceptances could be magnified by competition for citations, and the publication (and subsequent promotion) of researchers generating the most dramatic (but likely to be the most flawed) 'original' results (Fang and Casadevell 2011).

On another tack, about originality, critics also argued that grant funding driven by publication metrics produced large volumes of 'me too' research. Authors had to struggle to show that their results only minimally different from those already known were 'original' enough to justify funding and publication, and were almost always able to secure acceptance at an outlet somewhere down the hierarchy of progressively less prestigious journals. Some funding foundations experimented with randomly deciding which (legitimate

and well-qualifying) applications to fund, rather than trying to anticipate their likely success – although the results suggested no overall average improvements in impacts. Yet more high (and more low) performing projects occurred in randomly selected portfolios compared with projects in the set that was individually assessed (Phillips 2022).

A final big group of factors pushing OS agendas involved digital change processes. The rise of digital publishing led to rising demands for open access to scientific work – especially where it had been funded by 'taxpayers' money'. As the customs and costs of legacy publishing on paper proved increasingly pointless, and as STEMM readers shifted over to PDFs stored on computers, so academic criticism soared of the extraordinary profits made by global publishing conglomerates charging super-high paywall subscription fees for work submitted and refereed for free by academics (Buranyi 2017). Demands for free access to scientific papers rejuvenated and modernized the previously embattled tradition of scientific collegialism. It also drew inspiration and plausibility from open source IT developments, especially the rallying call that 'Information wants to be free' (see Section 3.3). Later on, the rapid growth of social media meant that scientists could easily talk directly to colleagues globally via preprints and online repositories as well as online conferences and seminars, making journal publication even more of the final 'tombstone' stage for research. And via social media (especially blogging) scientists could also reach academics in cognate disciplines, relevant civil society communities and the public at large, without having to filter their knowledge first through journals, journalists, commercial media or public broadcasters (Dunleavy and Tinkler 2021, Chapter 7).

In response, academic publishers progressively shifted to open access formats financed by billing authors and universities a fee for accepted articles (APCs – article processing charges). Governments in Europe and the UK who publicly funded almost all scientific work, and major foundations (like Wellcome) and funding agencies in the US, all saw open access as a way towards increasing the efficiency of their inputs. They lead the way in pressuring STEMM professional bodies (who ran many journals) and universities to switch to open access arrangements. For all but the smallest science/tech projects (and the most costly journals, like *Nature*) these fees added fairly small extra costs onto research grants (Tennant et al. 2016). Open access publishing introduced many STEMM researchers to the broader OS agenda, creating a new front to penetrate and challenge conventional, closed-science ways of working (see Chapter 15).

The digital revolution also facilitated much more detailed lay scrutiny of research outputs, especially after preprints became universal and social media developed as key channels for the dissemination of STEMM work and its critique. Like other digital developments some major positive and negative implications followed for OS. Feedback speeded up about what worked and what did not, improving methods and curtailing wasteful duplication, while more scientists could comment earlier on findings, and university scientists

(at least) had strong incentives to be open throughout on their analyses. At the same time, strong drawbacks emerged in terms of vested interests contesting results on ill-informed or disinformation grounds, and populist movements, religious interests and others openly seeking to denigrate and intimidate science research they disliked, often with endorsements from top politicians, leading to unprecedented partisan polarization over STEMM in the USA (Dunlap, McCright and Yarrosh 2016; Rekker 2021). In spring 2025, the populist U.S. Vice President J. D. Vance (2025) argued on X: 'There is an extraordinary "reproducibility crisis" in the sciences, particularly in biology, where most published papers fail to replicate'. This misinformation formed part of a bizarre justification for the second Trump administration to assert a right to directly micro-manage all federally funded scientific research.

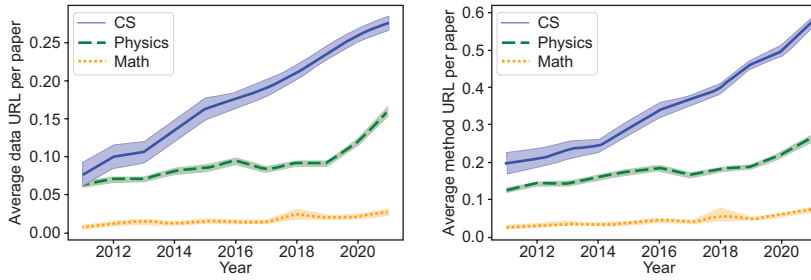
3.3 Varieties of OS in STEMM disciplines

Four main areas with diverse impulses driving open science changes can be identified.

Computational and mathematical sciences

Since Euclid, mathematics has always had a strong emphasis upon the full publication of proofs and the logical reasoning underlying arguments, before any knowledge can be accepted as valid. The discovery of any mistakes in proofs or operations immediately vitiates published work until defects can be remedied in a transparent way. So, the payoffs from OS initiatives here might seem to be far smaller than in other STEMM disciplines. And, indeed, Figure 3.2 shows that in 2021–2022 only around 3–4% of mathematics papers

Figure 3.2: Trends in the citing of data or methods URLs across computer science, physics and mathematics, 2012–2022



Note: The dark central line for each series shows the mean decimal percentage frequency of data URLs (top chart) or methods URLs (bottom chart) per paper. In each case, the paler shaded area around the dark line shows the 90% probability range.
Source: Reproduced with permission from Hancheng et al. (2023).

cited data or methods URLs, and there has been a very small increase in this number in recent times, in acute contrast to a surge in OS practices in computer science.

However, embedding FAIR (findable, accessible, interoperable, reusable) data principles in mathematics is still a work in progress (Conrad et al. 2024). These authors argue for further steps to ensure that symbolic data (like formulae, full proofs and functions), numeric data (like number sequences, matrices, finite lattices), geometric data (about surfaces or high-dimensional objects), maths models and empirical or observational data (such as simulations and sequences) are all deposited satisfactorily for the long term in data repositories – as well as occurring in research papers and appendices.

> Scholarly publications mainly articulate theoretical advancements, methodologies, and analytical insights, whereas research data repositories can provide various other types of data, for example, empirical datasets underpinning these academic discussions. This distinction emphasizes the importance of research data repositories in enabling empirical scrutiny and verification, a fundamental aspect of the scientific process. Thus, the integration of both research outputs enhances the robustness and credibility of mathematical research, highlighting the essential role of data repositories alongside traditional scholarly narratives (Conrad et al. 2024).

Their prescriptions stress the importance of researchers making good initial choices, such as:

- opting for user-friendly platforms (like Zenodo) rather than highly technical ones (like the Archive of Formal Proofs);
- fostering far greater data interoperability and standardization across sub-fields and topics in how data was stored;
- using good metadata to aid discoverability, such as the features provided in Harvard Dataverse or Wikidata (see Chapter 7).

Other OS advocates have also stressed the potential gains from using open notebooks (Rule, Tabard and Hollan 2018) and open code practices in maths, as well as more freely sharing advice to other maths scholars on how to handle unfamiliar materials or operations. In addition to traditional reliance on seminars and conferences, mathematical research data and collective knowledge of trends in the discipline are often published online using wiki-based systems. They help mathematicians to operate in collaborative ways, publicizing new developments (albeit only once proved) and following rumoured new discoveries or lines of advance in ways that are also a little more accessible for outsiders (also see Chapter 5).

Both in computer science and in many different STEMM disciplines within and outside Simon's *Sciences of the Artificial* (2019) computational work has

been by far the most advanced and rapidly growing area of open science, as Figure 3.2 shows. The frequency of URLs for data sources and methods in computer science papers has virtually tripled in the decade 2012–2023. In principle, all algorithms and operations using pure logic, computer programming and large-scale quantitative data manipulations should be fully documentable and classically deterministic in their operations. Partly this rapid surge in OS practices in computer science reflects the success and prestige of open source technologies discussed in the next sub-section, and partly the lesser importance of journals and the dominance of digital-only conferences and papers in the discipline, which facilitates extensive use of URLs.

Yet it also represents strong commitments to OS amongst academics via important OS initiatives such as the Turing Way (2024a) collaborative project. These reflect the inescapable problem that even well-documented code can generate unappreciated interaction effects when multiple elements are combined into whole data science programs and activities. Early Turing work stressed major bugbears in data science, such as:

- version control, problematic because analysts often quickly change computer programs in ways that meet an immediate need but then are not documented for future users;
- testing programs when assembled, which requires careful lab organization, with many commercial temptations to skip some onerous processes in the rush to get software betas out; and
- reproducible computational environments.

Later Turing work stressed that data depositing hugely simplified the problems of publishing negative research, helping prevent other researchers wasting time on fruitless duplication of effort. Documenting in OS ways also facilitates tracking complex project histories more effectively; makes collaborations easier (especially across different research sites); helps the writing of papers to be more efficient; allocates credit more fairly (especially in large teams, discussed in Chapter 16) and ensures better continuity within labs from one generation of work (or set of ECR researchers) to another (Turing Way 2024b).

b. 'Open source' technology

Reassessing what openness means within modern STEMM science has coincided with and been driven by digital change, in many diverse ways. The rise of open source tech doctrines is an important aspect.[1] 'Open source' is an apparently paradoxical movement occurring in purely formal and commercially critical fields, where secrecy might have been expected even in academia, as still happens to some degree in wider engineering. Instead, the Free and Open Source Software (FOSS) community spanning across the IT industry,

universities and (university-educated) freelance programmers contributed strongly to the wider momentum of open science. Many overlapping projects, personnel and concepts in FOSS intersected to produce multiple convergences (and some divergence) between it and the wider open source community. Many core propositions are underpinned by open source software, both for infrastructures and practical implementation.

FOSS originated in the 'hacker' cultures amongst programmers and computer scientists reacting against the increasing corporatization and commercialization of software development (Moody 2002). This powerful pushback stressed that software should be free for anyone to share, use and modify through access to the source code. The free software movement is primarily politically motivated and unrestricted access is a point of principle.

> These are the so-called four freedoms: (1) freedom to use the programme for any purpose; (2) freedom to study the programme and modify it to adapt it to your needs; (3) freedom to freely distribute copies of the programme; (4) freedom to improve the programme and to publicly distribute the improvements, so that the whole community could benefit. The so-called copyleft, therefore, uses copyright to achieve the opposite of what it usually does: instead of imposing restrictions on rights of use, it extends these rights (Ramella and Manzi 2021, p. 82).

By contrast, the 'open source' term was coined in 1998 to rebrand the free software movement towards a focus on technological development that was able to work more easily alongside and with commercial interests (Raymond 1998). Despite these differences, the FOSS movement has been united in seeing software development as a collaborative endeavour in which communities of developers can create and maintain software transparently, so that it can be examined or contributed to by anyone with the appropriate technical skills, contrasting strongly still with proprietary approaches, as Figure 3.3 shows.

Developing FOSS software has a wide appeal for programmers for different reasons, from those motivated by the politics of free software to those (or their employers) who simply see FOSS as a more efficient and practical way to develop new software because it allows code for common problems to be reused, rather than being rewritten from scratch. New software can be built on top of existing FOSS programs, and many diverse contributors can work together to tackle complex goals or catch errors in the code. The popular OS dictum 'With many eyes all bugs are shallow' highlights that with many different contributors working on a piece of software, critical faults that one programmer could easily overlook will be caught by someone else. The extent to which FOSS underlies modern computing is often underestimated by those who are not actively involved in its development. For example, all Android phones run on the FOSS Linux kernel, as do the majority of the world's

Figure 3.3: The contrasting approaches of two FOSS streams and two proprietary modes of developing software

For users	Free software	Open source	'Free' platforms	Paywall proprietary software
Costs	Free to use	Free to use	Free for end users	Paid
Code	Accessible and modifiable	Accessible and modifiable	No access	No access
Restrictions	No restrictions on use, derivatives may not be relicensed or made proprietary	No restrictions on use, derivatives may be relicensed or made proprietary	Use at the discretion of the owner, no modifications	Only usable with a paid licence, no modifications allowed
Motivating philosophy	Openness is best for software development. As a political principle all software should be free and open	Openness is best for software development, often supported by corporate developers	Users fund the platform company's business model (via advertising, data capture, future fees)	Costs of software drive the provider's business model

Source: Authors.

websites on which FOSS encryption packages (such as OpenSSL) are used to secure personal details and transactions. For the FOSS community, developing software leads to the production of higher quality software, and the success of this approach strongly influenced OS advocates, who argued that an open approach to developing, conducting and publishing research simply produces better science (Watson 2015). In addition, open science itself has been a relatively small influence on FOSS, although the creation of STEMM-specific software has formed only a small part of the overall FOSS ecosystem. A 2020 survey found that only 7% of FOSS project maintainers worked in either scientific or education settings (Linus Foundation 2020).

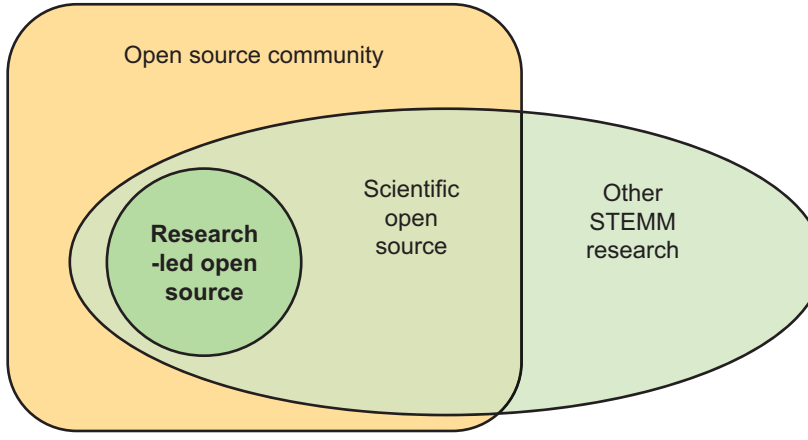
The FOSS model also helped many different specific areas of research build atop a wider ecosystem of open source and open science software, so that programming languages (such as R and Python) are widely used across the STEMM disciplines. Scientists can rely on a relatively consistent 'stack' of libraries that provide basic functionality. For example, the Python library

NumPy is designed for easily coding and efficiently computing numeric calculations and will be used by almost all software that is written in Python where users interact with charts, tables or numeric data. The programming languages, like Python and R, and the stack of general purpose libraries providing a wide range of functionality (from converting location coordinate systems to making websites), form the diverse eco-system of open source software (Figure 3.4). They are widely used, supported and taught within their respective research communities who design them to last and to be reliable, and are invested in the ongoing development. New research can then build on top of these libraries to draw on their specialist functionality.

Furthermore, code can easily be changed or adapted to the needs of the project. (By contrast, research reliant on proprietary software is at the mercy of the company to provide updates, add new features and maintain the product. In the worst-case scenario, an entire programme of research can be scuppered if a feature is removed by a commercial provider, or if the software is discontinued.) In a wider ecosystem of open source code, research also benefits from improvements in the stack of libraries that it draws upon, including maintenance through bugs correction by the FOSS community and continually added improvements. For example, the optimization of calculation algorithm in NumPy will speed up the code for all projects that use the library without the maintainers needing to be aware of this change or updating their own code. Existing software libraries can be built on or modified by future researchers and used in new contexts beyond the goals of the original researchers.

The advantages of open source software development for OS also extend beyond the technical to the practice of science by researchers. The predominance of a few programming languages within scientific research and stable

Figure 3.4: How open source supports STEMM research



Source: Authors.

stacks of commonly used libraries enable scientists to develop interoperable skills between a wide range of research settings. Introductory courses for languages like R and Python can be taught to scientists as part of broad interdisciplinary training programmes – greatly easing how researchers transition between projects, and also reducing some barriers to interdisciplinary collaboration. The continued availability of open source software, along with the numerous scientists with the training and skills to use it, all aids in furthering the reproducibility, replicability and robustness of scientific research.

Yet there have still been many problems in persuading academic scientists to acknowledge the key roles of the open source movement (and of background digitization trends) in encouraging the shift towards implementing open science internationally. FOSS is often seen as just a smaller component of the system-wide changes that are needed. For example, the European Union (EU)-funded FOSTER project mapped the field of open science, and yet did not list FOSS development directly, instead splitting multiple open source effects across several third-level components of OS practice (see Moody 2019).

Experimental and laboratory-only science

Many STEMM disciplines are those where either fully controllable systems (usually technologies) are being investigated or where natural phenomena can be isolated sufficiently to enable a *ceteris paribus* clause to stand in an unproblematic way. The latter case typically means operating in a very reductionist way, setting up or controlling initially complex materials, systems or phenomena so that a single effect of causal interest can be investigated. In classical science the search was for a fully determining causal process, but modern STEMM research mostly treats of probabilistic connections and causal connections.

An important moment was reached in the mid-2010s when conventional or Standard Model defences of a general 'scientific method' began to lose credibility with scientists themselves, as well as with the general public and some funders (as discussed above). Instead of just accepting on faith the claim that STEMM disciplines would evolve ineluctably towards accuracy, some scientists began to look empirically at evidence about the actual tracked and modelled evolution of the research-and-publishing chain. For instance, one early formal (mathematical) model sought to understand the science research and publication process using realistic parameters for publication success and promotion. Given the overwhelming career importance of original findings, the authors' model predicted that rationally motivated researchers would undertake many exploratory studies with datasets that were under-powered for their purposes and likely to generate misleading results (Higginson and Munafo 2016).

Within a few years of this rather more self-critical turn, the problems of reproducibility in medical research and drugs testing created a wider feeling

that all was not well with many STEMM disciplines. Mirowski (2018) argued that almost all contemporary policy documents addressing scientific funding and reform shared a 'taken-for-granted premise that modern science is in crying need of top-to-bottom restructuring and reform'. He saw the agenda as 'effectively to re-engineer science along the lines of platform capitalism, under the misleading banner of opening up science to the masses' (p. 171). So what open science meant, and what problems it was claimed to solve, constantly shifted within these policy prescriptions, and the remedy to the 'crisis in science' was OS as a panacea to cure all its ills, a stance mirrored in Wikipedia's (2023) account. Yet –

> Open science practices have not been adopted to the same degree in all research areas. In psychology, which was strongly affected by the so-called reproducibility crisis, the open science movement initiated real practical changes leading to a broad implementation of practices such as preregistration or sharing of data and material. By contrast, biomedical research is still lagging behind. (Diderich et al. 2022).

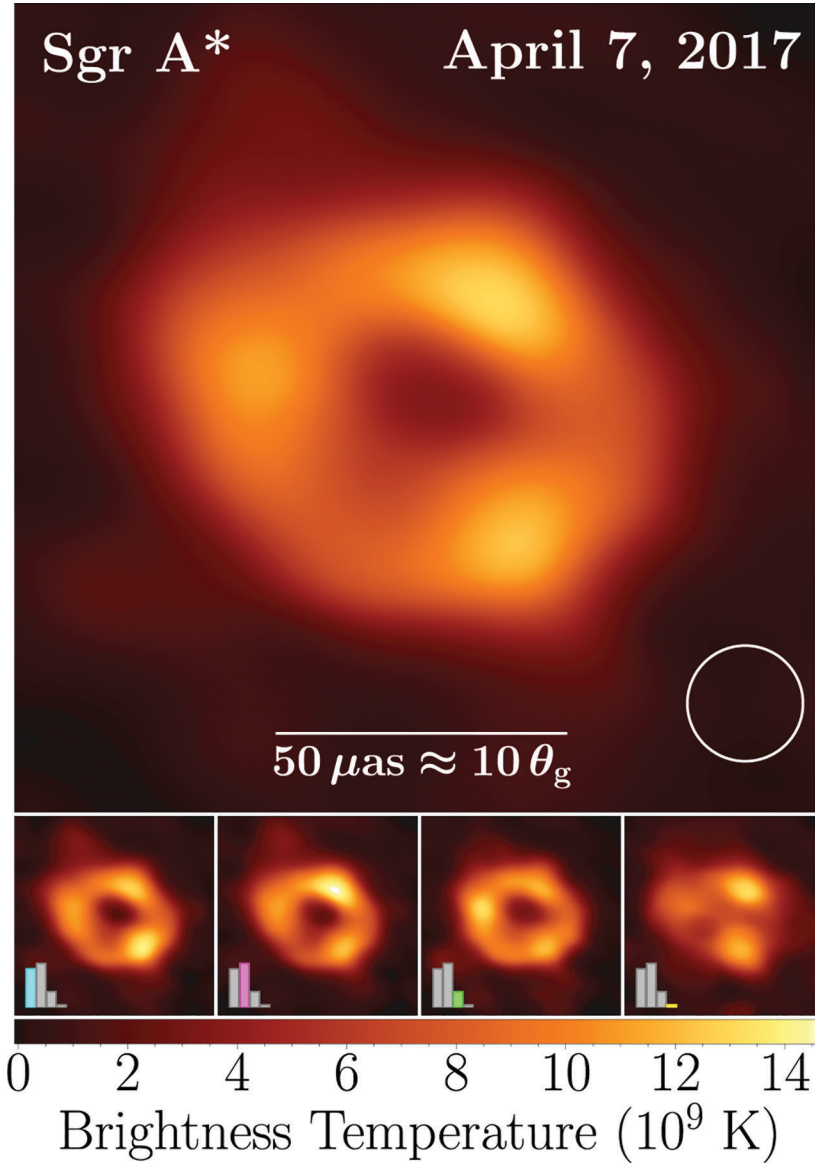
In some other catching-up OS fields the stimuli for change were smaller, as in physics, where citing of data and of methods in URLs rose little across the 2010s, albeit speeding up after 2020 (see Figure 3.1).

Instead, using resources more efficiently, effectively and collectively often played a larger role. Where there were serious risks of duplicating already falsified hypotheses, funders expressed repeated concerns over the costs of ultra-large projects like CERN, involving hundreds of researchers and sometimes thousands of authors per paper. The strong overlap between the free and open source movement and OS created reinforcing advances in each. For example, consider how many people it takes to photograph a black hole, of which the first ever images were published in 2022 (Event Horizon Telescope Collaboration 2022; and see Figure 3.5 below).

Each of the Event Horizon Telescope's (EHT's) black hole image papers had over 200 authors listed – reflecting the complexity of the research and collaborative efforts needed to tackle it. But, in addition, thousands of technicians, support staff and administrators across the 60 institutions in the project helped make the research possible. One computer scientist (Katie Bouman) who helped develop the image processing algorithm that made the photograph possible estimated that 23,000 people in fact played some role in enabling the image below (Rao 2021). In conglomerated teams this vast, most contributors will have worked on a single piece of general purpose software (most likely not specifically built for scientific purposes), and so are likely to be unaware of the small role they played in the scientific breakthrough.

More specialist extensions for more general programming languages (like Python and R) extend the diverse eco-system of open source software specifically for scientific applications. For instance, Astropy was developed by the

Figure 3.5: One of the first ever pictures of a black hole published in 2023



Note: An EHT project image of a black hole designated, Sgr A* (top; Paper I). Ring-like images dominate the wide range of images obtained across multiple methods; however, variability and sparse visibility domain coverage make selection of a single image impossible (Paper III). The inset images represent different imaging solutions and their associated frequency (histograms).
Source: Event Horizon Telescope Collaboration (2022a).

astronomy community to address common research tasks. Its broad reach covers functions to perform calculations commonly needed by astrophysics researchers, such as non-Earth-centric coordinates systems or specialist measurements of time like atomic or relativistic time, which those outside the astrophysics community are unlikely ever to need. So empirical open science is now mostly built on open source software. For example, the Python library NumPy is ubiquitous amongst anyone using Python for scientific purposes (although it was not originally designed for this), as it was in the EHT team's image-processing pipeline. In particular, 11 scientists (including Bouman) created the *ehtim* library for imaging radio interferometry, a piece of highly specialized software, but one that drew both on wide open source community efforts, and those of a smaller astrophysics development community, and was used to create the image of the black hole in Figure 3.5. In every STEMM discipline similar or parallel packages are widely used, supported and taught within the research communities, who built them to last and to be reliable, and are invested in their ongoing development.

The field sciences

A large group of STEMM subjects, including some of the most applied work, are field sciences like geology or ecology. Here data gathering work chiefly happens outside of lab settings, in evanescent conditions, in a 'dappled world' suffused by hundreds of causes operating at once (Cartwright 1999). Some observations (such as the output plume of a volcano or the nesting behaviours of sibling birds) can only be captured once (Fowler 2016). And fully specifying control variables in order to focus down on a single possible causal connection, excluding all others, is often impossible. Field research rarely establishes fully determinant findings, since the *ceteris paribus* clause is not fully operationalizable (although many conditions can be substantively controlled via careful study selection processes). Field research establishes statistical probabilities only, and statistical controls for possible confounder variables (including ineradicable spatial and temporal variations) are always partial at best and only as good as theory models can make them.

These are also often more 'holistic' sciences. Even where some data can be exactly measured (e.g. the dispositions of bones in an ancient burial site), a full explanation may rely on examining multicausal variations, paying attention to complex interaction effects that are rarely appreciated prima facie, and fitting available evidence within a wider, sometimes qualitative explanation. Some physicists (and other observers with 'physics envy') have sometimes suggested that fields like archaeology, palaeontology, ecology and conservation science are simply early in development – 'held back' from fully realizing HD and falsificationist 'standard science' by poor data which partially confines them to Black's (1962) 'metaphor' stage. Other observers argue that translational

studies necessarily lack external validity – for instance because 'preclinical animal models can never be fully valid [for humans] due to the uncertainties introduced by species differences' (Pound and Ritskes-Hoitinga 2018). Yet while the space for employing precise techniques and new technologies has constantly expanded in field disciplines, and means of recording one-time data have hugely improved, such changes may not in themselves resolve the problems of holistic explanation. For example, using LIDAR remote sensing laser tech to find new archaeology sites has only tended to increase the open questions that archaeologists debate.

In general, the costs for data gathering and assembling any large dataset in the field sciences are also substantially greater than they are for many kinds of experimental work (and all the computational sciences). The conditions for operationalizing research in even roughly comparable ways are often made more demanding by weather and seasonal timing variations, administrative difficulties in securing access to sites in other countries or continents, and problems in operationalizing research processes in reliably professional ways across difficult sites and terrains and using locally available labour or other resources. Consequently, there is a strong rationale for developing field science databases, reusing data wherever possible and ensuring that the creators of datasets receive full recognition. The incentives for opening up datasets in field sciences are also strong ones. For instance, a Canadian neuroscience open data portal now seeks to allow all qualified researchers to help decode mental illnesses even partly in terms of biological mechanisms – and its initial impetus came from a major hospital team acknowledging that 'we can't do it all' (Bahlai et al. 2019; Poline et al. 2023). A possible downside of greater data science/artificial intelligence (AI) use may be that more dataset creators feel that they can analyse larger amounts of data in-house, and so retain only single-site investigation. However, in many field sciences the scope of investigations remains so wide that even with widespread AI use multiple teams need to be enabled and mobilized by creating open data sources.

Field researchers have also argued strongly for the value of pre-registration and enhancing reusability, especially via greater use of registered reports and audits of field studies linked to systematic reviews. Some writers acknowledge that for 'strained researchers' many factors pull in opposed directions: '[So] while ecologists and evolutionary biologists have somewhat improved research transparency over the past decade (e.g. more data sharing), major obstacles remain' (O'Dea et al. 2021). The potential gains are magnified by the closeness of most field sciences to wider applications of great value outside academia. For instance, in conservation science:

> The long-term adoption of open science practices would help researchers and practitioners achieve conservation goals more quickly and efficiently and reduce inequities in information sharing (Roche et al. 2022).

Finally, in ecology and evolutionary biology, critics of closed science have also promoted other strong reasons for OS, as public controversies over using up live or dead animals in 'unnecessary', 'pointless' or purely commercial experiments surged.

> Unreliable research programmes waste funds, time, and even the lives of the organisms we seek to help and understand. Reducing this waste and increasing the value of scientific evidence require changing the actions of both individual researchers and the institutions they depend on for employment and promotion. (O'Dea et al. 2021).

Animal researchers themselves progressively shifted towards accepting that OS methods could most effectively ensure that data from previous studies was better reused, that negative results were effectively communicated and that no unneeded animal deaths or suffering were caused by their work (Muñoz-Tamayo et al. 2022). Thus: 'Open science might be of high value for research in general, but in translational biomedical research, it is an ethical obligation' (Diderich et al. 2022).

3.4 The integrated rationale for OS and its relevance for the social sciences

Early work on how OS has developed across STEMM disciplines suggested that it was 'one term but five schools of thought' (Fecher and Friesike 2013). These typologies partitioned the literature by the chief motivations and tools cited by advocating authors – described as building science infrastructures, growing pragmatic collaboration, diversifying impact measurement, democratizing access to science, or forms of OS stressing 'citizen science'. But these initially separated ideas, which the authors presented as supposedly distinctive to each 'school of thought', are now generally advocated by all proponents of OS. More recently, some authors have sought to distinguish between 'open science' and 'reproducibility' fields, but it is hard to see any worthwhile separation here (Murphy et al. 2020) – nor the differential participation by female authors, who are very active in both sub-areas. Equally, Mirovski's (2018) claim that OS is just a veneer label masking eclectic and unconnected concerns seems thoroughly outdated now.

To better capture the increased common purposes and yet varied character of the contemporary OS movement as a whole, we identify three top-level integrated OS themes with application across the whole of the STEMM discipline group:

1. Concerns of research integrity lapses or adverse research distortions that seek to improve scientific practice via better checking and concern for robustness;

2. Efforts to improve the efficiency of science by making better use of expensive resources (funding, equipment and expertise) and fostering reuse of earlier work; and
3. Changes that 'democratize' innovation and insights by opening STEMM research up to the widest feasible audiences, especially for applied work, where researchers can gain insights from other disciplines, and relevant and well-based 'ordinary knowledge'.

Inside these main categories we distinguish 12 distinct effects in all and we quickly reprise key aspects in STEMM subjects, before considering their applicability (or not) in the social sciences.

Within the integrity theme are four key motives:

1a. *Strengthening integrity in individual research teams* fosters norms and practices to discourage individual-level researcher behaviours that may subvert good work in an intensely competitive, mass-science setting. For instance, strengthening reproducibility efforts helps combat *p*-hacking (manipulating data to make results appear statistically significant) (Head et al. 2015) or harking (hypothesizing after results known) and data-dredging. Normalizing OS practices also reinforces ethical practices, highlighting the seriousness of competitive excesses and lapses before researchers are tempted by them. It also strengthens the chances that any marginal or sub-ethical behaviours used to push articles into publication will subsequently be discovered, potentially at high costs for the individual researchers responsible.

In principle, this broad ambition and the proposed good effects are equally applicable in the social sciences. Yet there is an important qualification. Both human-dominated and human-influenced systems constantly change across space and time and in response to literally thousands of different stimuli simultaneously – whether societal, technological, political, international or changing 'natural' circumstances. Empirical social science research never accesses an invariant 'natural order', the only possible guarantee of 'laws' operating inter-temporally. Exact reproducibility of the same study in the same conditions is accordingly an elusive goal across many social sciences. Proving the non-reproducibility of a previous research team's work is more difficult, is often contested by authors criticized (as the psychology example shows), and may less often reflect badly on them.

1b. *Combatting emergent 'bad science' effects at collective level* caused by intense, flawed or sub-ethical scientific competition seeks to change how scientific institutions operate. An influential Royal Society (2013) report argued that:

> Open science … may also help to deter, detect and stamp out bad science. Openness facilitates a systemic integrity that is conducive to early identification of error, malpractice and fraud, and therefore deters them. But this kind of transparency only works when openness meets standards of intelligibility and assessability – where there is intelligent openness (p. 8).

For instance, encouraging journals to embrace 'registered reports' commits them to publishing the results of studies that are well reviewed at an early stage, whatever the outcomes that the finished research generates. This can help counteract positive results bias, further reducing researchers' incentives for *p*-hacking, and ruling out harking (Chambers 2019). Around 300 STEMM journals have embraced registered reports (Turing Way 2024c).

In principle, this point could gain traction in the social sciences. Registered reports certainly have expanded in some areas of social psychology. However, the unceasing impacts of new innovations, conflicts and behaviours in modern societies renders the potential usefulness (and finiteness) of any large-scale (let alone comprehensive) *inventory of non-results* hard to envisage. Lindblom and Cohen (1979) argued that the full ensemble of truths established by professional social inquiry could only ever amount to a set of isolated (and in themselves not-very-meaningful) pinpricks in a far larger landscape picture – one that can only be understood in any effective or dynamic way by using ordinary knowledge to connect the dots. Of course, this claim is speculative and remains unproven; but so too does any counter to it affirming that social science laws are attainable. No aspect of society is yet effectively describable or predictable in any useful way *solely* using professionally validated or invariant social science propositions.

1c. *Improving reproducibility* aims to counter replication crises sapping trust in research findings. Again, this imperative is fully applicable to the social sciences, as progress in social psychology shows. But will replication work in OSS ever attract the level of attention that it has increasingly sustained in the STEMM disciplines? Again, societal flux has so far been a disincentive for researchers to do more than check on particularly controversial or influential prior research. In addition, most social science research claims have only ever enjoyed a conditional or contested acceptance from western public opinion, even amongst university-educated people or non-academic 'experts' – in part explaining why replication crisis charges have occurred less in many disciplines. For instance, social science research claims are rarely presented by the media as authoritatively established 'findings' or 'discoveries', but rather as just what the researchers 'argued' or claimed (Bastow, Dunleavy and Tinkler 2014, Chapter 8).

1d. *Increasing surveillance* by funders, professions or universities over researchers and mandating OS changes has been more important in STEMM research because of the huge investment of resources entailed, running into tens of billions of dollars in areas like medicine, pharma research, new IT systems, data science/AI, armaments, biogenetics and so on. Also very costly are some STEMM databases of previous research literatures and previously accumulated datasets, and the first-stage exploratory search needed to surface potentially promising areas of work from core resource banks (see e.g. Leonelli 2010, 2020). External interference by well-informed foundations or funders relying on advice from expert committees has also been less controversial in STEMM, since funders generally seek scientific consensus.

There are few remotely parallel whole-field databases in the social sciences, although some multi-author cross-national surveys and relatively standardized systems of national statistics have developed. In general, however, the discipline group still mainly has to rely on highly fragmented, project-specific datasets, some of which OSS advocates have reused and extended, often without much involvement by professional associations or research funders. Resource-orientated OSS pressures by governments/funders have been more evident in quantitative empirical sub-fields (which are generally costlier). Attempts to 'direct' research priorities in social science areas have generally used a lighter touch, but have also been relatively unsuccessful and more controversial in liberal democracies. However, populist politicians and some far-right governments have been much more interventionist – for instance, the Orban government made extensive efforts to 'clean up' (or more aptly censor) Hungarian contemporary history resources, ranging from museums and funded research through to textbooks and statues in public settings (Trencsényi 2021). And Republicans in the U.S. Congress have regularly debarred political science and other disciplines suspected of 'liberal biases' from receiving federal funding, a partisan 'crusader' approach that later expanded to affect a wide sweep of environmental and health research under the second Trump administration.

Within the efficiency subset are three key motives:

2a. *Producing speedier debugging of mistakes* is an increased priority, especially in tech fields like countering cyberattacks, where collaborative responses by many actors can only be achieved via open source approaches. More broadly, though, the speed of science-based changes (such as the adoption of new AI, new materials in construction, or the spread of new medical or illicit drugs), plus the history of past regulatory lags generating huge social costs, has generated societal demands for improving field-level resilience across many STEMM areas.

In the social sciences faster changes now happen in human and social behaviours – such as the viral spread of online social harms like disinformation, online bullying and harassment, new illicit drug uses, malign deployment of AI and robotics innovations by criminals and terrorists, and so on. A constantly and radically changing social world also generates new opportunities for better goods, services and public policies, so long as emergent trends can be understood and regulated in a timely way. OSS is a critical component of a 'forward shift' in the discipline group as a whole, towards studying contemporary and future-trending phenomena. Yet this is not just 'current affairs' or a shift of research towards applied uses. OSS enables faster evolution of the inherently collective research waves that characterize the social sciences – with fewer 'breakthroughs' realistically attributable to single-team innovations than may occur in STEMM disciplines.

2b. *Making more efficient use of research funding and resources* is a far stronger motivation in STEMM disciplines, where some key innovations or centres (like CERN) are only fundable with cross-national funding, and forefront tech and science databases have costs running into hundreds of millions of dollars. Set against the spiralling growth of researcher numbers and journals, public sector austerity pressures strongly boosted funders' insistence on combatting non-use of exisiting resources (Bowers et al. 2023), plus not duplicating past research, reusing existing datasets wherever feasible and increasing the cumulation of work.

The social sciences are not quite at the lowest cost levels in universities, but many departments are well down the list. Dependence on collective resources has historically been less because research is less expensive to do than in most STEMM subjects, and significant theoretical and empirical work is regularly done by academics with only tiny levels of own-funding. Yet because they have much lower funding levels, OSS pressures for efficiency and demands on scant funding are almost as strong as in STEMM subjects, especially in Anglosphere countries. And, with faster, wider, now often globalizing changes across technologies and social behaviours, the research tools needed in social science have grown costlier to implement (despite open software).

2c. *Expanding the realm of in-use knowledge over knowledge not in use* is a key OS driver within social science, but less so in most STEMM disciplines where the model of a rapid research frontier advance with high consensus behind it, discussed in this chapter (Collins 1994), has generally marginalized past understandings. The history of science has expanded as a separate sub-field, recognized non-western sources more (Poskett 2022) and exerted somewhat more collective influence on science advances than in the past. Some citizen science projects have sought to retrieve historical data and understandings to

better chart long-run climate changes or the episodic histories of volcanoes (see Chapter 12). But so far these are relatively small changes and likely to remain so, unless some areas of STEMM or tech efficacy go backwards into crisis – such as the declining effectiveness of antibiotics widely impairing health, or generative AI creating a 'dead internet' flooded by its own made-up outputs rather than human creations (Mariani 2023). Only in such circumstances are STEMM researchers likely to revisit earlier approaches to knowledge for new inspiration.

By contrast, Bastow, Dunleavy and Tinkler (2014) stress the contemporary importance of the social sciences as pervasive inputs into a dynamic knowledge inventory (DKI) vital for civilizational stability. Thanks to digital decompression in the modern era (Dunleavy 2014), the DKI is constantly expanding rapidly, with new social science understandings widely underpinning social and technological changes. The growing STEMM–social science convergence in human-dominated systems charted in Chapter 2 also implies that applied social science ideas may strongly condition issues at the STEMM research frontier. For instance, combatting antibiotic resistance is most likely to be achieved by states reshaping the investment choices for pharma companies to produce 'last resort' drugs (to be deployed sparingly only where conventional antibiotics fail), and stopping use of antibiotics in agriculture through government regulation – rather than by trying to find new antibiotics for mass use (Wellcome 2020, 2023). The border between knowledge that is actively being used or is not also adjusts all the time, with a risk that past knowledge that is decreasingly used can be lost, without research actively retrieving and reconditioning lessons relearned. Maintaining the dynamic knowledge inventory is greatly enhanced under OSS approaches and by some digital era developments, like cheap mass storage and digital archiving (see Chapter 12).

Within the democratization/transparency subset are five key motives:

3a. *Strengthening interdisciplinary understanding* in STEMM has been developed most outside academia by tech firms, other businesses and state agencies building large, purpose-focused teams to do collaborative research that is 'applied and basic combined' and brings together 'science, engineering and design' (Shneiderman 2016, 2019). Academic STEMM has also adapted in parallel ways, albeit more slowly, as university–business interfaces have lengthened and deepened. Yet despite strong research funder pressure, most faculty research remains single-discipline driven, with ancillary inputs, and there are still few high-prestige interdisciplinary journals (Daniel et al. 2020).

Siloed knowledge is just as serious an issue for OSS, since discipline boundaries have shifted little in 120 years (see Chapter 1). Human behaviour cannot

easily be separated out for single-focus studies on experimental lines. (Indeed, experimental social science designs relying on reactive methods may often trigger artefactual adjustments of research subjects' understandings or actions that can 'contaminate' the research data). Changes in the social world are baffllingly shaped by literally dozens or hundreds of variables and circumstances. Siloed, single-subject approaches that arbitrarily marginalize 'out of scope' factors, or rely on simplifying 'rational action' assumptions, are equally open to critique.

OSS opens up huge potential dividends for researcher-led (not funder-mandated) interdisciplinary convergence across the discipline group. Desiloing is fostered best where researchers steeped in one disciplinary base can nonetheless find out easily about alternative approaches, reusing and cumulating data more, checking for reproducibility, and involving research subjects in designing and interpreting studies. In the analysis of human-dominated systems:

> hybrid and cognate disciplines [*such as*] demography, (human/political/economic) geography, anthropology, public administration, social policy, political economy, town planning, urban studies, transport studies, business studies, socio-legal studies, sociolinguistics, social psychology, clinical psychology are prime examples of disciplinary integration. These fields of research (interdisciplinary by definition) make the social sciences ideal partners in projects tackling global societal challenges (Daley and Hantrais 2024).

The same strong momentum for greater STEMM and OSS cross-fertilization also applies in tackling issues in human influenced systems (such as antimicrobial resistance, climate change, or water insecurity) discussed in Chapter 3. A key OSS premium lies in fostering improved, joined-up, societal guidance.

3b. *Building stronger links between scientific/academic knowledge and external expert/professional 'ordinary knowledge'* is central to OS both in STEMM and social science subjects, bringing together academic and external expertise more in applied ways, both within universities and in linked businesses or agencies. Especially important here has been a growing recognition by academics that project-led 'innovation by doing' is increasingly key for applied STEMM and social sciences. Past 'ivory tower' cultural attitudes, with academics looking down on industrial or cultural change-makers, have largely disappeared, even if academics in many top departments remain resolutely wedded to the myth that only academic jobs count as the end goals of doctoral or early career research.

3c. *Widening public engagement with research* to everyone interested in and competent at diverse fields seeks to maximize the talent pool of people able to contribute to continuous knowledge development.

> This might seem a utopian goal, but it is one that digital progress with search engines, open access publishing, open educational resources (OERs), online education (especially free, massive open online courses (MOOCs)), open source software, generative AI and expanded social media in each subject area have all helped make a great deal more achievable. Historically, the development of even very technically challenging STEMM disciplines has often been driven by 'outside' actors coming up through unexpected or hard-to-predict pathways. In advanced industrial societies the professionalization of science has routinized and narrowed domestic pathways into academia. Yet there is a pervasive anxiety amongst scientists about apparent reductions in the proportion of university entrants willing to endure the hard slog of reaching research frontiers. Continuously attracting talented people from across the world has kept established STEMM centres of excellence growing (albeit with adverse consequences for equalizing global academic development across countries). OS changes in STEMM and social science subjects already appear to have contributed to helping maximize the development of talent, wherever people are located in the world.

In addition, OSS is arguably a prerequisite for any effort to boost 'useable knowledge' in recent years, given that even in advanced industrial societies large social groups have demonstrated incredibly strong and embedded tendencies to deny all forms of scientific advances and expert knowledge in favour of populist and actively disinformative counter-beliefs. Improving translational resources for STEMM subjects has had at best limited impacts. But similar efforts in the social sciences could yield larger payoffs, since giving intuitive explanations of social phenomena is easier than with some STEMM knowledge.

3d. *Globalizing access to research to all regions of the world* is directly fostered by open access publishing, OERs and other changes discussed above that mean that anyone appropriately qualified by education or self-effort with an internet connection can in principle access many times the knowledge bases previously available. Such recent changes have yet to work out in any easily measured ways. Of necessity, the OS gains involved have not yet been reliably quantified or metricized quite, even within advanced industrial states (Cole et al. 2024). But the distributional gains of open access publishing alone for less advantaged universities, students, academics and countries worldwide are unambiguously positive for all disciplines (Watkinson 2024).

3e. *Strengthening public belief and confidence in academic research.* All the public understanding benefits of OS changes increasing the robustness and quality of STEMM research are equally applicable

in the social sciences. Indeed, in liberal democracies there may be higher and quicker dividends from OSS, perhaps enabling the public at large to achieve better or greater oversight of powerholders and professionals in government and business.

Conclusions

'When will "open science" become simply "science"?' was a question originally posed when OS was still in its early stages (Watson 2015), and remains hard to answer even now that some early, troubling inequalities in open access publishing and other OS practices have been partly tackled (Bahlai et al. 2019). Even in STEMM subjects it seems that the distinction will not disappear soon. Rather OS is a developing project that is still working its way into all the component areas of the disciplinary group, changing as it does so (Armeni et al. 2021). Yet the sustained progress of an integrated OS conception, and the varied patterns and trajectories charted in this chapter, still hold out many lessons for the social sciences as they too (after a lag) make the transition to greater openness.

Notes

* 'Open source' is widely abbreviated in relevant literature to OS, but we have avoided that usage here – because in this book OS means 'open science'. To avoid any risk of confusing readers here we avoid using OS at all in this sub-section, spelling out the full words 'open source' and 'open science'. However, we do use the acronym FOSS meaning 'free and open source software'.

[1] This chapter draws on research and seminars forming part of the CIVICA Research programme, which was funded in 2022–2024 by the European Union's research and innovation programme Horizon 2020, under the chapter 'Science in Society'.

References

Acharya, Anurag; Verstak, Alex; Suzuki, Helder; Henderson, Sean et al. (2014) 'Rise of the rest: the growing impact of non-elite journals', *arXiv*, 18 October. https://doi.org/10.48550/arXiv.1410.2217

Armeni, Kristijan; Brinkman, Loek; Carlsson, Richard et al. (2021) 'Towards wide-scale adoption of open science practices: the role of open science communities', *Science Public Policy*, vol. 48, no. 5, pp. 605–611. https://doi.org/10.1093/scipol/scab039

Bahlai, Christie; Bartlett, Lewis J; Burgio, Kevin R; Fournier, Auriel et al. (2019) 'Open science isn't always open to all scientists', *American Scientist*, vol. 107, no. 2, p. 78. https://doi.org/10.1511/2019.107.3.78

Baker, David P. and Powell, Justin J.W. (2024) *Global Mega-Science: Universities, Research Collaborations, and Knowledge Production*, Stanford, CA: Stanford University Press. **$** https://perma.cc/2A48-Z87N

Ball, Philip (2021) 'What the COVID-19 pandemic reveals about science, policy and society', *Royal Society blog*, 12 October. https://doi.org/10.1098/rsfs.2021.0022

Bartling, Holman Sönke and Friesike, Sascha (eds) (2014) *Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing*, Berlin: Springer. https://perma.cc/F2RN-B3EH

Bastow, Simon; Dunleavy, Patrick; and Tinkler, Jane (2014) *The Impact of the Social Sciences: How Academics and their Research Make a Difference*. London: Sage. **$** https://doi.org/10.4135/9781473921511

Begley, Glenn C and Ellis, Lee M. (2012) 'Drug development: raise standards for preclinical cancer research', *Nature*, no. 483, pp. 531–533. https://doi.org/10.1038/483531a

Black, Max (1962) *Models and Metaphors*, Ithaca, NY: Cornell University Press. $ https://doi.org/10.7591/9781501741326

Bowers Emma C; Stephenson Jemima; Furlong Melissa; and Ramos Kenneth S. (2023) 'Scope and financial impact of unpublished data and unused samples among U.S. academic and government researchers', *iScience*, vol. 26, no. 7, article 107166. https://doi.org/10.1016/j.isci.2023.107166

Brandt, Mark J; IJzerman, Hans; Dijksterhuis, Ap; Farach, Frank J et al. (2014) 'The replication recipe: what makes for a convincing replication?', *Journal of Experimental Social Psychology*, vol. 50, pp. 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Buranyi, Stephen (2017) 'Is the staggeringly profitable business of scientific publishing bad for science?', *The Guardian*, 27 June. https://perma.cc/78D9-LQ9T

Cartwright, Nancy (1999) *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press. **$** https://doi.org/10.1017/CBO9781139167093

Chambers, Chris D. (2019) 'What's next for registered reports?', *Nature*, no. 573, pp. 187–189. https://www.nature.com/articles/d41586-019-02674-6

Cole, Nicki L; Kormann, Eva; Klebel, Thomas; Apartis, Simon; and Ross-Hellauer, Tony (2024) 'The societal impact of Open Science: a scoping

review. *Royal Society Open Science*, no. 11, article 240286. https://doi.org/10.1098/rsos.240286

Collins, Harry; Evans, Robert; Durant, Darrin; and Weinel, Martin (2020). 'What is science?', in Collins, Harry; Evans, Robert; Durant, Darrin; and Weinel, Martin (eds) *Experts and the Will of the People: Society, Populism and Science*, London: Palgrave, pp. 47–62. https://doi.org/10.1007/978-3-030-26983-8_5

Cologna, Viktoria; Mede, Niels G; Berger, Sebastian; Besley, John; Brick, Cameron et al. (2025) 'Trust in scientists and their role in society across 68 countries', *Nature Human Behaviour*, vol. 9, pp. 713–730. https://doi.org/10.1038/s41562-024-02090-5

Conrad, Tim O. F; Ferrer, Eloi; Mietchen, Daniel; Pusch, Larissa et al. (2024) 'Making mathematical research data FAIR: pathways to improved data sharing', *Scientific Data*, vol. 11, article 676. https://doi.org/10.1038/s41597-024-03480-0

Daniel, Kristy L; McConnell, Myra; Schuchardt, Anita; and Peffer, Melanie E. (2020) 'Challenges facing interdisciplinary researchers: findings from a professional development workshop', *PlosOne*, vol. 17, no. 4, article: e0267234. https://doi.org/10.1371/journal.pone.0267234

Demortain, David (2011) *Scientists and the Regulation of Risk: Standardizing Control*, Cheltenham: Elgar.

Dunlap Riley E; McCright Aaaron M; and Yarosh Jarrod H. (2016) 'The political divide on climate change: partisan polarization widens in the US', *Environment: Science and Policy for Sustainable Development*, vol. 58, pp. 4–23. **$** http://dx.doi.org/10.1080/00139157.2016.1208995 Open access: https://perma.cc/QS96-LR8L

Dunleavy, Patrick and Tinkler, Jane (2021) *Maximizing the Impacts of Academic Research: How to Grow the Recognition, Influence, Practical Application and Public Understanding of Science and Scholarship*, London: Palgrave (now Bloomsbury Press). **$** https://www.bloomsbury.com/uk/maximizing-the-impacts-of-academic-research-9781352010992/

Event Horizon Telescope Collaboration (2022a) 'First Sagittarius A* Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole in the Center of the Milky Way', *The Astrophysical Journal Letters*, vol. 930, no. 2. https://doi.org/10.3847/2041-8213/ac6674

Event Horizon Telescope Collaboration (2022b) 'Astronomers reveal first image of the black hole at the heart of our galaxy', Event Horizon Telescope, Blog post, 12 May. https://perma.cc/47ML-UGXZ

Fanelli, Danielle (2017) 'Is science really facing a reproducibility crisis, and do we need it to?', *PNAS*, 3 November, vol. 115, no. 11, pp. 2628–2631. https://doi.org/10.1073/pnas.1708272114

Fanelli, Danielle (2022) 'Is science in crisis?, in Jussim, L., Krosnick, J.A., Stevens, S.T. (eds), *Research Integrity: Best Practices for the Social and Behavioral Sciences*, Oxford: Oxford University Press, pp. 93–116.

Fang, Ferric C. and Casadevell, Arturo (2011) 'Retracted science and the retraction index', *Infection and Immunity*, October, vol. 79, no. 10, pp. 3855–59. https://doi.org/10.1128/iai.05661-11

Fecher, Benedikt and Sascha Friesike, Sasche (2013) *Open Science: One Term, Five Schools of Thought*, Berlin: SCIVERO Press, German Data Forum, RatSWD Working Paper 218. https://perma.cc/S83A-A4S5 Also Chapter 2, pp. 17–49 in Bartling and Friesike (2014). https://doi.org/10.1007/978-3-319-00026-8

Dr Foster (2024) Dr Foster website. https://openscience.eu/article/project/foster

Geiger, Roger L. (2017) *To Advance Knowledge: The Growth of American Research Universities, 1900–1940*. London: Routledge. **$** https://doi.org/10.4324/9781315135694

Hancheng, Cao; Dodge, Jesse; Loc, Kyle; McFarland, Daniel A.; and Wang, Lucy Lu (2023) 'The rise of open science: tracking the evolution and perceived value of data and methods link-sharing practices', *arXiv*:2310.03193. https://perma.cc/6H8Y-W9KD

Head, Megan L.; Lanfear, Luke; Kahn, Rob; Jennions, Andrew T.; Jennions, Michael D. (2015) 'The extent and consequences of *p*-hacking in science'. *PLoS Biology*, vol. 13, article e1002106 https://doi.org/10.1371/journal.pbio.1002106

Higginson, Andrew D and Munafò, Marcus R. (2016) 'Current incentives for scientists lead to underpowered studies with erroneous conclusions', *PLoS Biology*. vol. 14, article: e2000995. https://doi.org/10.1371/journal.pbio.2000995

Ionnidis, John (2005) 'Why most published research findings are false', *PLOS Medicine*, vol. 2, no. 8, article e124. OA at: https://perma.cc/4LT6-GWMW

Kitchin, Rob (2014) 'Big data should complement small data, not replace them'. *LSE Impact blog*, 27 June. https://perma.cc/7GBH-BVDN

Kuhn, Thomas S. (2012) (1963) *The Structure of Scientific Revolutions – 50th Anniversary Edition*. Chicago: University of Chicago Press. Originally 1963. **$** https://perma.cc/ABR4-C8G3

Leonelli, Sabina (2010) 'Documenting the emergence of bio-ontologies: or, why researching bioinformatics requires HPSSB', *History and Philosophy of the Life Sciences*, vol. 32, no. 1, pp. 105–125. https://perma.cc/PA4F-Q36R

Leonelli, Sabina (2012) 'When humans are the exception: cross-species databases at the interface of biological and clinical research', *Social Studies of Science*, vol. 42, no. 2, pp. 214–236. $ https://doi.org/10.1177/0306312711436265

Leonelli, Sabina (2023) *Philosophy of Open Science*, Cambridge: Cambridge University Press.

Leonelli, Sabina (2020) 'Scientific research and big data', in *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), Zalta, Edward N. (ed.). https://perma.cc/WC8G-ZJMW

Lindblom, Charles E and Cohen, David K. (1979) *Useable Knowledge: Social Science and Social Problem Solving*, New Haven: Yale University Press. $ https://perma.cc/4YWQ-F8UE

McLaren, Duncan and Markusson, Nils (2020) 'The co-evolution of technological promises, modelling, policies and climate change targets', *Nature Climate Change*. https://doi.org/10.1038/s41558-020-0740-1 OA at: https://perma.cc/UV7A-MKPV

McNeely, Ian and Wolverton, Lisa (2009) *Reinventing Knowledge: From Alexandria to the Internet*. New York: W.W. Norton.

Mariani, Robert (2023) 'The dead internet to come', *The New Atlantis*, no. 73, pp. 34–43. https://www.jstor.org/stable/27244117

Mayden, K. (2012) 'Peer review: publication's gold standard', *Journal of Advanced Practical Oncology*, vol. 3, no. 2, pp. 117–22. https://perma.cc/4DRQ-QUK

Mirowski, Philip (2018) 'The future(s) of open science', *Social Studies of Science*, vol. 43, no. 2, pp. 171–203. https://doi.org/10.1177/0306312718772086

Moody, Glyn (2002) *Rebel Code: Linux and the Open Source Revolution*, New York: Basic Books. https://www.hachettebookgroup.com/titles/glyn-moody/rebel-code/9780738206707/

Moody, Glyn. (2019) 'Open science means open source-or, at least, it should', *Experfy Insights* blog, 13 May. https://www.experfy.com/blog/bigdata-cloud/open-science-means-open-source-or-least-it-should/

Morris, Peter J. (2021) 'The history of chemical laboratories: a thematic approach', *ChemTexts*, vol. 7, no. 3, p. 21. $ https://doi.org/10.1007/s40828-021-00146-x OA at: https://pmc.ncbi.nlm.nih.gov/articles/PMC8218977/

Mullard, Asher (2021) 'Half of top cancer studies fail high-profile reproducibility effort', *Nature News*, 9 December. $ https://perma.cc/T3AP-43Y9

Mullen (2024) *Open Access, Scholarly Communication, and Open Science in Psychology: An Overview for Researchers*, Sage Open Long Form: Open Access Monographs, SAGE Open, January–March 2024, pp. 1–123. https://doi.org/10.1177/21582440231205390

Muñoz-Tamayo, Rafael; Nielsen, Birte L; Gagaoua, Mohanned; Gondret, Florence et al. (2022) 'Seven steps to enhance Open Science practices in animal science', *PNAS Nexus*. vol. 1, article pgac106. https://doi.org/10.1093/pnasnexus/pgac106

Murphy, Kevin R. (2022) 'Descriptive statistics are essential to making complex analyses useful', *LSE Impact blog*, 25 April. https://perma.cc/EU58-V55J

Murphy et al. (2020) 'Open science, communal culture, and women's participation in the movement to improve science', *PNAS*, vol. 117, no. 39, pp. 24154–24164. https://doi.org/10.1073/pnas.1921320117

Nature Materials (2005) 'The cost of salami slicing', *Nature Materials*, vol. 4, no. 1. https://doi.org/10.1038/nmat1305

O'Dea, Rose E.; Parker, Timothy H.; Chee, Yung E. et al. (2021) 'Towards open, reliable, and transparent ecology and evolutionary biology', *BMC Biology,* vol. 19, article 68. https://doi.org/10.1186/s12915-021-01006-3

Pennington, Charlotte (2023) *A Student's Guide to Open Science: Using the Replication Crisis to Reform Psychology*, London: Open University Press/McGraw Hill. **$** https://perma.cc/5FQM-NZN9

Phillipps, Axel (2022) 'Research funding randomly allocated? A survey of scientists' views on peer review and lottery, *Science and Public Policy*, vol. 49, no. 3, pp. 365–377. https://academic.oup.com/spp/article/49/3/365/6463570

Poline, Jean-Baptiste; Das, Samir; Glatard, Tristan; Madjar, Cecile et al. (2023) 'Data and tools integration in the Canadian Open Neuroscience Platform', *Science Data*, vol. 10, p. 189. https://doi.org/10.1038/s41597-023-01946-1

Popper, Karl R. (1979) *Objective Knowledge: An Evolutionary Approach*. Oxford: Oxford University Press.

Poskett, James (2022) *Horizons: A Global History of Science*, London: Penguin. **$** https://perma.cc/3PG5-RVDV

Pound, Pandora and Ritskes-Hoitinga, Merel (2018) 'Is it possible to overcome issues of external validity in preclinical animal research? Why most animal models are bound to fail', *Journal of Translational Medicine*, vol. 16, p. 304. https://doi.org/10.1186/s12967-018-1678-1

Ramella, Francesco and Manzo, Cecilia (2021) *The Economy of Collaboration: The New Digital Platforms of Production and Consumption*, London: Routledge. OA at: https://perma.cc/4V8W-26Q9

Raymond, Eric S. (1998) 'Goodbye, "free software"; hello, "open source"'. Web page, 8 February. http://www.catb.org/esr/open-source.html

Rekker, Roderick (2021) 'The nature and origins of political polarization over science', *Public Understanding of Science*, vol. 30, no. 4, pp. 352–368. https://doi.org/10.1177/0963662521989193

Resnik, David B. (2006) 'Openness versus secrecy in scientific research', *Episteme* (Edinburgh), vol. 2, no. 3, pp. 135–147. OA at: https://perma.cc/NE9Z-J3SX

Roche, Dominique G.; O'Dea, Rose E.; Kerr, Kecia A.; Rytwinski, Trina et al. (2022) 'Closing the knowledge action gap in conservation with open science', *Conservation. Biology*, vol. 36, article e13835. https://doi.org/10.1111/cobi.13835

Royal Society (2013) *Sciences as an Open Enterprise*. London: Royal Society. Report no. 3. https://perma.cc/R8GB-U6QX

Royal Society (2018) 'Integrity in practice toolkit', Web page, August. https://perma.cc/G6HN-AERZ

Ruff, Kathleen (2015) 'Scientific journals and conflict of interest disclosure: what progress has been made?' *Environmental Health*, vol. 14, no. 45. https://doi.org/10.1186/s12940-015-0035-6

Rule, Adam; Tabard, Aurelian; and Hollan, James D. (2018) 'Exploration and Explanation in Computational Notebooks', in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ACM, Montreal, QC, Canada, pp. 1–13. https://doi.org/10.1145/3173574.3173606

Santoro, Helen (2022) 'Open science is surging. Open science is becoming the norm in psychology – a trend spurred on by the COVID-19 pandemic', *American Psychological Association*, *Monitor on Psychology*, January/February. https://perma.cc/6ENJ-NHBW

Shneiderman, Ben (2016) *The New ABCs of Research: Achieving Breakthrough Collaborations*, New York: Oxford University Press. **$** https://global.oup.com/academic/product/the-new-abcs-of-research-9780198812173

Shneiderman, Ben (2019) 'Creativity and collaboration: revisiting cybernetic serendipity', *Proceedings of the National Academies of Science*, vol. 116, no. 6, pp. 1837–1843. https://doi.org/10.1073/pnas.1807200116

Simon, Herbert A. (2019) *The Sciences of the Artificial* (1st edn 1969), Boston: MIT Press. https://doi.org/10.7551/mitpress/12107.001.0001

Smaldino, Paul E and McElreath, Richard (2016) 'The natural selection of bad science', *Royal Society Open Science*, vol. 3, no. 9, article 3160384. http://doi.org/10.1098/rsos.160384

Soneryd, Linda and Sundqvist, Göran (2023) *Science and Democracy: A Science and Technology Studies Approach*, Bristol: Bristol University Press. **$** https://perma.cc/6LW3-F7ET

Stoddart, Charlotte (2016) 'Is there a reproducibility crisis in sciences?', *Nature*, 25 May. https://www.nature.com/articles/d41586-019-00067-3

Tennant, Jonathan P; Waldner, François; Jacques, Damien C; Masuzzo, Paolo et al. (2016) 'The academic, economic and societal impacts of open access: an evidence-based review', *F1000Research*, vol. 5, p. 633. https://doi.org/10.12688/f1000research.8460.3 OA at: https://pmc.ncbi.nlm.nih.gov/articles/PMC4837983/

Trencsényi, Balázs (2021) '"Politics of history" and authoritarian regime-building in Hungary', in May, Neils F and Maissen, Thomas (eds) *National History and New Nationalism in the Twenty-First Century – A Global Comparison*. London: Routledge. **$** https://www.routledge.com/National-History-and-New-Nationalism-in-the-Twenty-First-Century-A-Global-Comparison/May-Maissen/p/book/9780367520410

Turing Way (2024a) 'Overview of reproducible research'. https://perma.cc/6UYE-48HT

Turing Way (2024b) 'Added advantages'. https://perma.cc/THG6-QM72

Turing Way (2024c) 'What are registered reports?'. https://perma.cc/D67N-6XWN

Valentinuzzi, Max and Kohen, Alberto (2013) 'The mathematization of biology and medicine who, when, how?', *IEEE Pulse*, vol. 4, no. 1, pp. 50–56. https://doi.org/10.1109/MPUL.2013.2228591 OA at https://perma.cc/QTJ2-SA6F

Vance, J.D. (2025) Post on X, 24 May. https://perma.cc/2FK7-2DXZ

Watkinson, Charles (2024) 'University Press redux, Plenary 4; Reclaiming Open Access for the University Press Community', University of Michigan Press presentation to Association of Learned and Professional Society Publishers Conference, University of Edinburgh, 15–16 May. https://perma.cc/6BK6-2G7Z

Watson, Mick (2015) 'When will "open science" become simply "science"?' *Genome Biology*, vol. 16, no. 101. https://doi.org/10.1186/s13059-015-0669-2

Watson, Peter (2011) *The German Genius: Europe's Third Renaissance, the Second Scientific Revolution and the Twentieth Century*, London: Simon and Schuster. **$** https://www.simonandschuster.co.uk/books/The-German-Genius/Peter-Watson/9781416526155

Wellcome (2020) *The Growing Crisis in Antibiotic R&D: Opportunities for G20 Member Action to Support Sustainable Innovation*, Online discussion paper. https://perma.cc/6C2P-MLME

Wellcome (2023) 'Why is it so hard to develop new antibiotics?', Blog post, 17 November. https://perma.cc/5TKK-Z22J

Wethington, Elaine and Dunifon, Rachel E. (eds) (2012) *Research for the Public Good: Applying the Methods of Translational Research to Improve Human Health and Well-being*, Washington, DC: American Psychological Association. **$** https://doi.org/10.1037/13744-000

Wikipedia (2023) 'Open science', web page. https://perma.cc/F686-A4BJ

4. Pre-registration for confirmatory and exploratory research

The evolution of open science in science, technology, engineering, mathematics and medicine (STEMM) disciplines suggests that adopting a broad pre-registration approach can play an important role in enabling reproducibility, by helping to better specify the context, expectations and parameters of the research ahead of it being conducted. This move can improve the statistical rigour of quantitative work, better surface (and hopefully codify) tacit knowledge that has shaped the research question, and enable study findings to be evaluated more reliably in relation to the original research questions. This logic has been strongly accepted in some quantitative social science subjects, notably in social psychology and economics, where it fits well with hypothetico-deductive modes of theorizing and testing. For qualitative research, pre-registration also provides safeguards for maintaining a clear 'discovery chain' linking back to theory and safeguards against ad hoc or unself-conscious changes in concepts and methods by researchers (Safra 2022a, 2022b). Yet its adoption has been slowed or partial because there seems (at first sight) to be a contradiction between fully specifying hypotheses and undertaking exploratory work that is more typical of qualitative disciplines.*

In this chapter we seek to show that in all social sciences there is scope for both confirmatory work where pre-registration makes most sense, and for exploratory work, where it still has a role. We begin by examining the contribution that pre-registration approaches offers towards building more open social science (OSS) research in an overall hypothetico-deductive theory–test–evidence sequence, and where the strategy can operate successfully (Section 4.1). We also look more specifically at the most strongly defined confirmatory data analysis (CDA) expectations for new research, which apply in more experimental and rigorous quantitative work (Section 4.2). Next we look at exploratory data analysis (EDA) approaches where some elements of 'hard-core' pre-registration procedures are not feasible because the theory involved

How to cite this book chapter:

Dunleavy, Patrick and Monteath, Timothy (2026) 'Pre-registration for confirmatory and exploratory research', in Dunleavy, Patrick and Monteath, Timothy (2026) *Doing Open Social Science: A Guide for Researchers.* London: LSE Press, pp. 79–104. https://doi.org/10.31389/lsepress.dos.d

is less defined, social change has outpaced theory understandings, test propositions are multiple, or the data being used has already been generated (Section 4.3). Nonetheless, more varied forms of pre-registration are feasible in ways that build more OSS research and aid researchers in developing their work. Finally, we look at qualitative research that uses inherently exploratory or 'soft' theories and methods (some discussed in more detail in Part III). Even here we argue that it is valuable for analysts to pre-define their approach, the methods to be used and at least some expectations about findings, and then log any departures from these as the research progresses (Section 4.4). Overall, we argue that there is no inevitable tension between broadly construed pre-registration approaches and effective exploratory research.

4.1 The scientific method and pre-registration

Many social scientists agree with the STEMM view that 'science' must test potentially falsifiable hypotheses which are supported only so long as they survive attempts to refute them. In its ancient Greek origins, the term hypothesis means a base or foundation. More specifically, it denotes a precise empirical proposition or prediction, derived from a given, well-developed theory, and formulated so as to make practical comparisons with empirical evidence in particular conditions. This is especially the case for any quantitative research where the validity of a hypotheses is evaluated through statistical measures. In the hypothetico-deductive model, research hypotheses are distinctive predictions that are derived from a recognized theory, which should be logically (deductively) developed and self-consistent. Only then will the falsification or disproving of a logically entailed test distinctive to that theory clearly bear against the theory being seen as valid or plausible. Equipped with a set of relevant hypotheses (because a single proposition test is rarely illuminating enough), (social) scientists also seek to test their theories against empirical evidence.

In recent years, more attention has been given to more inductive Bayesian models of scientific endeavour where a deductive model is not seen as essential for advancing empirical insights. Instead scientists are seen as using evidence to assess competing interpretations of what is going on in complex processes, first formulating a set of 'priors' (expectation assumptions) derived from theory or from previous empirical work, and then using data analysis to update the priors in ways that fit with the data pattern, advancing the probability or weight to be assigned to some propositions, and reducing that of others.

What can go wrong with this model? First, in the hypothetico-deductive strands, the underlying theory may be underdeveloped, without a defined logical structure and thus potentially capable of generating multiple possible hypotheses consistent with the 'theory'. Alternatively, the theory may be set out in ways that are incapable of generating hypotheses specific enough to count as falsifying it. Researchers may also 'defend' their core hypotheses by insulating them from disproof by adding 'bolt on' modifications that make

core propositions more complex or less testable (O'Boyle et al. 2016). Subsidiary propositions may also be created and tested in ways that build up layers of 'insulation from attack' for the core theory or model (Lakatos 1970). In the Bayesian vein there may be multiple potential priors, and the evidence may suggest that the same outcomes can be generated in several different ways, no one of which distinctively speaks to the viability of one model alone.

More generally, quantitative research relies on analysts being able to sufficiently control for the conditions of a useful test to be sustained, either by setting precise laboratory setups or using randomized control trials in field research conditions, or by conducting a multivariate analysis within which all potentially relevant other interpretations are filtered out by the use of control variables. Omitted variable bias arises where the control variables used in a multivariate analysis are insufficient to rule out other relevant factors being involved in an apparent association between variable *X* and the dependent variable *Y*. In addition, testing may fail because the terms in which the explanatory theory or priors motivating research are defined may be so vague that they incorporate many different phenomena or outcomes, so that it is hard to see what could count as a falsification of it. However, some quantitative approaches work more inductively, using statistical analysis to sift for latent structures in data that might then become defined variables around which hypotheses can later be constructed.

Alternatively, theories or measurement indices derived from them can rather easily incorporate some circular elements that make them effectively tautological. For instance, a very crude version of evolutionary theory might posit 'the survival of the fittest' species in biology, or of well-adapted behaviour patterns in evolutionary biology and psychology, but then determine fitness or adaptation itself only by a species or a behaviour pattern having survived, and 'unfitness' by its having died out. Much better predictive propositions are needed to support the central thesis here (Hunt 2014). (By contrast, some advocates now argue that evolution is not a theory but a 'natural law' with such detailed and necessary underpinning in genetics that an ability to make specific predictions is not needed (Crouch and Bodmer 2024)).

Second, a debased form of the hypothetico-deductive model can be followed where the hypotheses to be tested are not really derived from any theory, nor specified in a distinctive way, but are simply stand-alone, potentially plausible empirical propositions, plucked from the ether of current academic debates or practical discussions. This may be a particular problem in rapidly changing social or technical processes, where 'sense-making' is still going on (with some parallels to doctors' intuitions playing a key role in clinical reasoning in medicine). Equally, inductive models can be formulated from previous experience that lacks a developed theoretical rationale. Here a 'null hypothesis' of the form 'variable *A* shapes (e.g. increases or decreases) the undertaking of activity *X* (other things being equal)' is tested with a dataset. Using the huge apparatus of confirmatory methods developed in modern statistics, the extent and direction of the association between variable *A* and phenomenon *X* can

be specified in quantitative terms, subject to a set of controls designed to operationalize the *ceteris paribus* (all other things being equal) clause.

How a hypothesis is formulated and precisely how it is seen as tested can make a great deal of difference to the usefulness of the evidence generated. Often, apparently disconfirming evidence has led social scientists who are exponents of a given model or interpretation to simply respecify what it implies. A cumulation of amendments may incorporate so many qualifications or tweaks into the original hypothesis that it is hedged around with caveats sufficient to explain away almost any disconfirming evidence, or elements of tautology have been added in measuring effects. Below gives a good meta example of this process in action with Duverger's Law (DL), a central proposition (sometimes called 'the only law') in political science.

Figure 4.1: Some difficulties in specifying a precisely testable and theoretically coherent null hypothesis – the example of 'Duverger's law' of electoral competition

In the mid-1950s the French political scientist Maurice Duverger advanced a 'law' (Duverger's Law, or DL, named after him) that countries using plurality rule elections in local seats (where the candidate with most votes wins, irrespective of whether they have an overall majority) would become or remain two-party systems. He argued that voters' experience of small parties standing and winning few or no seats locally, even when they piled up substantial national vote numbers, would put citizens off 'wasting their vote' on candidates from parties outside the top two parties. Subsequently, the proposition was extended beyond the misleadingly labelled 'first past the post' elections (there is no post) used in the US, UK, India, etc. to include other majoritarian systems (such as those in Australia and France under the Fifth Republic). Duverger (1964) also advanced a more tentative hypothesis that countries with proportional representation (PR) voting would move towards and remain multiparty systems (Duverger's hypothesis, DH). These propositions were enormously influential with both political scientists and politicians in the US, UK, other Anglosphere democracies and France. They were cited extensively and initially seemed to fit well with these countries' experiences.

Critics argued that both DL and DH were too vaguely formulated to be testable and relied only on 'impressionistic' or anecdotal evidence. In addition, during the 1960s and '70s voting for third parties increased strongly in Canada (for a regionalist Quebecois party), and in the 1970s for third and even fourth parties nationally in Great Britain (with nationalists in Scotland and Wales) and Canada. From the 1980s the number of separate parties in India also enlarged and later exploded, as the post-independence dominance of the Congress party faded,

Figure 4.1: (continued)

fragmenting into a complex multiparty system with many regionalist parties. Defenders of the Duverger propositions at first attempted to add a qualification to the law to predict two-party politics 'except in the case of regionalist social cleavages', a bolt-on widely denounced as an obvious attempt to evade refutation. It anyway did not explain enduring support for non-regionalist third parties such as the Liberals (later Liberal Democrats) in Britain or the NDP in Canada.

Defenders of the DL subsequently argued that it should only apply at the level of *individual seat contexts*, and that properly reformulated in this way it in fact made no fixed predictions about party competition at national level. Thus a multiplication of parties nationally ceased to be a problem for DL. And, as disaggregated data became available for hundreds and then thousands of local contests across countries, DL exponents argued that empirical analyses showed far more 'two-party dominance' (or sometimes one-party dominance) in plurality-rule countries than was the case in PR systems. Critics replied that the concept of dominance in local contests was too vaguely specified to be refuted.

A final evasive manoeuvre designed to counteract this criticism was that DL exponents adopted a particular weighted index of the number of parties (the effective number of parties (ENP) votes). They used it to predict as their central testable hypothesis that in plurality or majoritarian systems the ENPv score statistic would be below an arbitrarily chosen level of 2.5 parties in the vast majority of local contests (and the mean level would be so also across all contests). Alternatively, the claim was that plurality rule local contests would far more often be below 2.5 than would local contests in PR systems. Empirical support for these propositions has been widely claimed in top political science journals.

However, critics now argued for an omitted variable bias effect – pointing out that the relevant ENPv statistic to test depends on the number of parties/candidates standing in local contests – for example, if only two candidates stand then the ENPv statistics logically *cannot* be higher than 2.0, however voters behave. And if three candidates stand there is an 85% logical possibility that the resulting ENPv will be 2.5 or below. So far only defective H_1 tests have been conducted, often including numerous constituencies with one, two or three parties competing in the local data for plurality systems, whose results are autocounted as 'DL compliant'. A theoretically valid H_1 test would instead need to show that the local ENPv scores in plurality rule or majoritarian systems are systemically lower in statistically significant ways than those in PR contests *with the same numbers of parties or candidates competing in*

(*continued*)

Figure 4.1: (continued)

local seats. No such evidence has yet been produced, but the 'law' is still given widespread credence after nearly 70 years of its being repeatedly reframed to evade falsification.

Meanwhile, India now has some 40+ separate parties in its parliament, and the UK, Canada and Australia all have permanent multiparty competition at elections. Only the USA retains near-perfect two-party dominance – but critics of DL argue that this again reflects omitted variable bias in previous tests. All U.S. local contests reflect the strong impacts of partisan 'electoral engineering' by incumbent Democrats and Republicans in U.S. state legislatures that they control. Both punish third parties in cartel mode and each shapes or gerrymanders state legislature and congressional districts and voting access rules in ways favourable to the incumbent party. In addition, these effects are reinforced by the impact of voters' experiences of one-shot presidential contests, where the Electoral College means that third parties stand absolutely no chance of ever winning. Thus U.S. patterns overwhelmingly reflect strongly partisan election engineering at work, rather than necessarily reflecting voters' reluctance to support third party candidates.

To explore this case further, see Clark and Golder. (2006); Cox (1977); Dickson and Scheve (2007); Dunleavy and Diwakar (2013); and Singer (2013).

The overall rationale for pre-registration

Pre-registration involves clearly formulating and then saving and recording in a formal way key elements of a research plan *before* that research is conducted. The most important thing is to write down, in a manner that is (potentially) publicly available, the set of hypotheses that the research will test, using a process of online registration. A second key element that can be pre-registered in the same way is the variable list to be used, especially for control variables used to stand up 'other things being equal'. This step guards against later impromptu or ad hoc additions of new variables into the mix, so potentially changing the nature of the analysis conducted. A third key element that researchers have pre-registered concerns the scope and nature of the data to be used. This guards against the researcher later changing the analysis by broadening or narrowing the range of data cases considered in ways that may tend to accentuate the chances of favoured hypotheses surviving falsification. In all three aspects, developing the research plan ahead of the research being conducted is important for the integrity of the research and for the statistical validity of the analysis to be undertaken. Saving the research plan as a digital document is the crucial step because it attaches a time stamp that evidences when it was developed.

Many platforms now exist that allow researchers to register their pre-analysis plans (PAPs) and create a date-stamped document, such as the Open Science Foundation site, osf.io. The exact platform that researchers choose to use, and the features offered by the platforms, vary by discipline. Registering a PAP on an open science (OS) platform does not necessarily mean that the plan must be made publicly available at once or in future. Researchers may choose to immediately make public their PAP or some components, or they may make it public later on in the research process. Alternatively, they may hold it initially as a private document, but one that can nonetheless be released to other researchers seeking to reproduce their research findings. Many researchers choose to keep PAPs private until the research has been conducted and disseminated or published, so that they can continue to compete for grants and go about getting published without worrying that their innovations or detailed designs will be 'scooped' or copied in detail by fleeter-footed competitors. A plan that has been time-stamped and archived protects these interests, and it can also be made available to reviewers or other researchers at a later date where it is appropriate to do so.

In those STEMM disciplines closest to the social sciences, the pre-registration of hypotheses and use of PAPs became widespread in the medical sciences alongside the use of more rigorous randomized control trials (RCTs) to test and develop new pharmaceutical treatments in clinical trials and obtain reliable findings when testing new drugs on patients. For scientists developing new drugs (a very lengthy and costly process), the registration and publication of information on clinical trials is important at every stage of the research process – before they commence, during their administration, and in disseminating their findings once ended. There are multiple benefits. Publishing a clinical trial plan before it commences sets in stone how the study will be run, what data will be collected, and the statistical methods and measures to be used to assess its success. The open registration of all clinical trials creates an incentive to publish null results along with the relevant data and analysis. This may help overcome 'publication bias' effects where only positive or novel results are considered worthy of scientific publication by journal editors or reviewers (see Chapter 3). Finally, the publication of past results can help improve the macro-efficiency of research and reduce patient harm. Running a clinical trial that tests a hypothesis already proven to be false may waste significant resources on duplicating work already done. It may also create a risk of active harm for the patients involved in the trial, who must forfeit other, better treatments. Furthermore, the pre-registration system for clinical trials requires that any adverse effects to patients are reported as soon as they are found, while trials are still ongoing. So if potential harms to patients are identified, the risks of them reoccurring can be minimized. Full disclosure also means that problems that may arise out of similar or future research can be anticipated and prevented more speedily.

The practice of pre-registering has since expanded into quantitative social research – particularly in health studies, psychology and economics, where

RCTs or experimental work are feasible. Many of the reasons that social scientists have cited for the growth of pre-registration mirror the debates that popularized their use in medical science. The most important motivation has been 'replication crisis' worries affecting many areas of quantitative research. Pre-registering hypotheses and PAPs have been advocated as ways in which the reproducibility – and therefore reliability and validity – of research results can be improved. In addition, pre-registration is increasingly seen in quantitative fields as a matter of good scientific practice that helps researchers think more carefully about the designs of their studies at an early stage. Amidst the many other pressures that academics face, getting investigators to engage in detail with their research plan early on and in an extended way, in a manner culminating in a formal outcome (the acceptance of the pre-registration) can be a productive exercise.

However, unlike clinical trials and RCTS in medicine, there has so far been no standardization in social science disciplines about what a useful PAP should contain. Consequently, researchers have a lot of flexibility about what they record, and the scope and usefulness of registered PAPs varies considerably. That said, a number of features do recur across most PAPs, which Section 4.2 discusses in greater depth, exploring both the requirements of each feature and the main goals that each of feature seeks to accomplish. The examples used are for null hypotheses significance testing (NHST), one of the most widely used quantitative frameworks in (social) scientific research. However, it should be noted that the use of PAPs is not limited to these methods.

4.2 Confirmatory data analysis (CDA)

The primary method for evaluating the validity of quantitative data in the social sciences (as well as many fields of bio-medicine) has been (and remains) NHST. To use this statistical measure requires that the researcher generates two specific forms of research questions. These are:

- a null-hypothesis – often referred to as the H_0; and
- an alternative hypothesis, known as the H_1.

Say a researcher is looking to prove a causal connection between two variables of interest, with variable *X* shaping *Y* – here a null hypothesis would state that there is no statistically significant relationship between the two variables. The alternative hypothesis would state that *X* partly causes or shapes *Y* in the light of a theory about why this should happen. When conducting NHST the point of the research is to disprove the null hypothesis – that is to demonstrate that there is a statistically significant relationship between the variables of interest. The calculation of this significance is tested against random variability in the

data – the alignment between the variables of interest must be shown to occur more than could be expected simply by chance or because of random variation or static in the data.

Critics of both the hypothetico-deductive and inductive versions of hypothesis testing, heavily reliant on quantitative methods, have pointed out that two types of failure can occur, shown in Figure 4.2. The first (type I error) occurs in the bottom row yellow-shaded cell where a null hypothesis that is true (*X* and *Y* are not related) is rejected, creating a 'false positive' result tending to lend support to retaining the alternative proposition H_1. The second type of failure (type II error) occurs in the top right yellow-shaded cell of the figure, when *X* does indeed shape or is related to *Y* but the analyst rejects the link as 'noise' in the dataset, one of many potential patterns that turn out on inspection to be only random variations in the data.

In specifying the hypotheses a threshold criterion – known as the significance level or α – needs to be set, against which the *t*-value is compared. The level of α is the threshold which researchers deem acceptable as the level for type I errors. In the social sciences this is widely accepted (by convention) as 5% (or two standard deviations) and corresponds to a computed *p*-value of less than ≤ 0.05. This threshold is thus usually not explicitly justified as part of the hypothesis. Instead, increased sample size increases the overlap between the distributions being measured, therefore reducing β and decreasing type II errors. Indeed, the hypotheses can be designed to reduce the ratio between mean differences and standard deviations to reduce β, thereby increasing its 'power' and reducing type II errors. It is, furthermore, possible to take this one step further – if it is possible to know or estimate population sizes and distributions – by computing a priori a power analysis to help design a set of

Figure 4.2: Two types of error in hypothesis testing

		Null hypothesis (H_0) is in fact	
		True	False
Analyst's view of the null hypothesis (H_0)	Does not reject H_0	*Correct decision* so H_1 rejected (probability = $1-\alpha$)	*False negative for* H_1 a type II error (probability = β)
	Rejects H_0	False positive for H_1 a type I error (probability = α)	*Correct decision* so H_1 not rejected (probability = $1-\beta$)

Note: The substantive proposition tested is H_1, claiming that explanatory variable *X* has an influence on the dependent variable *Y*. The null hypothesis H_0 is that there is no relation at all between *X* and *Y*.

research hypotheses and a sampling framework that will be able to generate statistically meaningful results – whether that proves or disproves the null hypothesis.

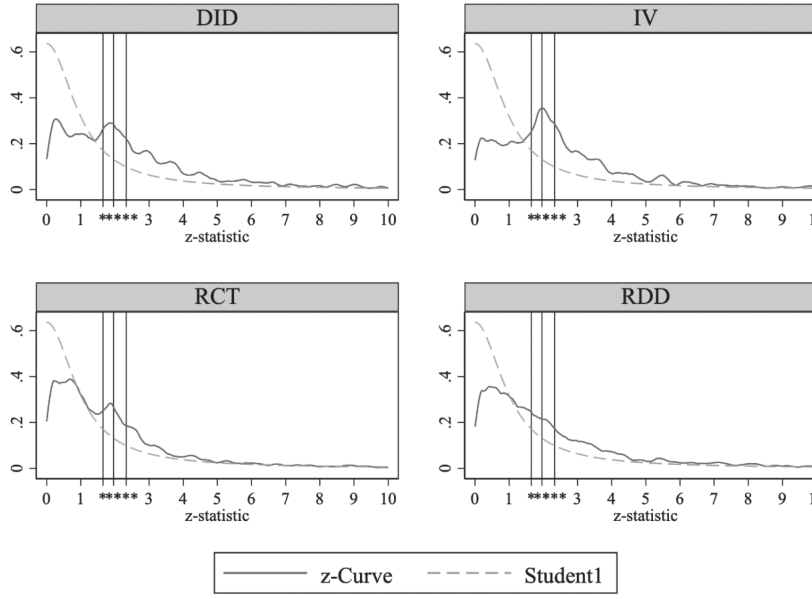
To assess the statistical significance of hypotheses when using NHST, researchers must calculate the -value. A significant result is conventionally held to be a *p*-value below ≤0.05 (or <0.01 if a higher bar is set), which demonstrates that any relationship between the variables of interest is not a result of random variability or static in the data since in probability terms it would arise only 5% or 1% of the time. A *p*-value below ≤0.05 is generally seen as allowing the researcher to (tentatively) dismiss the null hypothesis. Rejecting the null hypothesis does *not* prove that the alternative hypothesis is correct, but it *does* demonstrate that there is a statistically significant relationship between the variables being examined.

The use of significance levels in quantitative data testing approaches may add uncertainty here, since many test propositions different from $X \rightarrow Y$ (such as alternative links from $Z \rightarrow Y$ or $A \rightarrow Y$) may meet the same standard of significance. Equally, putting a lot of weight on *p*-values may alter researchers' behaviours (whether unself-consciously or deliberately) since their incentive is to find H_1 propositions where the null hypothesis can be rejected, creating a 'positive result' more likely to be published in a journal than a 'negative result' where the null hypothesis cannot be rejected. Researchers may be encouraged to make changes in what data is included in their work, or what methods are used to analyse it, in a *p*-hacking manner (see Chapter 3) – that is, making changes solely to try to bring a relevant H_0 to a point where it can be rejected and H_1 retained.

A key root of such deviations from sound methods in quantitative work lies with the 'publication bias' whereby journals are more likely to see as 'original' and publish papers with 'positive' results than those with only 'negative' findings. For top economics journals Brodeur, Cook and Heyes (2018) compared the reported *p*-values between papers which employed RCTs and those using laboratory experiment (Figure 4.3). For laboratory experiment data, they found a clear sub-peak in reported values just after the *p*-values cutoff of ≤0.05, suggesting that statistical results had been 'massaged' to fit just under this threshold. While instances of *p*-hacking cannot be identified in individual studies, their distribution in aggregate is telling.

Similarly, Open Science Collaboration (2015) sought to investigate the replicability of established research findings in psychology. With the collaboration of 270 contributing authors, their study conducted quasi-random, 'high-quality' replications of findings published in three leading psychology journals. Of these reproducibility efforts, only 36% reported findings consistent with the original research. Even more concerningly, many of the look-again studies were conducted with larger sample sizes. Overall, the findings also found no statistical relationship between the *p*-values of the original and replication studies.

Figure 4.3: The distribution of z-statistics for different sub-samples of economics papers, by type of data



Note: Four types of studies are covered, using DID (difference in difference) approach; IVs (instrumental variables); RCTs (randomized control trials); RDD (regression discontinuity design). The vertical lines show the conventional significance levels.
Source: Reproduced with permission from Brodeur, Cook and Heyes (2018).

In addition, of course, wider problems may arise around the limited or zero reproducibility of test outcomes, the skewing of empirical testing towards clearly proved or disproved propositions, and the temptations for researchers to engage in the various post hoc analytic manoeuvres to make their findings look more 'interesting' or convincing, such as *p*-hacking, but also 'data-dredging' and 'harking' to find hypotheses that 'work' (see Chapter 3). Kerr (1998) argues that harking post hoc damages research because it introduces type I errors into theory that are later hard to eradicate; it takes unjustified statistical licence; it presents inaccurate models of the scientific method to students; and it implicitly violates basic ethical principles. (He also lists four other harms introduced by the practice.) To revisit and modify the hypothesis is therefore seen as creating an explanation that fits the data rather than one that seeks to build or advance new scientific theories. Thus, in quantitative research settings, research questions and design should normally now be set in advance through the use of pre-analysis plans. Pre-registration brings transparency and openness to the first stage of the research lifecycle, in the process aiding the reproducibility of research.

Pre-registration to strengthen CDA research design

In pre-registering a PAP a key consideration is the sample size to be used, which determines the statistical 'power' and validity of the analysis conducted. As with the hypothesis wording, the digital time stamp created by the pre-registration process is key because it enables researchers to show that as the research evolved any steps needed to make legitimate alterations in the sample's size were specified prior to the research beginning. The form of the hypothesis used in a NHST requires an understanding of the sample being taken for the research. To statistically test the null hypothesis the amount of 'static' or random variation in the data must be computed as a specific value, which can then be evaluated. Known as the *t*-value or *t*-statistic, this number is calculated from the sample mean, the population mean, the standard distribution and the sample size. In specifying the parameters of the hypotheses we may know the true size of the population and the distribution of values within it – for example, if working with a form of comprehensive data such as a census. Or these values may be estimated from the data collected. Where the standard distribution and population mean are being estimated, the size of the sample collected can have a significant impact on the validity of the findings. Thus, the specification of the sampling framework that is developed as part of the hypotheses being tested influences the calculation of whether the findings are statistically significant.

However, specifying a sample size is not the same as collecting the data. Indeed, in social science research, it is common for the particulars of the sample actually gathered to vary from what was intended to be collected. A PAP can help demystify this variation, explaining how the research may adapt to foreseeable issues, and allow contingencies or limitations to be addressed transparently. For example, what happens if there are not enough participants in the initial sample, or too many? Should a second round of data collection be conducted to ensure that the research has the appropriate statistical power, and if so how and when should it be enacted?

These questions are not merely theoretical, but relate to 'optional stopping' in data collection which can severely undermine the validity of NSHT research. 'Optional stopping refers to the practice of peeking at data and then, based on the results, deciding whether or not to continue an experiment' (Rouder 2014). The issue with optional stopping is that it increases type I error rates, where the null hypothesis is incorrectly rejected. For an unscrupulous researcher, the benefits are clear. They can keep expanding the sample $n+1$ until the desired, positive result is reported. Without a pre-specified sample size, it is impossible to compute the type I error rate – another reason the intended size of the sample to be collected is set down in a PAP. Deviating from a pre-set sample size (regardless of its intention) can affect the calculation of type I errors. Consequently, a well-intentioned researcher can inadvertently commit optional stopping errors in their research if they expand or reduce their sample size for practical reasons during data collection and

this change is not factored into the subsequent statistical analysis. Without a PAP the introduction of this error may not be made clear to other researchers reading the results of the research. Nor may there have been a specified procedure for the researcher to follow should the actualized sample size change and the research plan and analysis need to be adapted to reflect that.

Furthermore, in registering a PAP it is also crucial to limit the statistical methods through which the data is to be analysed. Modern statistical software makes running multiple statistical tests extremely easy. So once data has been collected there may be a temptation to run it through a wider variety of statistical tests with different options to see where and if significant results are returned. However, searching for significant results in this way increases type I error rates and can undermine the validity of the research. Rather, prespecifying the analysis that will be undertaken, along with a justification for why the exact methods and parameters have been chosen, pre-determines the error rate and encourages researchers to have a good understanding of how the statistical methods being used work. For example, running a standard 2×2 analysis of variance (ANOVA) on a truly null dataset has a 14% chance of producing at least one statistically significant result (Herzog et al. 2019). In such an example, a researcher searching for significance without a predefined analysis plan, and a poor understanding of the mathematics which underly the statistical measures being used, could easily and incorrectly accept these results, leading to a type I error.

Closely linked to running many tests until a model produces a significant result is tweaking the parameters of the statistical analysis so as to return a *p*-value of 0.05 or *p*-hacking. Here researchers adjust how they analyse their data to ensure that their research produces significant results. When taken on their own, many of the tweaks made in the course of *p*-hacking are innocuous and they may even be well justified. For example, if response rates are lower than anticipated, extending the time or 'reach' being attempted to find respondents may seem sensible. Or if your data is not showing significant results, perhaps adjusting the threshold for outliers from three to two standard deviations might yield a different outcome. This practice, however, undermines the purpose of using the *p*-value as a measure of significance because researchers have purposefully adapted their data or analysis to ensure that it meets the established threshold.

Other safeguards exist. For example, the test for excessive success (Ioannidis and Trikalinos 2007) seeks to understand whether the reported success rate for an experiment fits within the parameters that would be expected for the sample size reported. In a social science setting, where researchers cannot recreate the exact conditions of an experiment and are often reliant on samples taken from a much larger population, there is an inherent error rate for the research being conducted. Research that reports excessively successful findings therefore diverges from what is known about the population, raising questions about the overall validity of the findings. For example, research by Francis et al. (2014) found that in a set of 18 psychology articles published in *Science* over four fifths reported statistically excessively successful results.

The limits of pre-registration

Pre-registration and filing pre-analysis plans are not a 'silver bullet' to fixing all the issues of around a replication crisis. That a study is pre-registered does not mean that it is free of biases or that it has been designed in a thorough and comprehensive manner. Research designs with important theoretical or quantitative issues can still be registered. Moreover, while pre-registration aims to ensure that other researchers can follow through and replicate all of the steps of the research process, much of the tacit knowledge that goes into actually conducting the research may still not be recorded. While pre-registration may assist researchers in developing their studies and improving the overall quality of research designs, there is nothing intrinsic about PAPs that guarantees this.

However, the pre-approval and peer review of PAPs by some journals can make a difference. In effect, it provides a significant incentive to researchers to both develop a PAP ahead of conducting the research and on completion to publish whatever findings they have – whether they prove an effect or support the null hypothesis. Allowing researchers to receive authoritative external feedback on their proposed research plan before it is conducted and alter their study design accordingly *before* the research is done may prevent research time being wasted. It also forestalls researchers being asked to conduct additional research to meet reviewers' requirements for their article later on. The most developed approach here is 'registered reports' where researchers in stage 1 submit a detailed and comprehensive PAP or protocol for their planned study, which is fully peer reviewed before it is implemented. If that is accepted, in stage 2 the team collect and analyse the data as promised and submit a final paper manuscript with the results and discussion, which is automatically published, whatever the findings show. Accepting studies with sound methodologies and PAPs means that a journal with registered reports moves away from just publishing studies with positive results.

A common criticism of pre-registration, PAPs and especially registered reports has been that they do not allow for flexibility in the research process. This is true, in so far as their intended purpose is to be able to hold up a record of the research plan against the research that was conducted as a means of ensuring validity and transparency in the research process. However, just because a research plan has been pre-registered this does not mean that it has to be strictly followed. Researchers can still deviate from a pre-registered plan for legitimate reasons, but what is important is that any deviation is explained and accounted for in the published results. The effect that changes might have had on the validity of the research then has to be acknowledged and discussed. Variation in how the research is conducted can also be accounted for in a PAP with the use of the correct methodological tools. For example, a preset sample size is important for frequentist statistical methods. For Bayesian statistics, however, optional stopping does not present the same problems

(Rouder 2014). So here not planning on a set sample size, or deciding to just work with as many participants as show up on the day, would not necessarily undermine the statistical validity of research.

4.3 Exploratory data analysis (EDA) approaches

Exploratory imperatives are also equally strong in many (perhaps most) areas of quantitative social science, reflecting the impact of the wave of critique and statistical innovation triggered by John Tukey's 'exploratory data analysis' (Tukey 1977; Mosteller and Tukey 1977; Tukey 1993). Long before the PC revolution in the late 1980s and the internet wave in the mid 1990s that gave researchers easy access to data inspection and summarizing tools, plus the visual representations of data on which Tukey laid huge stress (Blaha 2019), his EDA approaches stressed the need to disentangle compound effects often present in data and to look hard at apparent findings and patterns (Jebb et al. 2017). Subsequently EDA philosophies and tools were swiftly incorporated into standard software packages and analysis courses changed to reflect the lessons of EDA approaches, such as always plotting the data, thinking of modelling as fit plus residual and then replotting residuals, and so on.

In modern social science the relationship between CDA and EDA is sometimes described as 'a continuum', as if it varied along only a single dimension. In fact the differences between the two are far more complex than that, as Tukey (1993) made clear, varying along multiple different dimensions that cannot easily be summarized. However, Figure 4.4 shows a rough stab at mapping the differences between them onto a two-dimensional canvas that has useful insights. The vertical axis here focuses on the role that theory plays in generating precise hypotheses for testing, versus situations where hypotheses are culled from or supplement an established empirical literature, and versus situations where there is no 'consensus' theory, or where the viewpoints that do exist do not generate precise or testable propositions that can be addressed via data.

The horizontal axis in Figure 4.4 focuses on the status of the data that is available for analysis, with purpose-designed and original data as the highest value information in the rightmost column, contrasted with the central column categories of data generated atheoretically (e.g. for government or corporate purposes) or with data already available for secondary analysis. The rationale here is this:

> The scope for biases to distort research outputs from secondary data analysis is perhaps particularly acute, [*because*] researchers now have increasing access to high-dimensional datasets that offer a multitude of ways to analyse the same data [*and*] analytic flexibility can lead to different conclusions depending on the analytical

> choices made. [*In addition*] current incentive structures in science reward researchers for publishing statistically significant, novel, and/or surprising findings. This combination of opportunity and incentive may lead researchers – consciously or unconsciously – to run multiple analyses and only report the most 'publishable' findings (Baldwin et al. 2022). [*Our inserts*.]

The leftmost column on the horizontal axis concerns situations where only qualitative forms of information are available that do not lend themselves to numerical representation or quantitative analysis.

Confirmatory causal analysis occupies the top-right content cell in Figure 4.4 – here (i) precise hypotheses directly derived from a reasonably consensus theory can be tested on (ii) specifically produced new data, for example information collected in a specific lab experiment or via an RCT. Some forms of computational data analysis may also meet both the theory and the data criteria here.

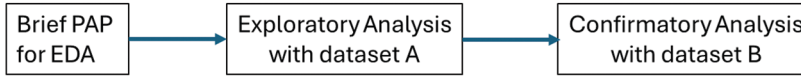
By contrast, exploratory data analysis occupies the middle column and row in the centre of the figure. Here the hypotheses to be tested are not directly theory-specified, but may derive largely from previous work in the field. So a new analyst necessarily comes to their task with a freight of expectations and prior knowledge that are likely to influence their choice of analysis approach and how the research exercise is framed. Similarly, the data used in exploratory data analysis work is not newly minted but instead involves secondary analysis of already produced quantitative data – either generated for official statistics reasons (including many recurring series), or from digital behaviour monitoring, or available from previous or larger data-generating research projects (Mertens and Krypotos 2019). Some exploratory data might be heterogenous in relation to likely significant control variables, and the data patterns uncovered may reflect multiple processes at work simultaneously, even after some major explanatory variables have been controlled for (Myatt and Johnson 2014).

The central position of EDA here is no exception, for there are good arguments to believe that it is likely to be and to remain the most widely practised form of contemporary social science research. We might realistically expect that even most causal analysis is likely to not fully meet 'hard' CDA requirements but to consist of some mix of EDA and CDA work. For instance, testing hypotheses closely derived from strong consensus theory but on previously generated or imperfectly controlled data (in the top contents row in Figure 4.4) is common. In a differently mixed way, the last column covers situations where hypotheses that are only rather weakly derived from a consensus theory (and hence where there are multiple rival interpretations) or essentially derived from previous data in the field rather than from strong theory may be tested. Even if such studies use strong quantitative methods (like experiments or RCTs) they must still only qualify as mixed CDA/EDA work because analysts' strong prior knowledge of what might qualify as statistically significant may shape the analysis.

Figure 4.4: Comparing confirmatory research, exploratory data analysis and solely qualitative research

Type of data / Theory status	Qualitative data or information	Data from previous studies, or recurring survey/behaviour series, or non-theory-derived statistics	Newly obtained quantitative data e.g. via lab experiments or RCTs
Hypotheses under test stem directly from strong, consensus theory	*Hypothetico-deductive approaches not used*	*mixed EDA/CDA*	*Confirmatory data analysis (CDA)*
Hypotheses tested chiefly derive from past empirical work	*generating hypotheses, sense-making*	*Quantitative exploratory data analysis (EDA)*	*mixed EDA/CDA* *e.g. two-stage EDA then CDA work; or training/holdout sub-sets (with large N data only)*
No consensus theory, or competing theoretical interpretations, or meanings need to be constructed	*Inherently exploratory methods of analysis*	*Inductive quantitative analysis, data-lead*	

A two-step research process has become an increasingly standard solution offered by researchers either operating with previous data, or with high-quality new data but in social science fields where there is no consensus theory, or where the social meanings of even quantitative data indices are complex and need deconstruction because of multiple causal routes to the same scores. In any of these three situations the recommended OSS sequence of work is:



Before undertaking even an EDA project, some authors suggest that researchers should file a mini PAP setting out the rationale, the preliminary data to be used (such as a subset of the full dataset) supplemented perhaps by qualitative methods, and the potential effects being looked for. After the EDA preliminary work is done, these elements are then updated in the PAP for the CDA phase, where a full analysis plan with data handling and any data exit or exclusion details are set out. The CDA phase is then implemented according to the commitments made and results (positive or negative) are published, even if no advance in causal thinking proved feasible (see Figure 4.5).

Figure 4.5: The relevance of pre-analysis plans and the role of pre-registration in CDA and EDA research

Type of data / Theory status	Qualitative data or information	Data from previous studies, or recurring survey/behaviour series, or non-therory-derived statistics	Newly obtained quantitative data e.g. via lab experiments or RCTs
Hypotheses under test stem directly from strong, consensus theory	*Hypothetico-deductive approaches not used*		*CDA* heartland – comprehensive pre-analysis plan (PAP) needed
Hypotheses tested chiefly derive from past empirical work		*EDA* – a PAP sets out the intended analysis, and any data edits/limits/upgrades, plus the rationable for new controls	*e.g. in 2-stage approach, separate PAPs for EDA and CDA stages*
No consensus theory, or competing theoretical interpretations, or meanings need to be constructed	*Inherently exploratory methods of analysis* – a PAP can record expectations to log reappraisal changes against	PAP assures against data-dredging or p-hacking	

A variation on running two separate studies (one EDA and the second CDA) is often feasible for large *N* datasets. Here a part of the available data is separated out in a randomized way to act as the 'training' dataset and used for exploratory analysis to surface patterns of relations and interactions. From this exercise a specific PAP is formulated and pre-registered. Meanwhile the bulk of the dataset is kept as 'holdout' and used in the second confirmatory analysis stage when the pre-registered plan is implemented. For survey datasets, which are generally smaller, the suggestion has been that perhaps a third of the total cases should form the training subset, and two thirds the holdout. However, for really 'big data' such as that generated from looking at social media messages, or from information on government transactions with citizens or firms, or businesses transactions with customers, the training data can be a much smaller subset of the whole.

Where really large datasets are involved a new problem arises with relying on conventional tests – namely that at large *N* statistically 'significant' results may be generated for almost any association between variables. Here different approaches to analysis are needed, such as running a tournament of alternative models in different subsets of the complete dataset, or paying attention to effect sizes not significance scores (Varian 2014). With internet data systems interacting with millions of users, implementing a form of online randomized

control trials may be feasible to assess how well different treatment effects deliver desired results.

Some critics have argued that even the two-step versions, seeing EDA as the first step towards pre-registered CDA causal models, underestimate the extent to which most data analysis is likely to remain exploratory on one or both of the dimensions in Figure 4.5. For example, Beecham (2023, pp. 1–2) calls for social scientists to go 'beyond the EDA/CDA dichotomy':

> The EDA → preregistration → CDA workflow is best suited to mature research settings where the theories and methods are well-developed and where prior studies and data exist that can be used to judge and evaluate observed effect sizes (McIntosh 2017). In modern applied data analysis settings, where administrative and passively-collected data are variously combined and repurposed (Singleton and Arribas-Bel 2021), these certainties are difficult to achieve: the potential analysis space is broad and there is no stand-out approach to formulating research questions, selecting datasets and statistical techniques. Additionally, too strict an adherence to separating EDA (informal pattern-finding) from CDA (statistically-grounded enquiry) may not always lead to strong theory and knowledge development. Attempts to impose pre-specification designs onto research settings that rely on secondary or 'found data' may result in knowledge statements that are overly specific or that do not express sufficient detail or richness in outcomes (Szollosi and Donkin 2021). [*References here are those in Beecham's quoted text.*]

In EDA-only contexts, where confirmatory methods are not attempted, a less demanding form of PAP can still be useful, with researchers committing on a limited number of precise propositions to be evaluated and on criteria and methods for evaluating them. Such a PAP can successfully fend off suspicions of the researchers making ad hoc variable choices, or 'method shopping' so as to maximize anticipated or desired apparent effects.

Inductive approaches to data analysis are especially open to criticisms that the analysts have engaged in data-dredging, *p*-hacking or the suppression of supported null hypothesis tests. Especially in contexts where the appropriate theoretical lens is controversial and still debated, there is more potential for confirmation biases (researchers wanting an expected effect to be present), and for critics to see data inclusion or shaping decisions as linked to it. Here a PAP that is time-stamped before the analysis starts and to which researchers stick in their implementation can provide an important role in keeping the analysis honest and underpinning the sound provenance of results. A Delphi study of researchers using qualitative methods concluded that 13 items are appropriately covered by pre-registration, beginning with the research aims, research questions and study design (Haven et al. 2020). Equally important were how sampling and case selection was done, the data sources and types,

the methods used for collecting data, collection tools or instruments (e.g. survey questionnaires or interview schedules) and stopping criteria. Finally, pre-registration should also cover the approach used for data analysis, and the actual process of analysis and credibility checks undertaken.

4.4 Exploratory research generating evidence but not data

There are many forms of exploratory research in the social sciences which rely solely on qualitative data (shown in the yellow cell of Figure 4.4 above). They include, for example:

- research that sets out only to describe the social world in descriptive/tracking or initial 'sense-making' ways, which may be inevitable in areas of social life where qualitatively new social phenomena have arisen, such as demand for goods that did not hitherto exist;
- research into rapidly changing and continuing social behaviours, for example uses of new technologies and their impacts on social norms or contexts;
- projects investigating phenomena or questions in areas where established or agreed theory is vague, indeterminate or undeveloped;
- research into aspects of society where there is no 'consensus' theory, or no theory yet exists, or where existing theory approaches seem inadequate or are vague, not suggesting precisely testable propositions;
- pilot research projects where the central aim is to generate plausible hypotheses that can then be later analysed in confirmatory research;
- research rooted in non-falsifiable epistemologies – such as psychoanalysis.

In all such areas social scientists typically use qualitative research to investigate what is going on, creating valuable (but not numerical) evidence, and relating it to other social phenomena using theoretically grounded methods. Traditions of analysis have often been strongly developed over decades and proven consistently valuable in generating reliable understandings and insights.

For this kind of essentially or inherently exploratory research (the bottom left cell in Figure 4.4) the pre-registration of research questions may seem obviously inappropriate at first sight. This position has been most associated with authors using methods that require them to be 'saturated' in knowledge of the research field and to internalize the complex meanings of events, perceptions, experiences and relationships to different kinds of research subjects, as with grounded theory, for instance (Rachel 2018). Many analysts here seek to understand in detail or to deconstruct the complex social meanings of actions, opinions or other social phenomena, arguing that the messiness of everyday life situations can only be understood through 'grounded' analysis and cannot

be reduced to causal flows between 'variables' (Clarke, Friese and Washburn 2018). A smaller literature seeks to apply 'interpretivist' approaches to quantitative analysis itself by contextualizing the ways that data are constructed, understood and used in societal organizations – tending to dismiss as naïve claims to have achieved a single 'objective' analysis. However, other qualitative social scientists are still interested in unpicking causation, but stress that there are dozens (perhaps hundreds) of causal factors in play in most social situations. As a result, there are often multiple different pathways of causation that lead to the same outcomes (Ragin 2000), a problem tackled empirically by qualitative comparative analysis (see above).

However, a limited form of pre-registration may still be useful even in the most resolutely qualitative-only work. A mini-PAP here can essentially log analysts' expectations of what they will find before research starts, and clarify the rationale for initial choices made about what to study and how. This look-ahead can then be updated as the analysis unfolds, in particular so as to be able to pinpoint how and why deviations from the expected track occurred. Indeed, some qualitative methodological approaches (such as situational analysis) already deploy similar iterative approaches to mapping the research process (Clarke, Friese and Washburn 2018). In particular, researchers' journalling about projects as they go, especially logging how and when changes of perspective were triggered, can help them to reappraise their understandings of the situations and processes being analysed. It may also assist other scholars seeking to reproduce a similar study approach.

Conclusions

Across the social sciences concerns about reproducibility and replication have fuelled some major reappraisals of how well-founded research should be carried out. In quantitative fields researchers are expected to clearly separate out work that uses exploratory methods to surface interesting potential areas of investigation from work that seeks to use confirmatory data analysis. CDA studies increasingly need to be well documented from the outset and conducted according to criteria and using methods that do not change as the study progresses (Safra 2022a, 2022b). Yet the sphere of exploratory data analysis and more inductive research has also expanded and become increasingly critical in the social sciences, and indeed in large areas of STEMM also (e.g. in drug development and rapidly understanding drug side effects). Past sharp divides between exploratory and confirmatory research have blurred, but the challenge on the EDA side and in areas of secondary data analysis has been to provide safeguards against data-dredging, harking, optional stopping, the non-reporting of negative results and other illegitimate post hoc alterations of planned analysis. Even in resolutely qualitative work, pre-registration of core anticipated approaches has a role to play in improving how research is undertaken.

Note

* This chapter draws on research and seminars forming part of the CIVICA Research programme, which was funded in 2022–2024 by the European Union's research and innovation programme Horizon 2020, under the chapter 'Science in Society'.

References

Baldwin, Jessie R; Pingault, Jean-Baptiste; Schoeler, Tabea; Sallis, Hannah; and Munafò, Marcus R. (2022) 'Protecting against researcher bias in secondary data analysis: challenges and potential solutions', *European Journal of Epidemiology*, vol. 37, pp. 1–10. https://doi.org/10.1007/s10654-021-00839-0

Beecham, Roger (2023) ' Scientific reform and visual data science: retiring the EDA/CDA dichotomy', Open Science Foundation preprint, 27 March. https://perma.cc/8YPG-M5Y2

Blaha, Leslie M. (2019) 'We have not looked at our results until we have displayed them effectively: a comment on robust modeling in cognitive science', *Computational Brain & Behavior*, vol. 2, pp. 247–250. https://doi.org/10.1007/s42113-019-00059-6

Brodeur, Abel; Cook, Nikolai; and Heyes, Anthony (2018) *Methods Matter: PHacking and Causal Inference in Economics*, Bonn: Institute of Labor Economics (IZA). IZA Discussion Papers 11796. http://hdl.handle.net/10419/185256

Clarke, Adele E.; Friese, Carrie; and Washburn, Rachel A. (2018). *Situational Analysis: Grounded Theory After the Interpretive Turn* (2nd edn). Thousand Oaks, CA: Sage. **$** https://uk.sagepub.com/en-gb/eur/situational-analysis/book238990

Clark, William Roberts and Golder, Matt (2006). 'Rehabilitating Duverger's theory: testing the mechanical and strategic modifying effects of electoral laws', *Comparative Political Studies*, vol. 39, no. 6, pp. 679–708. **$** https://doi.org/10.1177/0010414005278420 OA at: https://doi.org/10.1177/0010414005278420

Cox, Gary W. (1997). *Making Votes Count: Strategic Coordination in the World's Electoral Systems*. Cambridge: Cambridge University Press. **$** https://doi.org/10.1017/CBO9781139174954

Crouch, Daniel JM and Bodmer, Walter F. (2024) 'Evolution by natural selection is a scientific law and not just a theory', *Academia Biology*, vol. 2, no. 1. https://doi.org/10.20935/AcadBiol6158

Dickson, Eric S. and Scheve, Kenneth (2007) 'Social identity, electoral institutions and the number of candidates', *British Journal of Political Science*, vol. 40 no. 2, pp. 349–375. $ https://doi.org/10.1017/s0007123409990354

Dunleavy, Patrick and Diwakar, Rekha (2013) 'Analysing multiparty competition in plurality rule elections', *Party Politics*, vol. 19, no. 6, pp. 855–886. $ https://doi.org/10.1177/1354068811411026 OA at https://perma.cc/3RSK-Z2BP

Duverger, Maurice (1964) *Political Parties: Their Organization and Activity in the Modern State. Internet Archive*, London: Methuen.

Francis, Gregory; Tanzman, Jay; and Matthews William J. (2014) 'Excess success for psychology articles in the journal *Science*', *PLOS ONE*, vol. 9, no. 12, article me114255. https://doi.org/10.1371/journal.pone.0114255

Haven, Tamarinde L. et al. (2020) 'Preregistering qualitative research: a Delphi study', *International Journal of Qualitative Methods*, vol. 19, pp. 1–14. https://doi.org/10.1177/1609406920976417

Herzog, Michael H; Francis, Gregory; and Clarke, Aaron (2019) *Understanding Statistics and Experimental Design: How to Not Lie with Statistics*, Berlin: Springer. https://perma.cc/5HWH-TZA8

Hunt, Tam (2014) 'Reconsidering the logical structure of the theory of natural selection', *Communicative and Integrative Biology*, vol. 7, no. 6, article e972848. https://doi.org/10.4161/19420889.2014.972848

Ioannidis John P.A. and Trikalinos Thomas A. (2007) 'An exploratory test for an excess of significant findings', *Clinical Trials*, vol. 4, no. 3, pp. 245–54. $ https://doi.org/10.1177/1740774507079441

Jebb, Andrew T; Parrigon, Scott; and Woo, Sang Eun (2017) 'Exploratory data analysis as a foundation of inductive research', *Human Resource Management Review*, vol. 27, pp. 265–276. https://doi.org/10.1016/j.hrmr.2016.08.003

Kerr, Norbert L. (1998) 'HARKing: hypothesizing after the results are known', *Personality and Social Psychology Review*, vol. 2, pp. 196–217. $ https://doi.org/10.1207/s15327957pspr0203_4

Lakatos, Imre (1970) 'Falsification and the methodology of scientific research programmes', in Lakatos, Imre and Musgrave, Alan (eds) *Criticism and the Growth of Knowledge*, Cambridge: Cambridge University Press, pp. 91–195. $ https://doi.org/10.1017/CBO9781139171434.009

McIntosh, Robert D. (2017) 'Exploratory reports: a new article type for Cortex', *Cortex*, vol. 96, A1–A4. https://doi.org/10.1016/j.cortex.2017.07.014

Mertens, Gaëtan and Krypotos, Angelos-Miltiadis (2019) 'Preregistration of Analyses of Preexisting Data', *Psychologica Belgica*, vol. 59, no. 1, pp. 338–352. https://doi.org/10.5334/pb.493

Mosteller, Frederick and Tukey, John W. (1977) *Data Analysis and Regression. A Second Course in Statistics*. Reading, MA: Addison-Wesley.

Myatt, Glenn J and Johnson, Wayne P. (2014). *Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining* (2nd edn), Hoboken, NJ: Wiley

O'Boyle, Ernest H.; Banks, George C.; and Gonzalez-Mulé, Erik (2016). 'The chrysalis effect: how ugly initial results metamorphosize into beautiful articles'. *Journal of Management*, vol. 43, no. 2, pp. 376–399. https://doi.org/10.1177/0149206314527133

Open Science Collaboration (2015) Estimating the reproducibility of psychological science, *Science*, vol. 349, article 6251. **$** http://dx.doi.org/10.1126/science.aac4716

Ragin, Charles C. (2000) *Fuzzy Set Social Science*. Chicago, IL: University of Chicago Press. **$** https://press.uchicago.edu/ucp/books/book/chicago/F/bo3635786.html

Rouder, Jeffrey N. (2014). 'Optional stopping: no problem for Bayesians'. *Psychonomic Bulletin & Review*, vol. 21, no. 2, pp. 301–308. https://doi.org/10.3758/s13423-014-0595-4

Safra, Lou. (2022a) 'Pre-registration in the social sciences', PowerPoint presentation, Sciences Po, Paris. https://zenodo.org/records/7501036/files/ Pre-registration%20in%20the%20Social%20Sciences%20(Lou%20Safra).pptx?download=1

Safra, Lou. (2022b) 'Pre-registration in the social sciences', Sciences Po Seminar presentation, Civica YouTube page. https://www.youtube.com/watch?v=JaXHs5oirnc&list=PLTtFRsfDQ8MkUG8VRKUrp7C5O3z1v11YX&index=2&t=5s

Singer, Matthew M. (2013) 'Was Duverger correct? Single member district election outcomes in fifty-three countries', *British Journal of Political Science*, vol. 43, no.1, pp. 201–220. https://doi.org/10.1017/S0007123412000233

Singleton, Alex and Arribas-Bel, Daniel (2021) 'Geographic data science', *Geographical Analysis*, vol. 53, pp. 61–75. https://doi.org/10.1111/gean.12194

Szollosi, Aba and Donkin, Chris. (2021) 'Arrested theory development: the misguided distinction between exploratory and confirmatory research',

Perspectives on Psychological Science, vol. 16, no. 4, pp. 717–724. https://doi.org/10.1177/1745691620966796

Tukey, John W. (1977) *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Tukey, John W. (1993) 'Exploratory data analysis: past, present and future', Technical Report no. 302, Princeton, NJ: Princeton University. https://perma.cc/GFG3-X9JT

Varian, Hal R. (2014) 'Big data: new tricks for econometrics', *Journal of Economic Perspectives*, vol. 28, no. 2, pp. 3–28. https://perma.cc/FB3Z-H86B

PART II

Practical approaches to more open research

Conducting research is a complex task where priorities often evolve rapidly from the bidding stage, through starting projects off, then conducting the main body of research, and lastly finishing up. Projects can take years to develop and even longer to come to fruition. Nor do they often conclude neatly and on time, with personnel often moving on at different stages to other work and other topic priorities. Many research projects also continue to require further attention and work long after the bulk is complete (and even after they have 'officially' finished), because of publication demands, reports to funders and follow-on exploitation of data or other resources. In this part we examine four different practical ways of planning in advance and setting up research processes so as to minimize the whole-life workloads involved in this complex process. The shift towards open social science (OSS) strengthens this impulse by reinforcing researchers' incentives to design for these imperative from the outset.

Perhaps the most foundational issue for good research is how well and promptly it is documented. Nothing is more destructive of the extent and quality of recording information than allowing time to elapse between when events were documented, or changes made to data gathering and analysis, and when they are written up. Chapter 5 sketches some open science (OS) principles and tools for documenting as you go that are applicable to original research across the social sciences discipline group, whatever types of method are being employed. However, not all social science data is constructed afresh from the ground up, and a great deal of research involves reusing other people's data (examined in Chapter 6). Cleaning up, clarifying and checking

previously produced data, and then perhaps combining or 'mashing' it with other existing or newly gathered data, raises some additional issues.

If data or analyses are to be maximally useful to other researchers (and often if projects are to be successful in being funded in the first place) constructing an overall data management plan (DMP) is an essential step (covered in Chapter 7). Like the other steps above, writing a DMP may seem to entail an unwelcome 'frontloading' of effort at the busy beginning of projects. In fact, a DMP is a less onerous undertaking than it may sound. And it should also pay off in terms of anticipating and mitigating potential future problems before they accumulate, and helping deliver a smoother publication of outputs and depositing of full datasets in the open and reproducible ways that grant funders increasingly mandate.

Finally, in every research output and publication (whether a journal article, research book or PhD thesis) the analysis and every exhibit included should be easy for readers to understand and transparently assess. The information needed should come out along with the output itself, and in a form that is fitted to every exhibit in the authors' sequence (in formats that we cover in Chapter 8). For each journal article this requirement is met via a linked 'replication archive', either one housed on the journal's own site, or in a high-quality and permanent data repository. For theses and research books it is best to move away from past practices of trying to cram all the necessary detailed information into a single omnibus: 'Chapter 2: Methodology'. A section or chapter on methodology included in the main text should cover major methods design issues that are theoretically relevant, but not go into minute details. Instead, a research methods annex (RMA) located at the end of the book or thesis should house all the really detailed information needed by critical readers and valuable for any follow-on researchers reusing the same data, or inspired to undertake research that is similar in some other way.

5. Documenting research as you go – principles and tools

Documenting as you go through the many stages of the research process means keeping good records of what you have done, why you have done it and what you found out. These tasks are integral to progressing through every stage of the busy research life cycle, but in a manner that is both as comfortable and responsible as possible (given other task demands):*

> The single hardest part of making data useful involves disciplining yourself to make your data 'future-proof'. By planning for the long-term at the outset, you also achieve the happy accident of making your data amenable to re-use and distributed collaboration. (Shawn et al. 2020, Part 2).

However, the prospect of developing a system to properly document your research can feel like a daunting task. Even for experienced researchers there can be an intimidating gap between how they actually track and organize their research and how they think they really should be documenting it. In this chapter we cannot provide instructions on achieving an ideal system of documentation, for no such system exists that works across many diverse projects. Instead our more modest aim is to introduce a set of high-level principles and tools that can be used to structure your ongoing approach to documenting your research quickly and effectively, hopefully minimizing the slog involved (Shawn et al. 2020, Part 2).

Especially at the outset of research and in high-intensity periods later on there is always a temptation to move quickly and just rely on one's memory. It is easy to believe that things we have 'mastered at the time' we will retain and be able to recall accurately in the future, but this is usually a fallacy. At the time when we make a micro-move or an in-process decision things may seem 'obvious' or 'intuitively right', but in fact memory about micro-decisions is often highly fallible. Shallowly recognized decisions (e.g. the rationales for

How to cite this book chapter:

Dunleavy, Patrick and Monteath, Timothy (2026) 'Documenting research as you go – principles and tools', in Dunleavy, Patrick and Monteath, Timothy (2026) *Doing Open Social Science: A Guide for Researchers.* London: LSE Press, pp. 107–128. https://doi.org/10.31389/lsepress.dos.e

fixing a recode or leaving out a case from analysis) are especially likely to drop out of sight. Relying on memory for the integrity of your research will always lead to problems in the future.

Yet it is also vital to minimize the transaction costs of your chosen system for documenting research as you go, to keep down the regular time or effort costs to what is strictly needed – because that is the best guarantee that you will stick with the system. The higher the routine costs of your approach, the more the documentation process may lead to a friction with other priorities – when corners can be cut, steps be skipped, gaps appear and ultimately the system of records collected breaks down.

Early decisions and technical changes to your workflow made at the start of new research can have long lasting benefits. For example, picture a messy workshop with multiple ongoing pieces of work. Around the workshop are strewn various tools, nuts and bolts, instructions manuals in an apparently chaotic fashion. For a skilled artisan in their own workshop everything may nonetheless have a purpose and a place. Yet others cannot understand the organization as it exists in the artisan's mind, making the workshop difficult for even skilled people to use and an inscrutable environment unsuitable for teaching new skills. Now imagine a workshop that has a clear system of organization, where similar tools and parts are sorted into containers organized into logical groupings in transparent containers – so that anyone new to the workshop can see what is inside each box and how many parts are available. The workshop still reflects the idiosyncrasies and priorities of the skilled artisan in control. For instance, regularly used tools are kept close at hand, several different projects are ongoing on different workbenches, and how they wish to organize the space may still be unique. A good research workflow should mimic this image, reflecting the needs and sensibilities of the researcher(s) and their work, while at the same time being legible and transparent in ways that help explain what was done, fully establish the provenance of evidence, and share research methods and findings with others.

We begin by considering some of the reasons that documentation can go awry, looking at examples of easy mistakes to make, at the potential problems and benefits of larger research teams, the importance of naming conventions, and ensuring backup copying of materials (Section 5.1). The second section of the chapter looks at how to develop an appropriate way forward suited to your own research needs, and some generally useful tools to assist in this.

5.1 What can go wrong in documenting

In 2010 two distinguished economists, Carmen Reinhart from Maryland and Kenneth Rogoff from Harvard, published a paper entitled 'Growth in the time of debt' in the *American Economic Review* – one of the world's most prestigious economics journals. They presented evidence that 'when gross external debt reaches 60 percent of GDP, annual growth declines by about two percent;

for levels of external debt in excess of 90 percent of GDP, growth rates are roughly cut in half' (Reinhart and Rogoff 2010, p. 573). Published in the aftermath of the global financial crisis, when government debt levels were rising in many advanced and emerging economies, these apparently clear-cut findings were hugely influential amongst economists. They were also widely cited by politicians and policymakers in national governments and international bodies like the International Monetary Fund (IMF) and World Bank as empirical evidence that austerity policies and cutbacks in public spending were needed to control levels of national debt.

Problems with the paper came to light in a chance way. At the University of New England, as part of class exercise, a graduate MSc student, Thomas Herndon, was asked to pick a well-known economics paper and replicate the results from scratch. Herndon settled on 'Growth in the time of debt' but soon found that he could not replicate the results, even though the paper was based on high-level and easily available national economic indicators. Assuming that he must be getting something wrong he emailed the two authors asking for a copy of their data and analysis so that he could correct any mistakes he was making. After not receiving a response, Herndon went to discuss his replication – or lack thereof – with his professors, Michael Ash and Robert Pollen.

Pressed by these more prestigious professors, Reinhart and Rogoff eventually shared the Microsoft Excel spreadsheet which they had used for their analysis. When Herndon got to inspect it in detail, along with Ash and Pollen, they found multiple problematic Excel coding mistakes, apparently ad hoc exclusions of some countries' data points, or erroneous inclusions of others. They re-ran the same analysis, correcting for the mistakes and data exclusions, and found that 'countries with public debt/GDP ratios above 90% averaged 2.2% real annual GDP growth, not [minus] 0.1% as published' (Herndon, Ash and Pollin 2014, p. 1). They thus concluded that the earlier published results were not secure.

This episode highlights a number of issues that can go wrong when the research process is not documented or changes made during analysis are only poorly recorded. First, the ad hoc exclusions from the dataset made by the authors were not documented or justified either in the data work which was not shared, nor mentioned and explained in the methodology section of the published paper – implemented (like most journal articles) within restrictive word limits. Second, the lack of documentation of decisions was missed by reviewers because of a lack of transparency during the publications process. The original paper was published in a special issue and not peer reviewed, nor did the authors make the analysis dataset available to other scholars for them to inspect and replicate. In combination with the high prestige of both the authors and the journal, a collection of relatively small errors helped create a highly influential paper (unusual for its apparently clear-cut findings), which then came to be used as a faulty basis for austerity policies implemented in many countries, often with damaging results and penalizing people at the bottom of the income scale in particular.

Third, Microsoft Excel and now Google Sheets are often used for data assembly because they are the most widely available and perhaps the easiest to use numerical software packages. And making downloads of final data available in Excel (or in CSV format that is readable in Excel, Sheets or other software) is a good thing because it uses the format most widely available to and understood by potential users. However, using this type of software to undertake analysis can have key disadvantages, particularly for non-expert users. Unlike other academic analysis systems (like SPSS, SAS, Python or R) Excel does not have a mandatory autotracking capability that records how codes are changed, or how the scope of analyses varies (although it does have a version control capability that properly trained users can access). All users can very quickly and conveniently make dozens of such instant changes in Excel. This can be great for producing a business report under time pressure, but it is potentially quite opaque if used for research analysis, unless *all* the changes are separately documented. This aspect, plus relying a lot on Excel's quite complex formulae functions, can make it hard to easily see what calculations are being done within the spreadsheet or how this may have affected other parts of the data. Hence data analysis mistakes are easier for non-experts to make in Excel, and harder to double check and audit.

The best context for using Excel or Google Sheets is within a well-controlled university or workplace with a cloud-based service, such as the Office 365 package with Copilot or a Google Suite server (both of which really need you to be trained to operate well). Using the software with version control will then save every edit in real time (so you can go back and hopefully recall why you made a change), adds a time stamp and IP address, and records which person changed things in team projects. However, if you cannot access such facilities, and more generally for new or non-expert users with small datasets, then re-saving a spreadsheet with a unique name every time it has been used can be good documentation practice. It helps to track file differences in simpler datasets across various sessions, although it may create larger storage problems (unless you can later on winnow or archive earlier versions without risk). But it does not eliminate the problems of potentially numerous instant changes. Alterations made within a given session (and these could run to tens or hundreds of amendments) are still not trackable unless you can use a version control capability.

Recording what was done, and the size of research teams

To avoid these kinds of problems, appropriately documenting your research as you go (in an economical and consistent way) forms an essential part of scientific enquiry. Yet (as with scientific and academic writing itself (Dunleavy 2003; Pinker 2014a, 2014b) there is no single 'scientific' way in which this

process can be systematized. Rather, a feasible aim is to ensure that the knowledge of steps made during the research process is not lost when that is no longer the researcher's immediate, active focus of attention. The research should remain fully comprehensible to, and reproducible and replicable by, the original researchers themselves later on. And future researchers should be able to access much the same knowledge about how the research was conducted, without having to rediscover it or reconstruct the analysis done from scratch. Documentation tries to maximize the explicit knowledge available about the research process followed, reducing the amount of reliance on the memories and implicit or tacit knowledge of the researchers alone – although there will always be some (Collins 2010).

Science, technology, engineering, mathematics and medicine (STEMM) laboratories take very seriously the importance of documenting the research as it is done and training people so that implicit knowledge can be transferred through training within and between labs (Aly 2018). The longevity of many STEMM labs, centres and research programmes has often reflected long-run state subsidies and continuous successes in securing funding. To give one indicator, 35% of STEMM scientists in the UK have research-only jobs, not requiring them to do any teaching (Bastow, Dunleavy and Tinkler 2014, p. 8). Near-permanent research (and research-only career paths in academia and industry) have also allowed the development of comprehensive procedures and lab handbooks to ensure that research within each lab and team of scientists remains consistent and reproducible. Handbooks can cover hundreds of different equipment setups, machine settings, handling processes and interim steps in evidence preparation and analysis, so that they are always made in standardized and recorded ways. Lab handbooks typically go into a lot of detail on what needs to be done – for example, how to prepare samples for use in animal research – and new staff are carefully inducted into procedures. Researchers must keep detailed lab notebooks that other team members can consult, either on paper or increasingly on computer as software for archiving and retaining lab notes in reaccessible ways continues to become more sophisticated.

In the social sciences the situation on recoding as you go has not been directly comparable for multiple reasons. The weaker and more episodic funding of social science research means that it is less common to find large groups of researchers who are consistently employed over many years in the same lab or centre. Only 11% of UK social scientists have research-only contracts, less than a third of STEMM science levels (Bastow, Dunleavy and Tinkler 2014, p. 8). The main trajectory of academic social science research has largely been one of individual researchers (with their own interests and trajectories) coming together in medium-term collaborations of between two and five years with one or two other investigators, creating small teams of other researchers that are project-specific and dissolve after two or three years. Consequently, the development of codified data recording and management approaches has been much slower, more partial (e.g. confined to larger, long-run projects) and generally

less advanced than in the STEMM disciplines. However, 'embedded' social scientists working in industry or government agencies have better-developed 'lab' standards and documentation practices.

Research teams in social sciences academia have grown in size throughout the twentieth century, and into the current era, but they have also remained far smaller than those in STEMM disciplines over time (Fortunato et al. 2018), and remain roughly half the STEMM size now (see Figure 16.6 below). However, the social sciences have clearly moved considerably away from the humanities where research remains predominantly a 'lone wolf', one person effort (Plume and van Weijen 2014) and multiauthor papers are also more common in the social sciences than in the humanities. In terms of specializations within teams, many STEMM outfits can perhaps assign one member to either lead on or handle completely data documentation issues, which would be rare in the social sciences, and not at all feasible in the humanities, beyond delegating mundane tasks to a short-term research assistant. Fortunato et al. (2018) also demonstrated that across all the discipline groups the papers produced by larger numbers of authors are more often highly cited, high-impact papers than is true for all papers taken as a whole. This may suggest that the ability for team members to undertake separate tasks, or the more collegial nature of multiauthorship, or both, helped increase the academic value of work done.

Having multiple members in research teams, and 'fractional' authors for articles and books, can have potential advantages for documenting research. If members are not multitasking they may have more time. There may also be more capability for one team member to act as the documentation expert who can be consulted on likely requirements and advise on issues that arise. There are also some possible drawbacks. Poor communication or 'siloing' between team members may hamper documentation as you go. In large teams in particular, assigning most documentation/recording roles to one member may seem to 'license' other team members to pay it less attention than they need to. This can be especially problematic if the nominated person is the most junior team member, or (even worse) the most recent or short-term research assistant on the project – likely to be someone who has the least background in the area, or who must also tread carefully around dominant research principal investigators.

File naming, codes and copy making

A first, but foundational, step in the documenting process is to make sure from the very start that everyone involved in the research is using the same appropriate and sustainable file-naming conventions for all copies or versions of the data – both originals or 'master' files that must not be changed, and working analysis files incorporating multiple edits and recodes as they evolve through multiple versions. Similarly, in reporting analyses file names should help researchers, giving immediate clues as to what results are being

reported, within any space limits on name lengths. It is important that file names should be:

- *distinctive* – avoid using file names that are mostly similar to each other, where only a few keystrokes separate one from another (increasing the risk of wrongly overwriting another file).
- *substantively descriptive* of their contents, giving real clues from the name alone what has been done or is being reported. Where multiple successive files do the same thing in different stages, and names reflect inclusion in a 'family', file names could include progress indicators (e.g. 'just started', or halfway', or 'nearly done') that help them stay distinctive while grouped together.
- *not reliant for distinctiveness just on date differences*. On their own, date names are easily mistyped and they are always obscure once some time has passed. All modern software already allows you to sort by date and time without putting these details in the file name.
- *archived or organized into folders in consistent ways,* for example using a 'Store' or 'Older versions' sub-folder to hold non-current files, and moving older files into archives consistently, so that your working folders stay accessible and easy to keep tabs on.

Also consider the following.

When choosing a new file name or assessing a set of existing file names, ask: how different are they? How many wrong keystrokes would it take for you to mistakenly overwrite an earlier file or load a wrong one? Whenever you start working again on an already existing file your first step should be to save it with a new progressed version number (e.g. 'version 1aaa' becomes 'vers 2bbb' and then 'version 3ccc') – so that from the outset of your new work session you run no risk of losing previous progress made.

As your research datasets and archive grow over time and across projects it will become ever more vital that all of the files, sources, notes and data you collect and store remain accessible and easy to find. A standardized system for how files are stored and archived is vital to keep things straightforward, and to avoid the later high time costs in trying to remember when you made changes (possibly years ago) and then 'rummaging' and searching for suddenly needed files at later stages.

As with all steps in documenting, there is a temptation that feeling 'mastery at the time' will lead you to assume that you will always know where to find the files you are looking for. However, over the course of a research career your personal archive will grow to encompass thousands and even hundreds of thousands of separate files, often multiple versions of the same piece of analysis or writing as it is reviewed, refined and edited for publication. Ensuring that you are returning to the right document and not repeating work that has already been done is crucial – as anyone who has entirely re-edited an earlier draft twice by accident will attest.

The schema for how you wish to catalogue your files is a deeply personal choice that will reflect what works best for you. No specific conventions, categories or formulae are applicable to all researchers. That said, two different high-level principles underlie most file naming and organizing systems – a library-like pattern (discussed here); or a full-text (search engine) approach (considered below). Organizing your files like a library means using a wide-ranging classification system to create a hierarchical system that mimics the codes one would use to find a book in a library. A code specifying an increasing level of specificity prefixes each file and folder name followed by a short description. Files can therefore easily be found by working down from the broad code to the specific, with the final number tracking the order in which files were created. Timothy says:

> This system allows me to easily know where to save and look up files by following the broad and comprehensive schema I have already created. In particular, I like that it enables me to easily to see where everything is located and browse between folders without being swamped by 1000s of files. Moreover, I have found that strict hierarchy of labelling files like a library allows me to work with files programmatically which can be very helpful for organising and tracking things like log files and automatic outputs. My system was broadly inspired by the johnny.decimal system created by the programmer Johnny Noble (2022, see https://johnnydecimal.com/) with some adaptations to fit my own research workflow. (See Figure 5.1).

Another, less strictly ordered, and potentially more flexible, way to organize your files is to think of accessing your archive as a search engine. In this case, it is still useful to have a rough schema of folders where the files are stored, but the main way you work with and find files is via searching for them by name components. For such a system to work it is imperative to use file names that are descriptive and allow you to easily see which is the most recent file. Such a system should have long descriptive file names, letting you know what each file is for, information on why the file was created or the research stage involved, and the version number of the file. With all this information, it will therefore be easy to search for the file you are looking for, and from the results, pick out the relevant version. Patrick says:

> My filenames tend to look like this:
>
> *Ch 5 Documenting Research late July 2023 version 22vvv* or
>
> *Data Science for PAPP resubmit early March corrected HM vers 2bbb*
>
> I always pick a file name that is intuitive to read and ends with a unique version number. I always track each version using a unique ascending number and three letters in lowercase (e.g. version 1aaa,

Figure 5.1: An example of how Timothy organizes his files

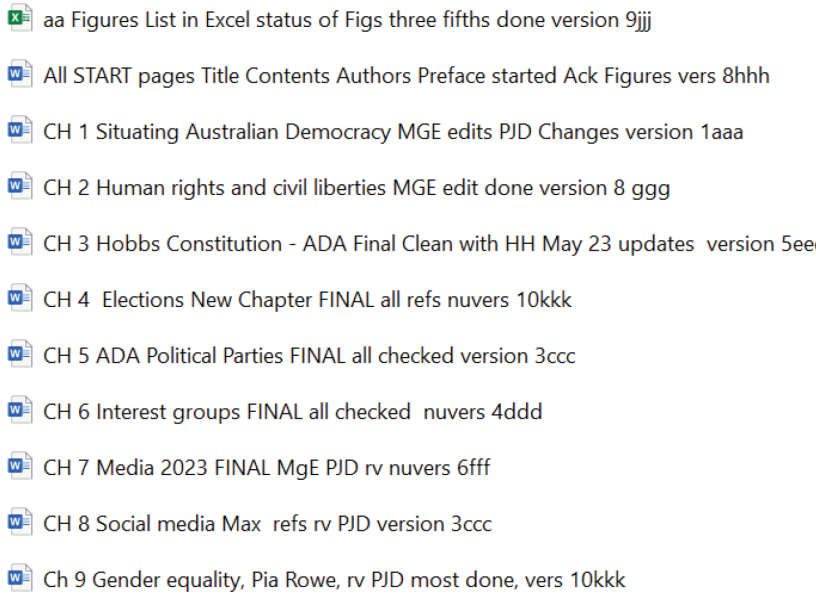


then I abbreviate 'version' to vers 2bbb, next is version 3ccc, and then vers 4bbb etc.). My names often start with a specific research stage or a chapter or section number (so they look organized in folder view) and then a description or short title, and perhaps a broad date (e.g. 'early May') when the file was worked on. I may also indicate a progress level within the file name (e.g. 'rv started' or 'checked halfway' to further increase the uniqueness of file names and reduce the risk of getting the wrong version. Sometimes it's helpful to even put a 'mood' note in a file title (such as 'frustrating' or 'very rough' that will help you remember what was going on then). If I am working with a co-author I include their initials in the names for any file they have sent over to me. I find that having these different elements makes it easy to find files quickly and harder to make mistakes (like inadvertently doing abortive work on an older version of a file that has already been revised). (See Figure 5.2)

Copying data

Backing up your files was once a laborious and time-consuming task. Yet, in the past, the failure to take this task seriously could result in the loss of whole unfinished books, theses or years' worth of research work. Today, however, backing up your files can be a seamless task. Multiple commercial and

Figure 5.2: An example of how Patrick organizes his files



open source offerings (such as OneDrive or Dropbox (proprietary), plus OwnCloud or NextCloud (open source)) can be set up to automatically back up files across secondary locations. They include file versioning systems that allow accidental deletions or overwrites to be undone. Many research institutions support and maintain one or more of these systems for their staff. We highly recommend that all researchers spend time setting up such software to automatically backup their files (twice *at least*). Take advice from organizational staff listed as data controllers who can advise on how to do backups and what files can or cannot be moved off university or workplace systems. There could be intellectual property right (IPR) issues in copying across to your own PC or datastore either commercial programs, or materials that the university may claim are their IPR alone. Some ethical issues may also arise (see below). PhDers and early career staff should especially take university or project team advice carefully here. But they should also ensure that if they move institutions they can take with them all their working and data files needed to complete publications and obligations, or to begin new lines of research.

When setting up an automatic backup system you should first be able to maintain access, migrate and share files throughout your career – during which you may work at many different institutions and collaborate with multiple sets of co-authors. Second, a backup system must be implemented in ways that take account of any sensitive or personal data used in your research. What these requirements are, and how your data is stored, should be covered by a data management plan – see Chapter 7. Where sensitive research materials must be handled it may be that you personally maintain

one kind of automatic backup system for your own personal files – notes, code, articles – and another provided by your institution, which stores the sensitive data used in your research. A well-documented research process will ensure that an appropriate version of your work is sharable with others and to your future self. As well as an inherent clarity, transparency and intra-operability that enables them to share and collaborate with others, each researcher's system of documenting research should also be personal to them, allowing them to take forward the skills and knowledge they have developed into new ventures.

5.2 Developing appropriate approaches for your research

How documentation is done is shaped by many combined influences – your field of study, topic of interests, doctoral training, colleagues' influence, personal preferences, working habits and practical concerns. Practice develops through doing and accumulates and is refined throughout your research career. So the exact system that works best for you cannot be taught – you will need to discover it for yourself. However, some broad principles and tools will be useful to all researchers. Similarly, it can be valuable to ask other researchers how they document and organize their work. Much can be learnt from other people, usually passed on informally in the lab, over coffee, or at the bar. We suggest five key questions:

- How will you be able to access your research, notes, data and files in the future?
- Will you need to pay/have an institutional affiliation to access and use these resources? (Don't just hold your data and records in expensive proprietary software that your next employer may not use, or that you cannot afford to access easily.)
- Will you be able to understand what you have done from the documentation, explanations of variable and value labels included and analysis steps made? Are any abbreviations used or naming clues currently included, and are they durable ones that a future self will understand?
- Do the software programs (and any other external resources) relied on have long-term support? Programs that are large and well used are always likely to be more translatable into currently unforeseeable new formats that will inevitably come along. But will that be true for smaller programs?

Documenting your research should cumulate in the creation of a personal research archive, a repository for the work you have done that you can draw on throughout your career. Achieving a long-term solution on how you document your research means taking a precautionary view of the messy and complicated research life cycle.

While it can be helpful to conceive of research projects as discrete pieces of work, the timescale on which they take place, and the work required often means that one effort blends into successor or alternative projects. For example, in 2021 the average economics article took 822 days to move from being uploaded to a preprint server (in pretty final form already) to full publication in a peer-reviewed journal (Xie, Shen and Wang 2021). During this time researchers will both need to move onto other projects and return to old work. To this end, the route from finishing primary research to getting a final publication can be seen as ongoing maintenance work. Text is redrafted, figures and data are revisited, and often additional analysis is conducted, as part of the work of writing, editing and navigating research through peer review. To easily return to work (perhaps done several years ago) and revisit, revise and redo parts of it, good initial documentation is vital, along with previous data being reliably stored and reaccessible in future.

Scratch pads and short-run records

Lots of things need to be written down, but not everything needs to be stored and systematized, so it is worth starting with what usually does not need to be too organized. 'To do' lists, quick notes from a meeting, emails to be sent, references to be looked up, and many other small tasks all need to be written down somewhere and kept for a while, but their purpose is ephemeral. Counter-intuitively, having a 'scratch pad' where you can group these notes together and then move on from them once they are done is the first step to more documenting your research permanently.

A scratch pad can be anything you want it to be, from a paper notebook to post-it notes stuck around your desk, or their digital equivalents. A common approach is to use a hardback notebook to log everything that needs recording for a while in a simple and eclectic data order sequence, and then go through later, ruling out tasks completed, or materials transferred into more organized files. Almost all PC notes software can operate this way too, but with faster and easier decisions later on concerning what to file by what subjects or headings. The immediate purpose of the scratch pad is to make a basic record of what you are doing, that can then be organized and moved into your system of documentation when the structure and labelling needed for the longer term are clear to you. And, when you are done with scratch pad materials, they can be discarded. Knowing that these notes are (mostly) transitory can help you think about what does and does not need to be stored properly. If your notepad was destroyed, would you be distraught about the research notes you had lost or be unperturbed? If it is the former then these notes need to be transferred into a proper system of documentation where they can be backed up and retrieved where needed. If you would be unbothered, and could simply start a new page, scribble down a fresh to do list, or leave a different post-it note reminding yourself

to save the PDFs you are reading or finish another current task, then you are using a scratch pad correctly.

A scratch pad can also help guard against the false expectation that everything done in the research process can or should be documented. Notes from a quick meeting with a collaborator on tweaking a presentation graph or submitting an abstract to a conference probably do not need to be stored in perpetuity. By contrast, it can be incredibly helpful to have typed-up minutes from an early meeting with the same collaborator where the scope of the research project was discussed, core research questions were refined, or lasting innovations in analysis categories were fixed. Computerized systems make it possible to record and archive everything that you do. But this just makes distinguishing between what is important to record, and what is not, particularly important – because, as the scope and size of your documentation grows, so too does it become technically more complex – harder to keep organized and unwieldy to use.

In moving beyond the scratch pad to create an enduring system for documenting research, choose carefully. The best system for documenting your research is not necessarily the most comprehensive, technically advanced, or expensive one – but the one that is the most straightforward (quickest) and intuitive method for you. This could be relatively simple and low tech. For instance, an individual researcher working in an archive could make handwritten notes when there, of just the reference numbers of folders and files, then photograph these notes pages, and store them in folders organized by date on a PC, along with a few 'reminder' keys or tags to assist later recall. They also could maintain a separate Word document with a log of archive reference codes examined each day. Finding their thoughts on a specific box at the archive years later would then be a relatively straightforward process of opening up the log file, searching for the archive code and the folder of that day, and finding the relevant pages of notes. By contrast a more 'perfect' digitization system, which required all notes to be typed up, imported to qualitative research software and then correctly tagged by topic and content, may seem ideal – but not if it breaks down or becomes erratic because of all the extra work required by the documentation process itself. Only if full text retrieval was likely to be needed a lot (e.g. because new and unforeseen topics crop up) might the extra work involved be worthwhile.

However, in other cases a more complex, higher-tech system may best fit the project's needs. For instance, long-running research involving complex statistical analysis might benefit from making use of software development tools (like Git) that allow multiple collaborators to work on the project at the same time, or for new analysis to be branched off in different directions, or for changes to be rolled back to earlier versions and rerun again on new data. In this case the upfront cost of each team member learning to use new software, and ensuring that all changes are correctly recorded and added to the project in small incremental steps, will be repaid as the work involved rapidly develops and becomes more unwieldy.

Logging what you did, when and why

Keeping an enduring record of research should cover:

- how and when each wave or activity was carried out, because that exact time context can often shape survey responses, people's objective behaviours and other relevant variables for field research;
- the data gathered;
- what was found out; and
- the researcher's contemporaneous thoughts, comments, understandings and analysis (which may be evanescent).

Such records have always been a fundamental part of the scientific method. The crucial question to ask is, 'In the future, what will you – or other authors or team members – need to know to understand the work you have done?' It is vital to move away from the illusory pursuit of a perfect system of documenting, and to transition to one proportionate to the needs of the research at hand.

An equally illusory belief is that documentation needs to meticulously record *everything* that was done. This falls into the opposite trap, by making the 'transaction costs' of documenting so high that they work against behaving consistently, instead creating backlogs and time pressures that actually encourage researchers to skip things that really need recording. Do you regularly create graphs in R? If so, then explaining in depth every line of code that makes the final visual is probably not necessary. However, spending time commenting on and justifying why you dropped certain data points earlier on in the analysis may be crucially important. The same is true of qualitative work. Making detailed notes on every document requested from an archive may be an unrealistic aim. But keeping a running log tracking what was in each document, and why it did or it did not warrant further investigation, can be crucial in helping you to retrace your steps at publication time, or even return to working in the same archive later on to meet reviewer comments.

A straightforward log file, where you keep simple bullet point descriptions of work done and decisions made, is an easy way to your research process. Including dates, file names and other findable or retrievable information makes your log file easily searchable, especially revisiting to look up otherwise potentially elusive information. For example, when did you move from using v1 to v2 of a variable in the analysis? Why was v6 recoded midway through analysis, and which other variable was it subsequently labelled as? And, on occasion, why did you feel it was necessary to start over again? You can remember exploring a potentially interesting dataset just before the start of a given busy semester, but even by the end of that time what information was used and where you found it may elude you. Like even the briefest of diary notes, a simple log file can put you back into the mindset of what you were doing at the time.

In quantitative work creating a log file can be interpreted literally as commenting on the code you are writing with plain language explanations of

what is being done, that make the work comprehensible. As Figure 5.3 shows, though, in a rush of ongoing work it is all too easy to make only enigmatic or bald comments that are pretty obscure if you later revisit that bit of code. Figure 5.4 shows that better comments are that bit fuller and clearer to read in retrospect.

Figure 5.3: Poorly commented code may not elucidate what is being done – but makes it harder to understand

```
# seperating out entires with dashes - god this code is becoming a mess

dash = land[land['Property Address'].str.contains('-') == True]

dash_df_2 = pd.DataFrame()

for index in dash.index:

    row = dash.loc[index]

    addr = row['Property Address']
    addr = re.sub(' ', '', addr)

    # taking out postcode which otherwise doesn't work for some reason
    postcode = row['Postcode']
    postcode = re.sub(' ', '', postcode)

    addr = re.sub(postcode, '', addr)

    addr = re.split('-', addr)
```

Figure 5.4: Well-commented-on code explains what is being done, why it is being done and the choices that were made at the time, matching the author's existing skill level

```
    # iterate through all the rows in the xlsm file
    for n in range(sh.nrows):

        # find the value contained in each cell
        cell_val = sh.cell_value(n,0) # 0 specifies column A, other cols not needed
        # extract the style infomation from the same cell
        cell_xf = book.xf_list[sh.cell_xf_index(n,0)]

        # conditional on font weight parsed through fontstyles for xlms sheet
        # 700 is bold (as save as rtf format text rather than excel native, this
        # is why other python excel parsers did not work)
        if font[cell_xf.font_index].weight == 700:

            # add the cell value and location to the dataframe
            bold_df = pd.concat([bold_df,
                                 pd.DataFrame([cell_val, n]).T])

    #format the outputted information
    bold_df.columns = ['BoldVal', 'BoldLoc']
    bold_df = bold_df.reset_index(drop=True)
```

For qualitative research, logging key documentation may mean recording the schema for coding interviews, the search strategy used to find documents and your own contemporaneous reflections on what survey questions worked well, or on refining the case selection. For a larger, multiple collaborator project, or one that you expect to be used by many other researchers, more 'official' documentation may be needed, laying out all aspects of the project or information gathered comprehensively in a separate document. A log file makes a good supplement to the other notes and analysis files that you produce as part of your research, but it is most effective when paired with effective documentation of your research itself in extended notebooks, especially digital ones.

Open notebooks

In a refresh of the traditional pen and paper lab book, many STEMM and quantitative social science researchers have shifted to using electronic lab notebook software (such as Jupyter, discussed below). These programs seek to build logs of 'science in action' that are fully accessible by follow-on researchers, and allow research to be more easily examined, rerun and modified. They allow text, computer code, raw data, results and data visualizations to be contained within a single file that can easily be shared and understood by their fellow researchers.

Within the lab notebook file, the content is then split into many smaller self-contained cells that are organized in a linear structure labelled with headings and titles. Each cell can contain code, text or results. For example, a notebook can start with a title and text passage laying out the research questions and strategy for analysis. Step-by-step computer code in following cells can then show how to download the data and follow the same analysis path. Particular attention is normally given to breaking up the steps undertaken in the cleaning, examination, and initial sense-finding phases, along with developing data displays and evaluation. Each part of the process can be interspersed with written comments, discussing choice options, analysis limits, possible objections and justifications. Electronic notebooks seek to seamlessly record research as it happens, with no delays or backlogs of unsaved materials, and then to document it for sharing saved copies with others at no extra cost in time or effort (see Figure 5.5). Many scientists using notebooks apply the idea of 'literate programming' (Knuth 1992) in which code is written with the *primary* intention of being human readable and easy to understand. For a very good illustration see the LIGO and Virgo Scientific Collaborations (2020) and also Abbott et al. (2020). Publishing notebooks showing the order and progression of the research is also good OS practice, 'telling a story' to readers in a way that fosters easier interactions with the code, results and data visualizations (Rule et al. 2019).

E-notebooks create a single, self-contained file that is easily made publicly hostable or shared with other specific researchers. Reusers can simply

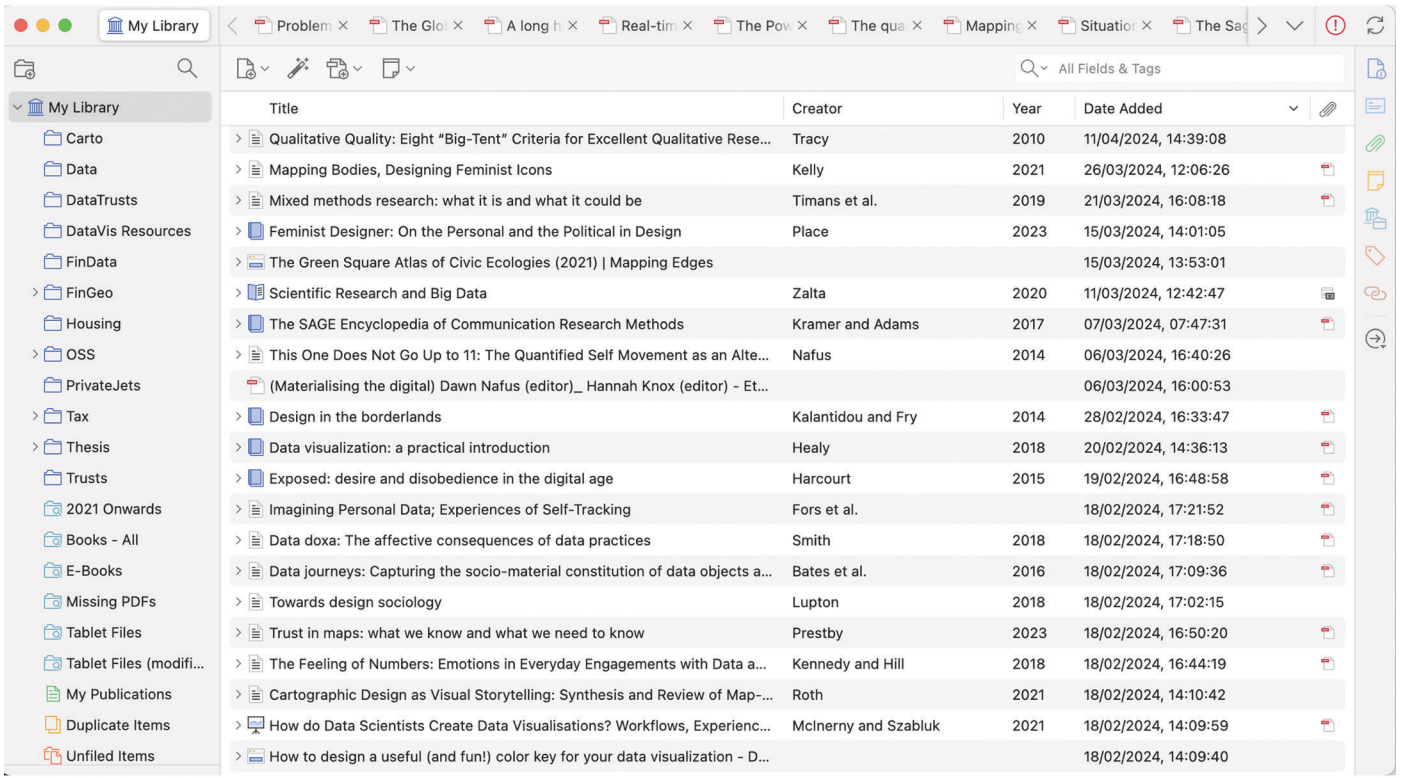
download a notebook to their PC and rerun the same code on the same data (assuming it too is publicly hosted), fostering instantly reproducible research. Other technical efforts (such as Docker (proprietary software) and Binder (open source)) aim to ensure that the computing environments on which notebooks run remain replicable and archivable – a key element since operating systems and software libraries are regularly updated. This intensive reproducibility also creates an inherent boost to replicability. All the code needed for analysis is contained within the notebook, and each step is split into separate cells, so modification of the existing code and workflows is easier. For example, altering a variable to see how it affects the results can be as simple as editing a single line of code, rather than a researcher having to reproduce the same analysis from scratch.

Reference management software

No matter what your discipline or methodology is, drawing on and reacting to the work of other scholars is a fundamental part of research. Without literature reviews, it is impossible to formulate a sensible research question or conduct research with any appreciation of what it may contribute to your discipline. And covering relevant literature is a key requirement in writing up research findings for publication (Chicago University Library 2024). So, referencing appropriately is an integral part of demonstrating scholarly involvement and competence, as the sociologist Robert Merton emphasized (Dunleavy and Tinkler 2020, Chapters 1 and 2). Yet some social sciences and most humanities disciplines still suffer from chronic under-citation (Dunleavy 2014). Some fields are still plagued by evident selection biases in what gets cited. These include 'me-tooism' (citing only relevant works that authors agree with or that are helpful for their argument); silo biases (citing only other people in your home discipline, and omitting relevant evidence from other subjects, a pattern rife in economics); and of course ethno-centrism (citing only authors from one country, language group or continent). Unsurprisingly, despite the popularity of well-known packages (Chicago University Library 2024) recent research has found that only one in five scholars in the social sciences use any form of reference management software (Watchorn 2022).

This slow uptake may be explained (in part) by misleading labelling. Contemporary referencing management software is far more sophisticated and capable of performing many more tasks than its name suggests. Modern software can be used to seamlessly create an archive of your reading and potential sources for research. For example, Zotero is the leading open source referencing management tool and can both capture bibliographical information for an article or book and store a PDF or HTML copy. (See Figure 5.5). Texts can be read within Zotero using the built-in PDF reader, which will also record highlighting and notes made, along with any other digital resource attached to items. Zotero can also be set to automatically back up all these extra resources, ensuring that

Figure 5.5: An example of Timothy Monteath's Zotero Library, sorted by date



they remain accessible in the future. And when it comes to writing up articles or books bibliographical information can be easily added into your word processing software in the correct format needed by a given journal or publisher.

Using Zotero or an alternative helps build up a personal research working library, providing comprehensive future access to all the things you have read. Files can be set to be automatically renamed as they are updated, keywords from the meta-data can be searched, and you can see a timeline of when files were added to your library. Because personal notes, observations and reflections are stored alongside the materials they relate to, you can return to the reading done for research at any time in future – so integrating multiple capabilities into the process of writing, rewriting and referencing effectively.

Conclusions

Most of this chapter does not explicitly link documenting your research to OSS, even though this is in fact a foundational step for all OS endeavours. The simple reason is that documenting is a critical foundation of *any* research, however open or closed the methods used. Better documenting as you go translates into better research (and an easier life for researchers) across the piece – an improved research process over the project's life, ensuring that the underlying data, analysis and code is already of a high standard, fostering better team collaboration, avoiding duplication of effort or confusion about what has been done, and greatly easing the path to publication and implementation of changes needed after reviews. The OSS bonus then adds on top of all that, by making the research undertaken much more legible and transparent for readers and reusers, and easily sharable with other researchers. So there are no extra burdens in documenting well, and there is no 'friction' at all between the goals of OSS and the practicalities of doing good research – the two are overwhelmingly complementary.

Annex 5A: Software resources

Zotero: https://www.zotero.org
A free, open source, reference manager (Harvard Library 2025). It allows you to easily 'collect, organise, annotate, cite, and share research'. It integrates with web browser plugins and can automatically collect PDFs, snapshots of websites and bibliographical information from almost all scholarly sources. It runs on all computer operating systems and includes tablet apps, as well as plugins that can connect to other software and word processors. It is actively developed and maintained by the Corporation for Digital Scholarship, a non-profit 501(c), which ensures its long-term support and independence. See https://perma.cc/3WGL-T3U8

Jupyter: https://jupyter.org
Open source software that allows code to be run in a 'notebook'-style environment. It is widely used in data science and scientific computing and supports all major programming languages, including R and Python. Notebooks can be run on computers either locally or remotely, connected to via a browser and easily published or shared with other researchers. Jupyter is actively developed and maintained by its own executive committee and is financially supported by the NumFocus non-profit ('Better Code = Better Science'), which supports a wide range of scientific Python open source software (Code Academy Team, no date). It has received funding from Microsoft, Anaconda, MIT and numerous other corporate, charitable and research organizations. See https://perma.cc/PD7H-6JSZ

Git: https://git-scm.com
Open source software that implements a 'distributed version control system'. It is primarily used for software development and is capable of managing very large and complex projects which are worked on simultaneously by multiple contributors. Git is technically complex to learn and use, but can be incredibly powerful when used correctly and for the right project. Git repositories can be hosted online. A range of for-profit firms (such as github.com) and non-profits (such as gitea.com) provide easy-to-use interfaces and web access. Git is a core part of the open source software community and is supported by the not-for-profit organization Software Freedom Conservancy. See https://perma.cc/4K7C-483F

OwnCloud and NextCloud: https://owncloud.com – https://nextcloud.com
These are open source software alternatives to the likes of Dropbox or OneDrive. They allow individuals and organizations to host their own automatic file backups which can sync across multiple devices and servers. They can be run at home or be scaled to thousands of users, and storage is only limited by the hardware they are run on. Many universities and research organizations are now using their software to ensure that data collected for research complies with data privacy and storage legislation. Both are developed and maintained by private companies: their business models rely on selling hosting and support to organizational partners.

Binder: https://mybinder.org
Project Binder is an open source tool that enables code to run in a computing environment that replicates the one in which the original code was written and run (Karoune et al. no date). It is developed by a subset of the same team that develops Jupyter and is primarily focused on creating an open alternative to commercial emulation environments, such as Docker, that is tailored towards scientific research. Its development has been supported by the Alan Turing Institute, UC Berkeley and other research institutions. (See Karoune et al., no date.)

Note

* This chapter draws on research and seminars forming part of the CIVICA Research programme, which was funded in 2022–2024 by the European Union's research and innovation programme Horizon 2020, under the chapter 'Science in Society'.

References

Abbott, B.P.; Abbott, R.; Abbott, T.D.; Abraham, S.; Acernese, F. et al. (2020) 'A guide to LIGO–Virgo detector noise and extraction of transient gravitational-wave signals', *Classical and Quantum Gravity*, vol. 37, article 055002. https://doi.org/10.1088/1361-6382/ab685e

Aly, Mariam (2018) 'The key to a happy lab life is in the manual', *Nature*, vol. 561, no. 7721, p. 7. $ https://doi.org/10.1038/d41586-018-06167-w

Bastow, Simon; Dunleavy, Patrick; and Tinkler, Jane (2014) *The Impact of the Social Sciences: How Academics and their Research Make a Difference*. London: Sage. $ https://doi.org/10.4135/9781473921511

Chicago University Library (2024) 'Citation management', web page, 8 October. https://perma.cc/Q8PG-RLGV

Codeacademy Team (no date) 'How to use Jupyter Notebooks'. https://perma.cc/PD7H-6JSZ

Collins, Harry. (2010) *Tacit and Explicit Knowledge*, Chicago, IL: University of Chicago Press. $ https://perma.cc/A9DE-9QGR

Dunleavy, Patrick (2003) *Authoring a PhD: How to Plan, Draft, Write and Finish a Doctoral Dissertation or Thesis*, London: Palgrave (now Bloomsbury). $ https://www.bloomsbury.com/uk/authoring-a-phd-9780230802087/

Dunleavy, Patrick (2014) 'Poor citation practices are continuing to harm the humanities and social sciences', *Impact of Social Sciences*, LSE blog, 9 December. https://perma.cc/U8VF-BAUX

Dunleavy, Patrick and Tinkler, Jane (2020) *Maximizing the Impacts of Academic Research*, London: Palgrave (now Bloomsbury). $ https://www.bloomsbury.com/uk/maximizing-the-impacts-of-academic-research-9780230377608/

Fortunato, Santo, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen et al. (2018) 'Science of Science'. *Science*, vol. 359 (6379): article no. eaao0185. $ https://doi.org/10.1126/science.aao0185

Harvard Library (2025) 'Research Guides: Zotero', updated 28 January. https://perma.cc/3WGL-T3U8

Herndon, T.; Ash, M.; and Pollin, R. (2014) 'Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff', *Cambridge Journal of Economics*, vol. 38, no. 2, pp. 257–79. $ https://doi.org/10.1093/cje/bet075

Karoune, Emma; Gibson, Sarah; Bachelor, Sophia; and Vilas, Martina (no date) 'Top tips for making the most out of Binder', Blog post from Software Sustainability Institute. https://perma.cc/C6UL-43UY

Knuth, Donald E. (1992) *Literate Programming* (1st edn), Stanford, CA: The Center for the Study of Language and Information Publications. https://perma.cc/PQ39-S5JX

LIGO and Virgo Scientific Collaborations (2020) 'A guide to LIGO–Virgo detector noise and extraction of transient gravitational-wave signals', *Classical and Quantum Gravity*, vol. 37, no. 5, article 055002. https://doi.org/10.1088/1361-6382/ab685e. For the notebook pictured see: https://perma.cc/B58M-E6C3

Noble, J. (2022) 'Johnny Decimal**:** a system to organise your life', Blog post, 21 April 2022. https://perma.cc/WM3K-UH7B

Pinker, Stephen (2014a) *The Sense of Style: The Thinking Person's Guide to Writing in the 21st Century*, New York, NY: Penguin. **$** https://perma.cc/2QGK-7DA6

Pinker, Stephen (2014b) 'Why academics stink at writing', *The Chronicle Review*, 26 September. **$** https://www.chronicle.com/article/why-academics-stink-at-writing/

Plume, Andrew and van Weijen, Daphne (2014) 'Publish or perish? The rise of the fractional author', *Research Trends*, vol. 1, no. 38, article 5. https://www.researchtrends.com/researchtrends/vol1/iss38/5

Reinhart, Carmen M. and Rogoff, Kenneth S. (2010) 'Growth in a time of debt'. *American Economic Review*, vol. 100, no. 2, pp. 573–578. $ https://doi.org/10.1257/aer.100.2.573

Rule, Adam et al. (2019) 'Ten Simple rules for writing and sharing computational analyses in Jupyter Notebooks', *PLOS Computational Biology*, vol. 15, no. 7, article e1007007. https://doi.org/10.1371/journal.pcbi.1007007

Shawn, Graham; Gupta, Nehra; Smith, Jolene; Angourakis, Andreas; Reinhard, Andrew et al. (2020) *The Open Digital Archaeology Textbook*. OA: https://o-date.github.io/draft/book/making-data-useful.html

Watchorn, Deirdre (2022) 'What does open science mean for disciplines where pen and paper are still the main working methods?' *Impact of Social Sciences*, LSE blog, 21 April. https://perma.cc/FA5A-H9W6

Xie, Boya; Shen, Zhihong; and Wang, Kuansan (2021) 'Is preprint the future of science? A thirty year journey of online preprint services', *arXiv*. https://doi.org/10.48550/arXiv.2102.09066

6. Reusing and mashing other people's data

Anything ready-made sounds attractive. It should be quicker and easier to work with, and avoid duplicating unnecessary tasks or previous work done, and above all it saves the costs of original data gathering. Of course, there are many different types of data available, and complications involved in reusing them. So trying to provide a comprehensive overview for even one discipline, let alone all the social sciences, would be a lengthy undertaking, and could quickly go out of date. Instead, we focus in a systematic way on some generic issues, applicable to any digitized secondary data source, considering some high-level questions and ways of engaging with data reuse.*

We begin by considering the difficulties of finding out what relevant datasets already exist and assessing how far existing data can fruitfully be brought to bear on your precise topic, which is at least partly new and different (Section 6.1). Evaluating possible data sources always involves going into details – for example, looking at the exact form of questions survey respondents were asked, or the specific nature of actor's digital social behaviour being tracked (Section 6.2). And many secondary datasets cannot be used as initially downloaded, but often need to be 'cleaned'. This activity is almost always as painstaking, taxing and boring as its housework namesake (Section 6.3). Finally, combining datasets is not straightforward. It can be done in many different ways, and needs some careful judgement if it is to be implemented appropriately (Section 6.4).

6.1 Should you reuse data?

Researchers who learn of 'possible' dataset relevant for their needs or addressing their interests need to ask five questions before resolving to use other people's data:

- What is the relevance of this dataset to my study? Can it help advance *specific* research questions that are important to me, and why? Are there also perhaps *faut de mieux* reasons to try reusing data (e.g. no better resources are available to you)?

How to cite this book chapter:

Dunleavy, Patrick and Monteath, Timothy (2026) 'Reusing and mashing other people's data', in Dunleavy, Patrick and Monteath, Timothy (2026) *Doing Open Social Science: A Guide for Researchers.* London: LSE Press, pp. 129–148. https://doi.org/10.31389/lsepress.dos.f

- Which specific parts of a dataset can be reused, and what advantages or time savings will flow from this choice? Are there constraints or limits set on how data can be analysed by reusers?
- Are there any problems with how the existing dataset specifies variables of interest for my specific purposes? Can they be worked around? Is it feasible to combine an existing variable of limited use with other data to create new variables that are better 'proxies' for issues of interest – that may be only partially or tangentially addressed in previous work?
- What data analysis skills and resources are needed to carry out this work? If they are not already present in the team or in your own repertoire, would it make sense to devote time to getting them (e.g. by learning to use new analysis software or techniques needed for a particular dataset)? Is this the best available use of team/personal time or resources?
- How will reusing datasets fit with your long-run research interests and career trajectory?

Before plunging into engagement with a particular dataset, consider if it makes sense within your development as a researcher. It often may, especially for funding-poor new researchers like PhDers, post docs or other early career researchers. Established research teams applying for limited grants or constrained by their current funding also have strong incentives not to 'reinvent the wheel'.

Thinking carefully about the alignment between these elements can help researchers to weigh up the suitability of reusing data for their research and avoid setbacks once major effort has been committed if problems emerge later on (as some surely will). For example, many academics possess the statistical and technical skills needed to implement a complex analysis of quantitative data at the professional level. However, without a theory that they are seeking to advance, or an understanding of how stats package tests and processes apply in the context when the data was collected, or awareness of how this relates to wider disciplinary concerns, a decontextualized statistical analysis will fall flat. Conversely, some researchers with the wider knowledge needed to formulate a research question exploiting secondary data in a theoretically informed way that advances the literature may lack the analysis skills needed to implement them at a practical level.

Of course, how to put a dataset to productive use can in principle be learned. But does the time commitment and difficulty of that task make it feasible to learn a new literature or technical skills? For a PhDer or early career researcher it may make sense to invest major time in developing knowledge or skills that will serve them well in future, going beyond the initial data reuse project. By contrast, a more established researcher with a developed relevant competence may see potential gains from reusing a dataset but lack the ability to conduct every aspect of the analysis for themselves, or the funding needed to employ a skilled researcher to help them. So committing to developing new knowledge

or skills here may distract from an already established research competence. Instead, looking for collaborators to complement existing strengths may be a better strategy for enabling data reuse here.

Initial prospecting for possible sources should also consider practical issues around accessing someone else's data and putting it to work. There are clear advantages in reusing readily accessible data without the time and costs of collecting it from scratch. However, there may be other costs in reusing data – such as delays in getting access from authors or securing access to a dataset that requires special permissions, plus concerns such as data storage or compute time with large datasets. And if the accessible information available is not *exactly* what is needed, coping with its limitations may suck in unexpected efforts that can mount up disproportionately to the gains. So, on a first dig around, be a little sceptical and realistic in determining whether reusing someone else's data is right for developing new research.

Finding feasible data to reuse

Whether you already have a specific data source in mind, or only a general idea of what data could help, it is worth investing some time in exploring the whole field of potentially relevant information available. It is especially key for new researchers in any area to find the optimal data source for research purposes, rather than just lighting on an 'acceptable' source that pops up early in the search process. In addition, the formal and informal sources helpful for finding relevant research datasets may not be obvious and may be unique to your discipline and sub-field of interest. Supervisors, colleagues and research collaborators are obvious people to start conversations with about useful sources or other teams active in your area, and you should chat widely with others within your own university, at conferences and seminars, or any social media networks you are part of – which are often unexpectedly helpful in response to a well-framed and reasonably precise question asking for help (see below). And for social scientists it is useful to tap experts or academic liaison people within national libraries, parliamentary libraries and relevant external archives. Even large government agencies or sometimes large corporations may have their own librarians or archivists who can sometimes be prevailed on to give valuable advice. In asking around, aim to 'snowball' recommendations of where to look for reusable data, other people's practical experience with datasets and questions, positives or pitfalls to look out for, and who else to talk too.

It is useful to also 'beat the bushes' with social media contacts, both about what is out there, and later on about how to evaluate the advantages and limitations of a particular source you have in mind. Being active within online academic communities is crucial here, since more experienced researchers may be good at giving short tips on other work or places to check. There are varying norms across sub-disciplines in the frankness of discussing the realities of research in public, so never plunge in with omnibus or overly large (hard to answer) queries. Instead use specific requests that people with

answers can comment on quickly, seeking to generate a conversation that can then range more widely on people's experience of specific datasets. For starter knowledge, academic search engines like Google Scholar, Web of Science and OpenAlex can be a good place to start. Increasingly, large language models (LLMs) like ChatGPT, Grok and scite.ai are being used to digest and regurgitate massive quantities of data – indeed, their results are now baked into conventional search engines like Google. Artificial intelligence (AI)-supplemented search may help surface useful avenues for further research, but any system's 'answers' always need further checking and verification, especially for news items (Jaźwińska and Chandrasekar 2025). Where it is possible to quickly download and probe accessible datasets in an exploratory way, then getting a 'feel' for the data and how you might work with it, can help generate specific feedback. It can also enrich the lessons you get from other people's often condensed (sometimes even 'gnomic') advice, whether on social media or in a bar or restaurant.

The next in-depth exploring stage has two differing search strategies – acting either as a *reverse engineer* or as an *explorer* – depending on the position you are starting from. In either case it is important from now on to keep a record of what you searched for, what you found and the thoughts and comments that occurred to you along the way.

Acting as a reverse engineer – If you already have a specific research question in mind this approach can quickly narrow down search results. Before you start intensive searching, first describe your research question on paper, and next describe the perfect dataset that would allow you to best answer it. Think freely and directly. Do not be constrained at first by what you know of what has already been done, the practical limits or resources that would be needed to assemble this data, or even how complex your final analysis would be if you could access and analyse all the data needed in optimal form. What would this perfect dataset contain? How would it have been collected? What variables would you be particularly interested in and how would you use them? The more detail you can capture here the better you will understand your research question and the data needed for a complete answer.

Of course, no perfect dataset does or will actually exist (even if you have resources to try and create it). The point of this exercise is to 'envision' what you are looking for. Then, you can search to find datasets that approximate the ideal or are similar to it in some respects – where reverse engineering some of that existing information might get close to some of what you need. Knowing that there will always be a gap between an ideal source and the limitations of data already created by someone else, should allow you to better critically evaluate what you find. Consider what you can (reasonably) do to narrow this gap. Could you use existing variables' data to construct new proxy variables for things you want to measure but are not directly addressed? Can datasets be combined to better approximate a 'perfect' dataset? Could additional data be collected to supplement existing data or to give extra assurance on how to interpret response patterns and the meanings of questions? Finally, what

compromises are feasible? Does adding some components at the expense of others take you closer to your 'perfect' dataset or further away? And at what point does the gap or the compromises you make undermine your ability to answer your research question? Working through as a reverse engineer can help clarify how to evaluate these questions while remaining tightly focused on your own research aims.

Acting as an explorer – The aim here is not to start with a specific research question in mind, but to develop one inductively from the datasets you find as part of your search. If you only know the kind of study you want to do, and the sort of question you want to tackle, this strategy will generally help to keep an open mind about what you are looking for and what you might do with viable data that you find. When searching widely, it is important to sketch possible research questions that develop and iterate in response to what you find. It can be helpful to do this comparatively if different datasets and authors come into frame. In picking between two potentially viable datasets X and Y, run a competition to surface their different (likely quite disparate) strengths or drawbacks. Write out and consider a comprehensive list of pros and cons for each contender dataset before deciding to focus on one.

An explorer search will draw attention towards particular topics and datasets that sync best with researchers' existing knowledge and focus. It can help to ask, 'What can I do with this data that the original authors did not explore? And, why did the original authors not look at this question themselves?' Normally they were just looking at different questions. But perhaps the question you want to ask was too hard to tackle? Approaching data from a different theoretical angle can add value and suggest new analysis pathways. Alternatively, viable analysis can be helped by having a comparative advantage in knowledge or alternative methodologies, or by new analytic developments that allow the data to be re-evaluated or analysed in a different way. You must be able to use data originally created to address someone else's research question for a purpose that you develop and is distinctly your own.

Approaching the search as an explorer and understanding the best available data gaps will help develop better research questions, making a novel contribution to a wider field and speaking to other researchers' concerns. And whichever approach you use, of course, you are not limited to the variables included in a single dataset. You may be able to bring two different datasets together either through merging them (Section 6.4) or by contrasting them in ways that help expand a research question or perspective going beyond the parameters of the original research.

6.2 Evaluating data sources

In reusing data, a first question to consider is how it was collected. *Reactive data* is usually collected with a specific (research) purpose in mind and requires the active involvement of its participants, so that collecting it is often

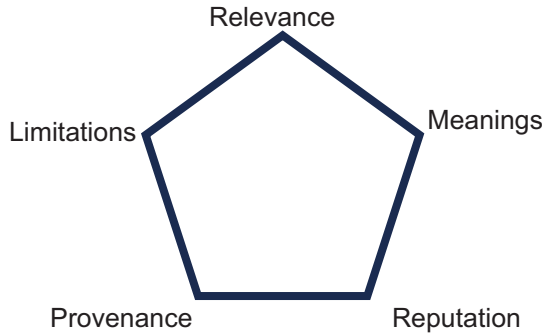
costly and time-consuming. Consequently, sample or dataset numbers in reactive data sources tend to be smaller than their non-reactive counterparts. For instance, in most experimental social science data the *N* sizes are often comparatively small – because participants need to be recruited (and often paid), experimental conditions established, consent granted and a wide range of variables recorded – many of which are needed as statistical controls, such as demographic information. Similarly, interviews, focus groups and deliberative assemblies usually involve small numbers of people as subjects. The largest reactive datasets come from surveys, whose size can vary greatly, from dozens or hundreds of responses collected on a one-off occasion by a single researcher, to thousands of responses collected over time or cross-nationally and on a regular basis by teams of established and well-funded researchers (both academic and external), such as official data surveys. In all these cases, the research purpose of the data is clear and the precise wording of the questions asked, prompts used or other framing given before subjects respond, strongly conditions the meanings of any data generated.

Non-reactive data is collected passively, without the active participation of the participants in shaping it specifically. Most non-reactive behavioural data is collected with a commercial or official purpose in mind rather than just for social scientific research. Transactional data for administrative purposes is routinely gathered by states and firms as part of their record keeping and delivery of services. In other cases non-reactive data is collected as by product of facilitating other social or organizational infrastructure. The passive, observed nature of non-reactive data does not mean that it is collected without consent but rather that it is gathered without specific permissions each time, and often unobtrusively. For example, phone or social media firms' contractual conditions give them a general access to user data. In the digital era, participating in any market tends to create data – for example, information on house sale prices is a by-product of housing market operations. In many states this information is recorded by the state or a company and made publicly available for administrative purposes. The cost of a single house may not be particularly useful or interesting to social scientists, but in aggregate, land registry data on house prices across the country can be extremely useful to researchers seeking to understand a wide range of social phenomena.

The growth of 'big qualitative data' has also added a whole class of systemic information that does not fit the standard quantitative pattern of a rectangular, tabulated dataset (Chandrasekar et al. 2024; Brower et al. 2019). There are now many non-numerical, text-based or video- or audio-based large datasets, and in Chapters 9–12 we explore many different ways in which systematically extracting analytic information from them can add to and strengthen qualitative research. For now, bear in mind that qualitative datasets also lend themselves to many of the suggestions below.

For either type of reactive or non-reactive data we suggest evaluating it across five dimensions (Figure 6.1).

Figure 6.1: Five key dimensions against which to assess reusing a dataset



Source: Authors.

The top criterion here is the *relevance* of the dataset for your particular research aims, which we considered above. Could it form either a central part of your work or underpin a discrete part of it? Regardless of its merits on other dimensions – such as being a prestigious or much used source in your subfield, or its impeccable reputation for using good methods – it must usefully address your specific research question to warrant further consideration.

Regardless of any claims to impartiality, all data is captured from a position and for a reason which it is crucial to understand. All data is imbued with *meanings* of which researchers must be aware in close detail (Barrowman 2018). What understanding of the social world was data collection based on, and how were these factors encoded in the final dataset?

In reactive research the precise form of words used in survey questions, the detailed setup of a lab experiment, or the sequence of questions asked in an interview may all strongly condition how useful the data can be. If it is data specifically created for the original academic study, then the meanings underpinning it are likely to be explicit – either in detailed expositions of the methodology or accompanying documentation justifying questions choices and study processes in terms of relevant literature. Some disciplinary norms and previous questions for which comparative data is available may also shape what is asked in less obvious ways. If survey or other reactive data is created instead by other actors (like official statistics bureaus, agencies or private corporations) there will be a practitioner rationale for how and why questions are asked – which may not be congruent or fit easily with academic interest in the data. So how have these background influences been translated into the framing of variables, what is covered, and what is not? How do these meanings affect the data's usefulness for your research purposes? For example, did the original researchers design questions in a survey in a slanted, biased or just inattentively worded way? Have critics argued that the survey question wordings had ambiguous meanings, or referred to other phenomena, or were

technically defective – and for new studies, could they? Was an experiment, focus group, interview programme or deliberative event explained to participants in ways that 'led' them, or 'framed' their responses or participation from the outset in particular ways? Do any failings here affect how far you can work with or approach this data?

In non-reactive contexts, how a market or state regulation process operates and what situation people experience as participants (with what interests and information) may deeply determine what a data measure means. In some ways non-reactive data provides 'objective' data, but its meaning may not be transparent. For instance, suppose we objectively record that millions of people using Twitter continue to do so after Elon Musk's takeover and changes to X. Does this mean that they welcome these changes to the site? Or that they do not care or have not noticed anything? Or perhaps they deplore the changes, but face large transaction costs in trying to relocate their social media presence elsewhere? Or again, they may deeply dislike the changes but believe that they can wait out Musk's perhaps doomed ownership period? Understanding the background to data's production is crucial to evaluating whether and how it can be reused for other purposes or combined with other data. As Bowker (2005, p. 184) remarks, 'raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care'.

Some *limitations* in reusing datasets inevitably arise in other issues or problems in how the original research was set up (the top left-hand element in Figure 6.1). How does this dataset rank in terms of its size and coverage? Is it more or less comprehensive than other datasets? The size of a dataset usually creates a trade-off between detail and coverage. Is the 'statistical power' of a given sample size enough to create meaningful answers to the questions asked by the original researchers, or in reusing the data – especially for surveys, experiments, quantitatively analysed interviews or cases, and so on? Analyses of objective data may have reverse problems if the dataset is so large that it is hard to analyse or generates 'significance' values that make every variable apparently 'relate' to every other one. Later on, when making a choice, focusing down on one or two candidate datasets can help to weigh up different limitations and decide those that are unavoidable, or can be mitigated or overcome, or would more seriously or fatally undermine the data's aptness for your research question.

The bottom two dimensions of evaluation in Figure 6.1 involve wider aspects. The *provenance* of the data concerns who produced it and how legitimately they went about doing so. Its *reputation* concerns how it is regarded more widely within your research community. In many cases, there may be a direct linkage or close overlap between these two considerations. For instance, nationally produced trade statistics may be carefully defined and have an impeccable provenance, such that their use is a fundamental cornerstone of the economics discipline.

However, there are also times where these two dimensions diverge. A data source may have a strong provenance in terms of the reliability of its data, for example geo-demographic data collected by advertising firms. But it may

come from an organization with a poor reputation in your discipline or that is seen by the media or public as funded by 'vested interests' in that area. Other researchers, and other readers with a different stake in an area, may object to the systems of classification used or see undesirable aspects in the ways that data was recorded. Similarly, it is important to consider ways in which research choices may affect wider social relations. If a researcher at a well-known university relies on a relatively unknown or obscure commercial dataset it may appear to be an endorsement of its overall soundness – even if this is not what was intended. Similarly, relying on a data source with a disputed or controversial reputation may inadvertently affect how research is seen, regardless of the design specifics or findings. For example, continuing to use Facebook data after the Cambridge Analytica scandal, or Twitter data after the Musk takeover of X, might easily be seen as problematic, even if the researcher has taken mitigating steps to avoid contamination from contextual changes.

Exploring a possible dataset

When you have found a dataset that fits your research interests well, scores strongly across all five evaluation dimensions and has been downloaded for further investigation, what comes next? Before finally committing, it is vital to first explore the dataset via some 'deep dives'; intuitive hunches and previous experience with other datasets can help here. Knowing exactly what it is you are looking for is key at this step, but it also hard to do when starting out. Some exploratory steps to work through can help get a 'feel' for a dataset, regardless of the data type or disciplinary focus.

Whatever software or language you have, try to get a *preliminary overall look* at your data. Make sure that you understand why, when and where a dataset was produced, with what initial research aims. It is important to always avoid treating datasets in a decontextualized way, for then major mistakes can creep in. Scroll through the columns and rows to see what variables look like, how they are labelled and what numbers or text they contain. To understand the size and character of your dataset, is it 'human readable' or so big that it can only be worked with in special software?

A quantitative dataset of a few thousand rows can easily be opened as a spreadsheet and worked through by a single researcher. Examining, correcting or classifying 3,000 rows of a variable by hand may be a tedious task, but it can realistically be done by a single researcher across a few afternoons. With 100,000 rows, however, even partitioning work amongst several people would still be unworkable. For very large datasets simply opening and examining the data in this way is not feasible – in Excel the 2020 XLXS document format had a hard limit of 1,048,576 rows (Microsoft 2023). (This catastrophically led to over 15,800 positive Covid 19 test results 'going missing' from official records in the UK during the height of the pandemic, after analysts failed to notice that the maximum limit had been surpassed and that 'surplus' records above the limit were simply being wiped (BBC 2020).)

For any large or 'big data' source, creating a smaller representative sample will yield a subset that you can look through and get a good feel for the whole. Either output a random number selection of rows or use a structured sample that works sequentially through the dataset – for example, outputting 10 sequential rows every 1,000 cases. If your dataset is not already in a simple tabular format, still try to find another way to browse through a broadly representative subset. For big qualitative data, a sample that can be examined and understood in depth is equally valuable, either through immersive understanding or using software packages that enable computer-assisted qualitative analysis (University of Surrey, no date).

On a first look at data, check for anything that stands out, may need further investigation or seems to be missing. The aim here is to surface issues that can then be followed up on more systematically as needed. Make notes as you go of what looks interesting or alternatively does not look quite right. Start by looking at the definition of variables or the classification of the data. Are the labels clear enough and do they make sense? Is there a codebook or extra documentation giving more information about how variables are defined – such as precise question wording in surveys, or exactly when and how 'objective' data was gathered or tracked, etc. Are all the variables and data rows you plan to reuse covered in the documentation, or are some not mentioned? Next, how are the relevant rows set up? Do they identify separate individual data tracks or respondents, so that each entry is a separate response? Or do clusters of rows have some form of 'family' or adjacency links? It is worth double-checking for duplicate or identical responses. Are there unique identifiers you can use for this purpose? Does the structure of your data appear more complex where identifiers are mostly unique but in some cases they appear in groups or areas? A preliminary look at a dataset will not leave researchers able to definitely answer these questions, but it can flag potential pitfalls with the dataset parts you plan to use, so that they are speedily resolved and do not undermine more committed exploratory analysis with the dataset.

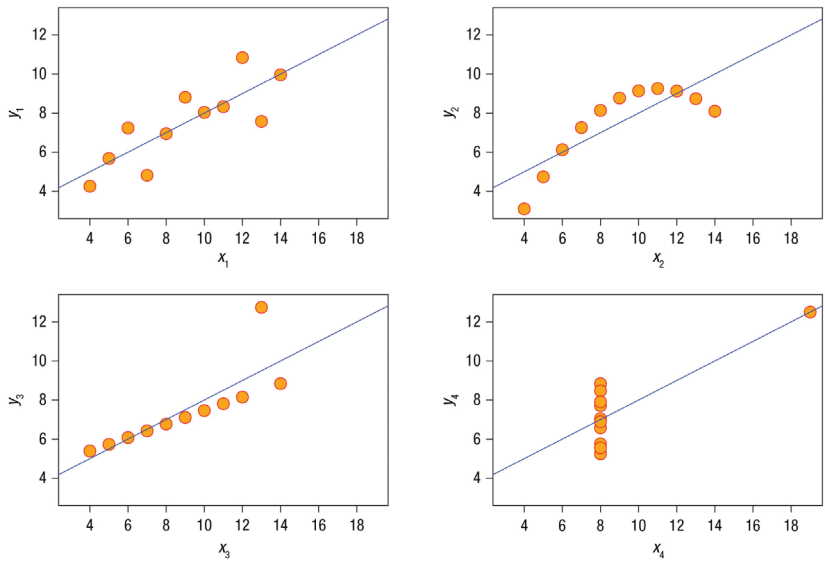
Creating simple quantitative counts, plots and tables is the next stage. How many missing values are there? Counting their incidence across variables of interest and across each row (or each unique identifier or group) will show if missing data will present issues when cleaning your data, merging it or starting your analysis. Ask too if coding has been consistent – for example, does it look like there was a lack of standardization across country or area names? A quick count of how many unique categories there are in each variable should let you know straight away. For instance, if there are more than 200 'country names' it will be apparent that this data either needs correcting or you are dealing with data that includes thorny issues of state disintegration/formation. A quick list of all the unique categories will also let you know if you are dealing with a few stray 'US' or 'U.S.' labels that need to be standardized into USA, or if you are dealing with a free text mishmash. On their own, glitches like this in relevant variables and identifiers do not mean that a dataset should be ruled out or discounted. But they can help you

evaluate the quality of the data and how much work will likely be needed to adapt the data for your own research.

Visualizing the core data you are interested in using (including demographic or contextual data) speedily and early on straddles the line between a first quick look and more sustained exploratory data analysis. A simple plot of a continuous variable will let you quickly assess what is in the data itself. Does it have the Gaussian distribution that you might expect, or are there floors or ceilings, or extreme outliers? Oddly high numbers may be the result of data input errors, or stem from using a separate category of error codes (like the still common practice of using 999 or other extreme numbers to indicate missing variables in survey datasets). On the other hand, very low or high values may be genuine data points that you will need to deal with as a methodological rather than a data cleaning issue.

Plotting or visualizing data quickly allows you to pick up on these issues in a way that simple descriptive statistics often cannot. For instance, Francis Anscombe formulated his famous 'quartet' display to demonstrate how four carefully constrained series can each generate the same statistical outputs, yet have completely distinct underlying distributions (see Figure 6.2).

Figure 6.2: Visualizing Anscombe's quartet – four different XY data relationships produce many of the same simple statistics



Note: These datasets have identical summary statistics (*X* and *Y* means, *X* and *Y* standard variances, correlations, regression lines and (rough) regression R^2). But the underlying data patterns are completely different.

Source: Wikipedia (2023) from Anscombe (1973). Reproduced with permission from Hehman and Xie (2021). Copyright © 2021 by The Authors. Reprinted by permission of Sage Publications.

If feasible, a helpful step before starting to go deeper is to run the replication code used by other researchers drawing on this dataset, seeing how they had to adapt a dataset to suit their research questions. Moreover, looking at how other researchers have used the same data may allow you to use their data versions if it is improved, or replicate early steps they took in cleaning and preparing the data, or find inspiration and ideas in how they have approached working with a given dataset. Finally, before you commit more effort to a source, it is important to have thought through any potential ethical intellectual property issues involved in reusing the dataset (see Chapter 7).

6.3 Data cleaning

Any researcher reusing someone else's data will likely have to spend a significant amount of their time on cleaning and preparing its information. As a rule of thumb, we suggest that you can expect to spend 80% of your time gathering, understanding, cleaning and preparing your data. It is better to overestimate the time these steps will take, because they are often tedious, time-consuming and unavoidable. Trying to rush through them or cut corners can easily lead to mistakes that can later fundamentally undermine the final analysis.

Sometimes it proves impossible to identify all problems in advance. But catching and addressing as many gremlins as possible prior to moving deep into your analysis can prevent the need for later revisions and reruns that absorb far more time. With potential issues identified – missing data, non-standardized or incorrect values, identifiers, mergers, additions, etc. – the challenge is to address them as efficiently and comprehensively as possible.

Tedious as data cleaning and preparation can be, automating tasks can make it far more bearable (or achievable at all in big data). You can build on your initial evaluation of the data to identify as many of these issues as possible and begin to think through how they could be done using a formula or automated sequence of steps.

In Excel 'automation' just means using fairly simple formulae to edit in easily trackable ways. If a complex recode is needed, make changes using successive 'IF' statements, creating intermediate variables that can be checked and corrected easily, and then cumulate to accomplish the whole change. Do not try to pack all the changes needed into a single formula whose meaning is hard to decode later on, and is always hard to correct if things go awry (they will). (As ever we caution against the use of Excel for analysis, see Chapter 5.)

In more advanced software packages, accomplishing complex recodes in one go may be easier so long as what each code line means is carefully and understandably documented. However, the process of implementing

automation itself can vary in its complexity, accuracy and time costs. For illustrative purposes we split this move into two categories:

- structured, strictly implemented automations – such as 'regular expressions' (or RegEX); and
- broader, heuristic automations – such as querying databases using application programming (API) interfaces.

Cleaning data by using Chat GPT, Claude, Grok or other LLMs or 'foundation models' is now feasible for expert users (Narayan et al. 2022; Zhang, Huang and Wu 2024). The conversational abilities of AIs mean that options for less expert users to clean data using ordinary language commands are still developing fast as we write. Google also offers free data cleaning tools in three different free packages that are easy to access but have not yet been independently evaluated. Sheets allows cleaning that is effective for small and well-formatted datasets and less expert users, working in a familiar spreadsheet mode that is easy to check. Gemini can work with Sheets and other packages, and is AI powered and able to work with ordinary language commands. Co-Lab is for expert coders and big datasets and is able to use powerful Python routines and link to computerized notebooks. For larger datasets Gemini and Co-Lab require subscriptions. Researchers should be wary of using any AI system to clean their data that they do not fully understand or control; the outputs obtained always need (quick but thorough) manual checks to ensure that the AI has worked as it was asked to do.

Regular expressions or RegEx

At a practical level lots of data cleaning is about tidying and standardizing information. At a technical level this often means working with *strings* – that is, data as sequences of numbers and characters. Often this can mean lots of small corrections, for example, removing trailing spaces after a variable or correcting capitalizations, so that when you come to your analysis the software you are using correctly recognizes the data. Many programming languages contain tools for addressing some of these problems, such as the command *trimws()* in R, which removes leading or trailing whitespace. However, being able to write your own code to tackle the wide variety of issues you encounter can be invaluable in easily and efficiently tackling complex data cleaning problems.

Regular expressions are a way of working programmatically with any form of string (Microsoft 2022). Almost all programming languages and some software can process regular expressions (although the particularity of their implementation may vary slightly) and they can be used to find, select and

replace anything that can be contained in a string. Moreover, because they can be written to included conditions and multiple statements they can be incredibly powerful tools that can accomplish far more than a simple 'find and replace'. As Friedl (2006) *Mastering Regular Expressions* evangelizes:

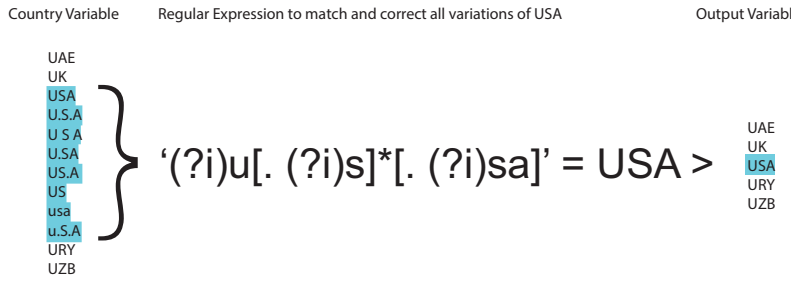
> At a low level, a regular expression describes a chunk of text. You might use it to verify a user's input, or perhaps to sift through large amounts of data. On a higher level, regular expressions allow you to master your data. Control it. Put it to work for you. To master regular expressions is to master your data.

By using regular expressions it becomes possible to construct complex commands for searching, matching and cleaning your data. For instance, rather than having to correct every variation of how a country name can be inputted, a regular expression could capture all variations and abbreviations (both correct and incorrect) that match a general pattern (see Figure 6.3).

Cleaning with APIs

Using APIs to clean and standardize your data can be a powerful tool, although they add an additional layer of complexity and data processing concerns to your workflow. (This technical use is distinct from downloading data from platform company APIs themselves to use in research; on which see Tromble (2021).) An API is a way for two (or more) pieces of software to talk to each

Figure 6.3: An example of a regular expression to match and correct all versions of 'USA'



Note: An example of a regular expression in use: '(?i)u' specifies an upper or lower case 'U/u', this must then be followed by a '.' whitespace or 'S/s', the * covers a repeat of any of these characters and the final part of the expression repeats this with the addition of an 'a'. All together this regular expression allows for many variations of how the USA could be inputted by users and allows them all to be quickly standardized without requiring the researcher manually identify each specific variation.
Source: Authors.

other. For data cleaning, it can be a good way forward to send data to be processed by a larger 'platform' that has resources or solutions to fix problems that cannot feasibly or easily be solved by individual researchers. For example, an API to a mapping platform could be used to turn messy non-standardized address data into longitude and latitude coordinates. Using APIs lets you draw on the greater resources of platforms. But it also means sending parts of your data to third-party sources, raising two main issues that need to be handled with care.

First, would sending any part of your data to an outside source compromise the anonymity or confidentiality of individuals in your dataset? Even if you are certain that you can maintain confidentiality, there may also be limitations from data protection legislation governing what data can be shared with a third-party platform. For some APIs you may be able to run all the code and services you need locally and so skirt these issues. For example, the open source mapping service OpenStreetMap (2023) can be downloaded in full to your own computer, and API requests can then be sent to a version running there – although this does require significant technical and computing capacity to set up. However, most commercial organizations force you to engage with their online platform as a core part of their business model.

This raises a second key issue with APIs in research, namely that they essentially work as a black box. You provide an input and get back an output. Even with open source platforms it is hard for researchers to audit or verify how the output received was arrived at, and with commercial APIs this is impossible. Finding ways to try and double check the results from an API can therefore be crucial in ensuring the veracity of this kind of data cleaning. Some of the same techniques used to first explore a data source can be helpful here – for example, manually inspecting samples and plotting distributions. It may also be possible to (partly) verify the results from other variables in your dataset. For instance, if you use an API to geo-code addresses, but the dataset also has variables for state/province, electoral district or local authority area, then you can partly check the accuracy of even large numbers of longitude and latitude points returned by an API.

6.4 Merging datasets

Bringing together two or more datasets can be an effective and innovative way to reuse and remash other people's data. Merging datasets, enlarging variable coverage for the same units of analysis, or adding existing variable data for new rows/cases, can all make novel contributions that extend a research field in new directions, and attract the gratitude of other researchers with attendant citations. Of course, what you want to add in will be motivated by the research question you are seeking to answer. And, while there are often salient theoretical reasons for bringing two (or more) datasets together, how is this going to be *practically implemented*? Merging two datasets is quite distinct from

using two datasets placed in comparison or conversation with each other in your analysis, and it requires a literal similarity in the data points that are to be matched.

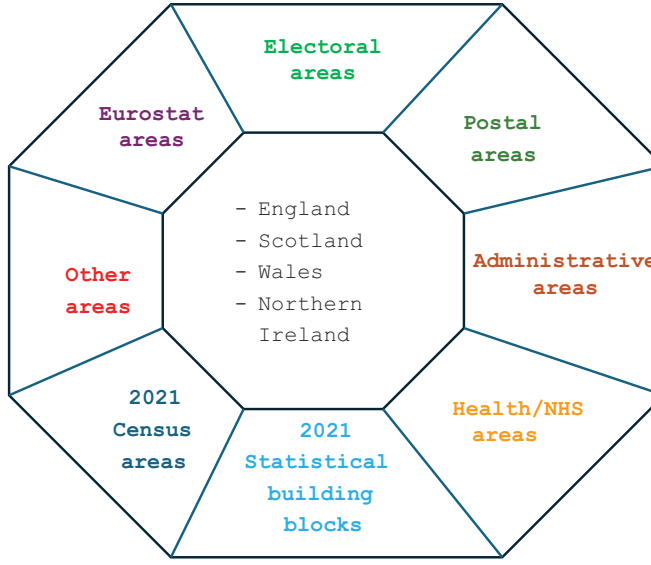
Merging two datasets requires there to be a common data point in both on which they can be matched, for example, codings for a city or organization name or another identifier present in both sources. Yet there does not need to be a symmetry between the two datasets themselves. One dataset could be at the level of individual respondents with the variable for 'place of residence' and the other be national statistics on metropolitan air quality across areas. Here a key part of the data cleaning process could be standardizing these two variables so that they are coded at the same geographical level with the same identifiers. Bringing the two datasets together also requires decisions to be made about priorities. If two datasets do not have an identical overlap, one must be prioritized over the other with non-matching points from one being discarded. Pushing forward with keeping non-matching data points can quickly lead to tricky issues where multiple null datapoints significantly duplicate data.

In bringing two datasets together there can be a temptation to merge in as much information as possible from a wide range of sources, with the justification that bringing to bear the most data feasible is likely to best advance research. However, the more unique data sources you bring together, the more complex and unwieldy it can become to work with the final dataset. In particular, when merging multiple datasets attention needs to be paid to the meanings that underlie why the data has been collected and how this has been encoded into each variable. Matching two datasets on the country code GBR is a trivial task, but if this code does not refer to exactly the same geographical unit it could undermine your analysis. For example, as an international standard the ISO-3166 country code GBR covers the United Kingdom and Northern Ireland. But in the World Values Survey the GBR code excludes Northern Ireland – here a separate survey was conducted and coded as NIR (a non-standard country code). This is not an isolated problem, even more so when considering official or government statistics. Figure 6.4 shows the eight different types of geographical areas used in UK official statistics, each containing many sub-areas, and also set up differently for different parts of the country. Different areas are used for eight different types of statistics, and converting area data for one purpose into area data for a different purpose is a complex operation, a condition that applies in most countries (Machin 2024, IDP blog 2024).

The specificity of definitions and meanings that are baked into the data you are working with can present both challenges and opportunities for bringing together data sources in novel combinations. That two datasets do not exactly align is not necessarily an impediment to your research because you can reshape the data to make merges possible.

Upscaling is one way to create a point of convergence by grouping or aggregating the information in one dataset so as to match the scale of another. It is

Figure 6.4: The eight main systems of areas used in UK statistical geographies by the Office of National Statistics (ONS)



Source: This is a very simplified version of the full image, for which see Office of National Statistics (2021) and theidpblog (2014).

particularly useful when working with temporal or spatial variables. Reactive survey data, which has used local administrative units, and non-reactive consumer data based around supply chains, may initially appear to have a non-compatible geographical variable. However, if both datasets are upscaled from their more granular original form to aggregates at the scale of large urban areas they might then be mergeable together. Of course, this aggregation changes the nature of the data, and cuts the ability to specifically examine each area, but it may be worthwhile if it gets you closer to answering the key research question.

In most cases, however, merging two datasets is best done on matching variables. There may be cases in which non-exact matches are useful. Approximate string matching or 'fuzzy matching' will match variables that are within a defined threshold of closeness to each other. Matching in this way can account for deviations or errors in the data you are working with that cannot be fully accounted for in the data cleaning process. For example, you may want to match individuals' names across historical records. Given the nature of this data, it is reasonable to assume that there may be mistakes in both the spelling of some names and errors in their digital transcription. A fuzzy match which can allow for one or two missing, additional or wrongly placed characters in each name would be able to account for some of these errors and make matches between names that are 'close enough'. What counts

as close is a matter of interpretation and fuzzy matching in this way usually requires experimenting with different thresholds to get the desired effect of catching errors without too many false positives. Running multiple iterations of merges, where the data allows, may also be a way to 'fuzzily' match data and increase positive matches between datasets. For example, two datasets may have an overall address split across multiple variables that are not always precisely filled out, such as 'building number', 'street name', 'city', etc. Matching across all variables and all possible combinations and their order and then dropping any duplicates may significantly increase the match rate between two datasets, without requiring the address data to be fully corrected and standardized by the researcher.

Matches can also be found between datasets by *downscaling*, or trying to disaggregate data into smaller units. While it is difficult to do (and hence less common), downscaling may present novel research opportunities so long as some conditions are met. It must be appropriate and ethically OK to make this move. Attempting to break down a dataset into smaller units may compromise the confidentiality and anonymity of the original participants. Even if you do not intend to share your results publicly, proceeding on this course here would be an ethical violation and should not be attempted.

However, if no such concerns are raised, there are a number of avenues for trying to downscale your data, and the information to do so may already exist in the data you are working with. For example, the only temporal variable in a dataset might be 'YEAR' but it may be possible to deduce a more precise date from information contained in other variables. Similarly, data could be merged-in from other sources that would allow you to make such deductions. Dependent on the type of data you are working with and the methodologies you are using, it may also be possible to estimate, infill, or estimate additional datapoints that allow you to downscale your data. Many sophisticated, semi-supervised and unsupervised machine learning (ML) techniques are capable of complex estimation or categorization – although using them always needs some verification and methodological justification.

Conclusion

Making use of other people's data to ground your own research is a central tenet of OSS. It offers an attractive prospect of making rapid progress to many researchers who cannot otherwise afford to generate specifically tailored datasets from scratch. And reusing data helps with the cumulation of information and integrative, multidisciplinary scholarship that has tended to be underemphasized in the social sciences. However, reusing other people's datasets can have problems of its own that are often overlooked. Just because data has been openly published there is no guarantee that it can easily be found, made useable, or effectively adapted to address other research questions. And as the next chapter discusses in detail, there may be ethical and intellectual property rights issues in reusing existing data. Nonetheless, we have shown here that

if sources are reused sensibly many strategies can help mitigate or solve the most common problems, and make reusing or mashing other people's data a beneficial and important OSS practice.

Note

* This chapter draws on research and seminars forming part of the CIVICA Research programme, which was funded in 2022–24 by the European Union's research and innovation programme Horizon 2020, under the chapter 'Science in Society'.

References

Anscombe, Francis. J. (1973) 'Graphs in statistical analysis', *American Statistician*, vol. 27, no. 1, pp. 17–21. OA at: https://perma.cc/2A3N-BE2D

Barrowman, Nick (2018) 'Why data is never raw – on the seductive myth of information free of human judgment', *The New Atlantis*, Summer/Fall. https://perma.cc/2VUP-933D

BBC (2020) 'Excel: why using Microsoft's tool caused Covid-19 results to be lost', *BBC News* online, 5 October. https://perma.cc/8R2B-664D

Bowker, Geoffrey (2005) *Memory Practices in the Sciences*, Cambridge, MA: MIT Press. **$** https://perma.cc/BSH9-73AV

Brower, Rebecca L.; Jones, Tamara B.; Osborne-Lampkin, La'Tara; Hu, Shouping; and Park-Gaghan, Toby J. (2019) 'Big qual: defining and debating qualitative inquiry for large data sets', *International Journal of Qualitative Methods*, vol. 18: 1–10 https://doi.org/10.1177/1609406919880692

Chandrasekar, Abinaya; Clark, Sigrún Eyrúnardóttir; Martin, Sam; Vanderslott, Samantha; Flores, Elaine C. et al. (2024) 'Making the most of big qualitative datasets: a living systematic review of analysis methods', *Frontiers in Big Data*, 25 September, vol. 7. https://doi.org/10.3389/fdata.2024.1455399

Friedl, Jeffrey (2006) *Mastering Regular Expressions* (3rd edn), San Francisco, CA: O'Reilly Media. **$** https://perma.cc/L25Y-5KCC

Hehman, Eric and Xie, Sally Y. (2021) 'Doing Better Data Visualization'. *Advances in Methods and Practices in Psychological Science*, vol. 4, no. 4. https://doi.org/10.1177/25152459211045334

IDP blog (2014), ONS Wallchart, 20 July. https://perma.cc/9KBC-7RJV

Jaźwińska, Klaudia and Chandrasekar, Aisvarya (2025) 'AI search has a citation problem – we compared eight AI search engines. They're all bad at citing news', *Columbia Journalism Review*, 6 March. https://perma.cc/NEF3-GQ4N

Machin, Sam. (2014) 'ONS-API, Understanding Geography', 20 August. ONS blog. https://perma.cc/9KBC-7RJV

Microsoft (2022) 'Regular expression language – quick reference', 18 June. https://perma.cc/N626-FYD4

Microsoft (2023) 'Excel specifications and limits', Web document. https://support.microsoft.com/en-au/office/excel-specifications-and-limits-1672b34d-7043-467e-8e27-269d656771c3

Narayan, Avanika; Chami, Ines; Orr, Laurel; Arora, Simran; and Ré, Christopher (2022) 'Can foundation models wrangle your data?', *arXiv*. https://perma.cc/7PTA-K3PD

Office of National Statistics (ONS) (2021) 'Hierarchical representation of UK statistical geographies', ONS website. https://geoportal.statistics.gov.uk/documents/338519c3569944f79aa34ea52a0a49c3/explore

OpenStreetMap (2023) Home page. https://www.openstreetmap.org/#map=5/54.910/-3.432

Tromble, Rebekah (2021) 'Where have all the data gone? A critical reflection on academic digital research in the post-API age', *Social Media + Society*, 19 January, vol. 71, no. 1. https://doi.org/10.1177/2056305121988929

University of Surrey (no date) 'Choosing an appropriate CAQDAS package', Computer Assisted Qualitative Data Analysis (CAQDAS) networking project. https://perma.cc/ESL8-293S

Wikipedia (2023) 'Anscombe's quartet'. https://perma.cc/3BTK-LCWM

Zhang, Shuo; Huang, Zezhou; and Wu, Eugene (2024) 'Data cleaning using large language model', *arXiv*. https://perma.cc/2XU4-QE7U

7. Writing a data management plan

Hannah Boroudjou

A data management plan (or DMP) is a structured document that asks researchers to broadly outline what data they will use in their research and how they plan to manage, store, secure and share it in the long term, in ways that meet their legal obligations for data protection and comply with other university, lab or funder policies on data handling. Coinciding with the introduction of data protection legislation, such as GDPR (General Data Protection Regulation) in the European Union (EU) or POPIA (the Protection of Personal Information Act) in South Africa, funding councils, labs and universities are increasingly asking researchers to submit a DMP in the early stages of writing new research proposals.*

However, these plans should not be seen as just a box-ticking measure for researchers to explain how they will meet the minimum requirements of international law. Instead, they aim to also be 'useful tools to investigate researcher's current plans and philosophies for how research outputs might be shared' (Bishoff and Johnston 2015, p. 1) – a key aspect for open social science (OSS). Increasingly, funding and academic reviewers regard DMPs as a key measure of researchers' commitment to open science (OS) and OSS in terms of depositing and sharing data outputs.

We begin by considering the high-level FAIR (findable, accessible, interoperable, reusable) principles (Section 7.1). We next consider what is needed in a DMP that fits with FAIR principles, along with the broad requirements that funding bodies and repositories generally ask researchers to explain (Section 7.2). The issues around identifying and bringing together all your data form the focus of Section 7.3. It is easy for oversights to occur at this stage. Lastly, we consider issues involved in depositing the final dataset (Section 7.4).

7.1 The FAIR data principles

Most DMP templates used by universities and funding bodies will prompt researchers with some variation on this phrasing: 'Outline how you plan to

How to cite this book chapter:

Boroudjou, Hannah (2026) 'Writing a data management plan', in Dunleavy, Patrick and Monteath, Timothy (2026) *Doing Open Social Science: A Guide for Researchers*, London: LSE Press, pp. 149–176. https://doi.org/10.31389/lsepress.dos.g

share the data outputs of your research. What aspects of the dataset will you share? How will it be made available and where?' Some plan questionnaires go a step further to ask how you will make any data shared consistent with the FAIR guiding principles for scientific data management and stewardship – that data should be findable, accessible, interoperable and reusable. This approach has been well advanced in science, technology, engineering, mathematics and medicine (STEMM) and some quantitative social sciences. Making data FAIR asks you to really consider how the discoverability of shared data can be maximized by making it readable by machines:

> Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. (Wilkinson et al. 2016).

Regardless of whether a particular DMP template mentions the FAIR principles directly, they constitute a useful framework for measuring the effectiveness of your data deposit plans. Being aware of FAIR also signals that a research team is engaged with OSS in ways that most reviewers and funding bodies rate as important. However, FAIR does not automatically equal open. It may be possible to share data in a repository under a restricted licence because legal, ethical or commercial concerns prevent sharing the data in a fully open format. Here the data is FAIR but not open, although it will be understood by funders and other stakeholders to have been made as open as can reasonably be expected.

The website go-fair.org has a comprehensive breakdown of each of the FAIR principles, how they work and what they imply (go-fair no date). However, existing guidance tends to assume that the data to be deposited is quantitative or numerical. It can be difficult for researchers using qualitative or mixed methods of analysis to apply FAIR principles to their own work. The following sections stem from the go-fair guidelines, but break them down in a way that is more applicable to the full diversity of research methods in the social sciences.

The *findable* requirement is that data must be easily discoverable, both by humans and computers. Any data that underpins an academic article or book must be linked to from that publication using a DOI (digital object identifier) – that is, a unique string of figures assigned to a publication or dataset. Ideally, each dataset should have its own DOI which makes it discoverable as a unique object. The dataset should ideally *not* share a DOI with any research publications derived from it. The best way to get a unique DOI assigned to each dataset is to deposit it with a recognized repository that automatically assigns a DOI and generates a citation to it. Once the data has been deposited you can link to the dataset from the article in a data access statement and in the reference section, where it may be more cited (STM, DataCite and Cross Ref, no date). DOIs work well if they are known by searchers, and they are widely supported in academia and publishing.

However, in addition, datasets should have rich, accompanying *metadata* that describe the nature of the study, any unique characteristics of the data, the broad demographics of the participants from whom data was collected (where appropriate), the methods of data collection used and any keywords that describe the subject of the study. Any reputable data repository should be easily searchable in multiple ways. A well-crafted data deposit will contain enough descriptive information and keyword prompts to ensure that it crops up in multiple relevant search results.

Once potential users locate a possibly relevant research dataset, it needs to be *accessible* so that they know how they can open and use it. Ideally, the data should be retrievable by anyone across the globe with an internet connection, simply by clicking on a link and without any need for payment, nor justification for why they want access, or any other manual intervention. Many different practices (still common in both STEMM and social science disciplines) are not in fact consistent with FAIR good practice – such as, requiring potential users to supply a name and email address to the data authors for permission to access the data. However, in social research there may be cases where commercial, ethical or legal restrictions mean that access to the data must be regulated or mediated. In these cases a good data repository will act as an intermediary contacting the author on behalf of the requestor. If granting permissions is unavoidable, then it is better to consider forming Data Access Committees with multiple authors or other institutional stakeholders involved, rather than using a single author as point of contact (who could also become a single point of failure for future data access requests).

A good repository will have these mediated or access-control processes already built into the framework. For example, the UK Data Service allows authors to deposit data on closed access, but with an option built into records where readers can request access, usually by a short application outlining the reasons they need to see or reuse sensitive data. Even with a mediated process in principle the data is still accessible, because a clear access route has been outlined and the authors are contactable. For the same reason it is important that all authors involved in data creation are named on data deposits. Data repositories hope that their stored data are sustainable and can be made accessible for many years to come. However, in the event that file formats change and links become unusable, the metadata record should still exist and all the data owners can still be tracked from that.

For data to be *interoperable*, it needs to be understood by most humans, but also by computers or machines, using commonly accepted or widely used formats for storing data. If possible, try to avoid employing file and data formats that can only be accessed via specialist software. Even quantitative analysis software that is commonly used by researchers (like STATA) will not be accessible to non-academics, such as members of the public, business executives or government officials. Many proprietary systems (like STATA or SPSS) have high access costs, far out of reach for individual researchers and students outside well-funded institutions, and access to relevant software is even more

patchy in less affluent institutions worldwide. Other free packages – such as R and Python – can have substantial skill barriers that prevent untrained people from accessing them, even though code that is well commented on and documented can mitigate this threshold barrier somewhat. Even a widely available lower cost software package (like Excel) creates a barrier if the data is only made available in a proprietary form.

One of the key purposes of OSS is to extend access to research data well beyond the rich countries' academies alone. This does not mean that you cannot use STATA or Excel in data analysis – although we caution against over-reliance on them (see Chapter 5). But when you are depositing numerical tabular data it should normally be feasible to convert it to CSV (comma-separated values) format, which is both readable by machines and more accessible to a wider number of people. (CSV files look like other spreadsheet files, but they have a .csv extension and can be read by Excel and Google Sheets, and many other free alternatives, as well as by academic software analysis packages.) Interoperability also means using commonly controlled vocabularies and thesauri when describing data. For example, authors can use international standards such as the Data Documentation Initiative (2025), a free international standard for describing data produced by surveys and other observational methods.

What if the data being deposited consists of documents and scripts, such as transcribed interviews, field notes and long-form or 'thick' descriptions and observations? Here use rich text format (.rtf) files, which are interoperable between different processing systems. Alternatively, commonly used software like Word and text-based PDFs (which are editable and searchable) are also considered acceptable. But other image-based forms of PDF are unchangeable and often uncopiable, and to be avoided. Any unsearchable form of PDF is not interoperable, because it loads in a 'slab' format that is unusable for other researchers. (If you have an image-based PDF – e.g. because it was scanned in – it should be possible to convert it to a text-based PDF. Most universities should have specialist software available to do this.) A comprehensive breakdown of the different types of data with recommended formats has been produced by the UK Data Service (No date b).

For qualitative social scientists and perhaps especially scholars working in overlap areas with humanities disciplines, it is also important to take a broad view of what kinds of 'data' will be generated, and to apply the FAIR principles to all of them: 'Humanities researchers create data too. These may include lists, tables, matrices, databases, maps, digital objects, images, and recordings and transcripts' (British Library 2021). To this list one could also add annotated bibliographies, or poems, plays and other creative text outputs. Most data depositing now focuses on digital data of all kinds, but managing research data also covers non-digital forms of information. In addition to completed tables, a lot of deposited data in the humanities and closely related social sciences consists of digitized images such as photographs, pictures, cartoons, maps, diagrams, text facsimiles, and so on. There are some

particular pitfalls here, such as weak metadata for individual images, still too often stored with only 'bloc' referencing of multiple images (Hansson and Dahlgren 2021).

If you have any linked datasets within the metadata of the data you are depositing – that is, if any of your data has been built on another pre-existing dataset, then users may need to access it also in order to gain a more complete understanding of your final dataset. In this case make sure to mention this in the introductory materials for your own deposit. Equally, point your users early on to any complementary or supporting information available in another dataset. Any such associated datasets should be linked to from within your metadata record using a unique DOI.

The final FAIR principle is that data should be *reusable* – meaning that it is so well described and documented that it can easily be understood, applied and put to good use by other people (scholars or citizens) in other settings. While the first FAIR element ensures that all data deposits are findable (thanks to rich metadata that gives them context and meaning), making data reusable means that readers can learn enough from the metadata to actively employ your dataset themselves – either to replicate your results, or to build a basically similar study. In practice, this means describing thoroughly:

- the origins of the data;
- all the software used to collect and analyse it;
- the scope, purpose and timing of data collection; and
- any limitations in the data or complications that may arise in using it.

The FAIR principles can appear to primarily apply to quantitative researchers, and at this point some qualitative researchers may begin to feel uncomfortable about using them. As Chapter 2 discusses, to demand that that all research must be replicable for it to be legitimate may raise fears that it excludes a great deal of qualitative and ethnographic research. Like many other types of field studies, they may not be capable of being exactly repeated (because 'time and tide' wait for no one and may introduce important variations). In addition, as Chapter 4 showed, social scientists have not reached a single consensus on what makes knowledge valid.

However, in qualitative social science reusability can simply mean providing enough information about the study that readers could recreate a *similar* study in their own environments, or use the data to answer a different research question, or even use the data for a completely different purpose. Some observers argue that for qualitative researchers it is more helpful to focus on the concept of transparency rather than reusability – meaning that there should be enough information to 'help put the reader back in the fieldwork scenario' (Corti 2018). Part III of this book considers such arguments in specific relation to each of the major qualitative methods and sources.

A final and essential element of reusability (or transparency) is that readers know accurately how the data can be reused and how the research and authors

generating the data should be acknowledged. There are two simple steps to doing this:

- in the introductory materials, include a section on 'How to cite this dataset' that sets out exactly how your data should be cited, in text that users can just copy into their own publications and reference lists; and
- when you create the metadata record for deposited data, make sure that you prominently apply an appropriate reuse licence to the data (such as one or another form of Creative Commons licence, discussed below (in Chapter 15, Figure 15.2)).

7.2 What should your DMP include?

Some universities have also made producing a DMP mandatory for doctoral researchers to transition to PhD status beyond the first stage of their studies. Increasing numbers of academic departments also require internally funded staff projects to be accompanied by a DMP. And the reviewers for funding bodies will want to gauge researchers' level of commitment to OSS from the earliest stages of a grant bid or application for resources, and to prompt researchers who have not previously thought through their stance to do so. So what should be included?

DMP templates (with notes and tips on what to write) are available to download and fill in for most major social science funders. (See Annex 7A for some sources.) Different templates place emphasis in different areas; for example, across data and security, data organization and description, open data arrangements, etc. So it is worth paying attention to each funder's key concerns and priorities. If a grant bid is turned down by one body and goes to another funder in a revised form, altering the DMP sections to match new requirements may make sense.

Most online and template advice is written with STEMM and quantitative subjects in mind; for example, assuming that researchers will be sharing code and depositing large quantitative datasets (see Annex 7A). Advice on how to deal with more nuanced or complex data from social research is harder to find – for example, how to prepare in-depth qualitative interviews for sharing, or how to calibrate instruments in the social sciences to enhance data quality. Yet in fact DMPs are not just used in large, funded and quantitative projects, although they are particularly essential there. All the Part III chapters underline that it is just as important to prepare qualitative research data so that it can be transparent to and shared by other researchers in the long-term.

Scope and overview

Plans take time to design and write up, but they can also have numerous benefits for researchers. First, a DMP prompts you to think critically about the

ins and outs of how you will collect, organize and share your data in the long term. Funders, institutions and reviewers often use an assessment of how well this is done to gauge how far researchers have thought through key practical aspects of their project.

An effective DMP also saves time later on by reminding you to describe and document data efficiently as you go along (see Chapter 5). In the past many researchers have used a kind of 'sandwich' approach to doing data-sharing work. Attention to data-sharing issues starts off high initially, as applications need to satisfy funders and reviewers. Yet then the topic can drop out of sight, only to emerge as a substantial issue right at the end when the data gathered must be analysed and put into robust forms that can be shared. An effective DMP asks you to document and describe data well as you go along, and make periodic checks on data quality, especially when more radical and unanticipated changes in data variables or analysis methods occur mid-course. Saving time needed at the end of projects is especially valuable, because this period is often super-rushed, with inflexible reporting or funder deadlines to be met, often with considerable changes in roles or departures amongst project staff, and the main researchers' attention moving on to new tasks. Prepping as much of the data as a possible ahead of 'closing time' can mitigate any stressful end-rush to complete multiple 'signing-off' obligations. In large projects in particular it helps ensure that the project's lead researchers make key decisions and complete the work involved while it is still uppermost or fresh in their minds.

Smaller individual projects can use a DMP as an organizational tool to help one or several researchers organize and plan how to conduct research at a more granular level. The main difference with unfunded research is that less elaboration is needed and submitting a plan may not yet be mandatory for smaller, individual, informally defined, or non-funded research, including PhD theses. However, rather than breathing a sigh of relief at being spared the DMP task, researchers should recognize that unanticipated issues around how to manage data (discussed with colleagues) have often proved a particular vulnerability for such projects. Initially, data issues may be addressed sketchily, but later prove more serious once the project is underway. And quite often numerous data 'time bomb' delays occur at the end of research as initially poorly implemented data gathering and preparation has to be rechecked or even redone. So, however small or apparently self-contained your project is, drafting an appropriate DMP (e.g. not over-lengthy for simple blocs of work), and then keeping it up to date, will always be worth the time and effort involved.

Because of the diversity of funder and institutional templates available, and the differing needs of PhD or small project DMPs, it is hard to provide an exhaustive list of questions and topics that any one must cover. A research study investigated a sample of data management plan templates from 51 different U.S. funders, and identified 43 different topics that were covered across

the plans (Williams et al. 2017, p. 133). In practice, a 'standard' data management plan for the social sciences should cover these topics:

- What existing datasets will you work with, and will they be used as they stand, or modified and adapted, or combined with other data?
- What new data will your project collect, and why is a new survey or a new compiling of information needed (or value for money) when compared with existing sources?
- How will you handle personal and special category data and the ethical issues that may arise?
- What arrangements do you have set up for storing the data securely – for example, on university or project-specific file stores (see below)?
- What plans do you have in place for describing and documenting the data? For instance, will you conform to FAIR principles in depositing the data?
- How far will your dataset be made open access, and with what access conditions or restrictions, if any?
- How are roles and responsibilities for dealing with data quality, documentation and depositing allocated within the research team, and at what times are they tackled during the project? If a project only has one or two researchers, it is important to assign feasible times needed for these roles.
- Where do you currently plan that the final dataset will be made open? For example, do you have a particular repository in mind already, or several possibilities?
- What copyright licence do you envisage for your dataset, and how will ownership be defined?

For anyone working on these issues for the first time, they can seem a daunting prospect. However, the following sections describe how an effective OSS plan provides answers about multiple issues.

Pre-existing data

Many DMP templates start by asking what already-established data sources you are already using or plan to use. An extensive literature review is not needed here, but just a broad outline of any specific datasets that can contribute to answering your research question, and any limitations they have. Funding agencies are especially averse to recreating very similar data to any that has already been paid for and produced elsewhere. So what secondary data will you need to support your research? And are there any terms and conditions that may limit your access?

If you are requesting data from third-party suppliers are you certain that they can in fact share that data with you freely? Or will they need a data-sharing agreement in place with you first? Also, are the datasets you will use

publicly available online? Or are there are special access conditions for their use, such as data-sharing agreements, access restrictions or licence agreements, etc.? Are you confident you can surmount such hurdles? It is a good idea to have explored these issues thoroughly as early as possible in a project's development. Some DMP templates have a question that asks you to clarify the copyright and intellectual property rights that apply to any secondary data you will use. This question often stumps researchers, but needs to be answered in two parts. In the first instance you must establish who the original or 'natural' data owner is. Often this is the original author or data collectors, but data ownership may also be vested otherwise (e.g. if a firm gave access to researchers).

Increasingly, some social scientists ask their research participants to produce original work to support their research, for example, photographs, video essays, diaries, and so on. Participants have a natural copyright over materials they have produced, so it is important to make sure that you have agreed with them in advance on joint ownership of such materials. Make sure to consult your university's or lab's intellectual property (IP) policy to check. Later you will need to specify the copyright and licence arrangements that will apply to any dataset that you will deposit (see Section 7.4). These decisions can sometimes get a little troublesome – for example, where you have merged other people's prior data that has some restrictions with your own data that is open access.

Briefly demonstrating an expert and wider knowledge of the practical research landscape that informs your work provides funders with assurance that any new data collection to be undertaken is a unique and necessary contribution to your knowledge area. In funding applications this information is covered in exact detail in the main body of the proposal, so the DMP section here can draw on that, and just provide a more practice-orientated summary. If you're submitting a plan to your university for a small, non-funded or PhD project, the reviewers in each of these cases will greatly appreciate you outlining the secondary data to be accessed (and how you will do this) plus the broad demographics of any research participants in any data you will modify and analyse yourself.

New data

Survey and field research are expensive to conduct, so project bids normally must outline and justify the minimum new data to be collected, and how you plan to collect it in terms of methods and methodology. Providing an overview of the existing data as above normally can be a key springboard for claiming that all your original data-gathering addresses important gaps in current knowledge. For brand new data include the obvious information that supports the substance of your project, for example, interview recordings and transcripts, survey results, observation notes, etc. But a DMP can also include less obvious forms of data, some of which will be offshoots of the main

datasets you collect, such as scripts or software code, workshop recordings, minutes or lists of participants, interview guides or questions, anonymization logs, read-me files, etc.

Ethical issues

The DMP is also a key tool in helping researchers meet their ethical obligations to the people or organizations that participate in or form the subject of research. In describing any new data collection, try to demonstrate a knowledge of the different type of data you will need to collect and their respective sensitivity levels. Some research topics produce more sensitive data than others, and any social scientist who has ever written and submitted a research ethics review should be familiar with concepts like vulnerable groups, high-risk research, the importance of informed consent, etc.

For instance, in any project that collects ethnographic data you will likely have to handle sensitive information about specific individuals, which must be done sensitively and responsibly. The Association of Social Anthropologists (ASA) guidelines provide a useful summary:

> Ethical, personal and legal dilemmas occur at all stages of research: in the selection of topic, area or population, choice of sponsor and source of funding, in negotiating access, making research agreements and while conducting fieldwork. These ethical dilemmas extend to the interpretation and analysis of results, the publication of findings and representation, and the storage and/or disposal of data (ASA 2021).

A data management plan should in particular outline how you will anonymize participants or explain why anonymization may not be sought, where permissible. Given the sensitivity of the data being collected, how can you effectively utilize responsible data storage and sharing methods to ensure that personal data is stored safely? What ethical, commercial, or legal restrictions on sharing apply to your research, and how will they be managed? Think carefully about the different types of data to be collected and the different sensitivity levels that may require handling, storing and depositing information in different ways (see p. 163).

Costing and resourcing data preparation

These are often overlooked by researchers but should be covered in DMPs. If seeking funding, factor in the costs of hiring a data manager or a research assistant with specific responsibilities for prepping data for long-term sharing. This helps give peace of mind to the principal investigator(s), knowing that costs for prepping data for sharing are covered, and mitigating any risk of last-minute additions to the essential budget, upsetting the overall project financing.

Large collaborative research teams will need to consider the costs of bulk-processing data and cleaning and correcting it (see below). Bear in mind too that additional costs may be involved in preparing data for sharing in high-quality formats – such as, hiring a separate person (or even several people) to act as data managers on the project. Often this role is assigned to a PhD or post-doctoral researcher (and here the costs involved should be based on the usual hiring costs for such roles at your university or lab). However, it is vital that all team members recognize that the day-to-day administration involved in effective data management happens on too granular a level for just one individual to take responsibility for it on behalf of the whole project. If a large-project DMP relies on one individual there is a risk of reviewers seeing this as a potential point of failure – for instance, if the key person moves on to other positions or their contract ends. It is important to make sure that the wider project team contributes to and is familiar and happy with the DMP.

For individual or small research projects consider the possible costs of any software needed to analyse or collect data. Some universities and research institutions will have subscriptions to analytical software like STATA or data analysis tools like survey and transcription tools, as will corporations for their in-house researchers. However, this is not guaranteed. So check what software is available within the team's home institution(s) and what needs to be bought, before filling out the DMP. You may also need to consider some labour assistance costs. On individual research projects these should not be large, but there may be some additional costs involved in hiring a one-off service provider such as a translation agency. These points only cover a few of the possible costs involved in data management. The UK Data Service (No date b) has a specialist costing tool that can be used to assess the needs of any project, and the likely costs arising for data management.

Ownership of the dataset

This and copyright issues (see below) should normally be covered by the DMP, which provides an opportunity to ensure that every researcher involved in a collaborative project (and linked PhD slots) has a shared understanding of where ownership of the data lies, and of how the data will be made discoverable and reusable for the long term. Clarifying the timings for data releases and depositing can be very important for individual team members' publication plans and their relative ability to deliver outputs. So these dates need to be set realistically and made crystal clear to all team members, especially when they are stipulated by the funder (often quite speedily following the project's windup). Being clear here can save long-term disagreements and confusion over data ownership and authoring, that can otherwise prevent data-sharing from taking place in a timely way at the end of the project. In addition, some forms of work – such as a project-linked PhD or post-doctoral project – may continue after the main project

effort has been completed. The early career researchers involved here need complete clarity on what data is 'reserved' for them to handle, and for how long, and what will be more quickly released.

Data storage, security and sustainability

DMP templates may not always refer to data storage and security practices explicitly. However, research teams of all types (large and small, quantitative and qualitative) will need a clear plan for where they will store and back up data in a way that meets all their legal and ethical obligations *during the lifetime of the project*, while allowing for the data concerned to evolve and keeping analysis findings safe and secure. This issue is separate from the long-term data storage solutions to be used for OSS or preservation purposes.

There is no one-size-fits-all solution for this topic. Most universities and labs have their own detailed data-security policies and guidance in place. So in the first instance, any researcher attached to such a research institution should study up on and rely on local guidance for advice. A good general rule is that data storage solutions should be safe from cyberattacks or tampering, secure against unauthorized access and fully compliant with the legal data protection regime in the country where the research is based. Where research involves international collaboration, the provisions must be compliant within the country that has the strictest data protection regime represented on the research team. For example, a researcher operating in a country with no solid data protection framework who is collaborating with researchers from South Africa, will need to follow the South African data protection regulations, POPIA – and all data should be stored in line with the POPIA guidelines. Similarly, American researchers collaborating with scholars or scientists in the European Economic Area would need to follow data rules compliant with the EU's GDPR (General Data Protection Regulation) rules.

A further useful rule is generally to employ at least two different data storage solutions – so that personal data can be stored separately from data that is anonymized. For example, when recording interviews with research participants, the original transcribed interview recordings should be kept separately from anonymized versions, and both should be kept separately from participants' signed consent forms. Although these datasets will still be linked by unique identifiers for each participant, separating out data stores can help ensure that even if one breach of a dataset occurs, the wider risks to participants are contained and mitigated.

Data security also means considering device security, especially for laptops, tablets or other devices used at home, or taken out of the physical security access measures that may apply in labs and university premises. Again, your local IT services team should have guidelines in place for the minimum standards of security protection needed for every device to be used in undertaking

the research – and devices need to be kept up to date as time goes by. Common sense measures here include:

- up-to-date antivirus software and malware detection to your local corporate standard or better;
- password policies that mandate long-form, hard-to-guess passwords; and
- general advice on encryption and security.

This guidance does not need to be set out in full within the DMP, indeed it generally would not fit in funders' restrictive word count limits. But a general assurance statement is needed, especially that all members of the research team have guidance on minimum device security guidance for their own personal devices, and that team members have passed appropriate cyber security awareness training at their university or lab. At the least, a single line in the DMP must state that all devices used in the project will be protected to institutional standards and in line with relevant data protection legislation, and follow through on that promise.

Normally this means that most data from research projects ends up being stored on the institutional servers of the relevant university or lab, usually a good solution because they are generally secure, backed up regularly, and constantly monitored by university IT departments for both technical problems and security breaches. However, this solution can also pose a challenge if some or all of the researchers involved then move on to a different university or lab. Most institutions usually want to delete the contents of any personal storage drives within a set time of someone leaving the university or lab. Taking precautions against the accidental or time-expired deletion of data that is no longer in use at your previous institution is important for OSS, because problems around people moving can result in the irretrievable loss of research data. Luckily, OSS also proffers the solution to this conundrum. By ensuring that all your data is prepped as you go for timely deposit in a recognized data repository, you can best ensure that your data is stored sustainably in the long term, in one place, and that all data authors retain full access even after moving to a different lab or university. Committing to data sustainability and moving your data into a repository as soon as it has been prepped for deposit is a confidence-building element of any DMP.

You can also mitigate against potential accidental data losses or damage from cyberattacks by planning ahead for contingencies. For example, there should be secure backups of all data, particularly any that that is being stored on physical storage devices such as USB flash drives and external hard drives (easy to use, often high capacity, but also easy to lose and subject to hardware corruption over time), or on local storage on laptops, home PCs or tablets. University and lab IT services should be able to advise on the automatic storage backup options they can provide and advise on the institutional policy on backups.

You might also seek out helpful advice on 'alt' solutions like paying for additional private sector backup packages – such as Apple, Microsoft or Dropbox, and other cloud storage. For non-problematic data these can be a sound supplementary investment at only a small extra annual cost. These services can be particularly useful for backing up *your* own research data – draft texts, PDFs to be read, research ideas, etc. – which is separate and distinct from collected research data that is regulated by legislation. There are still some risks, though. Apple and Dropbox are not automatically GDPR compliant, so projects involving European researchers will need to ensure that storage is allocated on a European server. This can often be arranged via direct liaison with the companies concerned. Any backup solutions need to be checked for compliance with national data protection regimes and local university requirements.

There may also be a requirement to access secondary data under specific terms. This can be commercially sensitive information, such as data from corporations or other business interests, or microdata that provides information about individual characteristics or households. Such data is usually subject to the signing of licence agreements, and sometimes there will be charges attached. Often, as with national government datasets, the information is free to use for research purposes, but requires a separate application form to be submitted justifying usage of the data and explaining how it will be accessed and stored safely during and after analysis. These access requirements often go beyond storage on the typical university server – for example, you may need to access the data via a trusted research environment (TRE) such as a SafePod or another safe setting. Some data suppliers may specify what country the data can be stored in, and that the data cannot be stored on a networked drive or can only be accessed offline. It is important in cases like these to seek out the advice of experts within your institution; this may be your IT department, Data Protection Officer, Information Governance Manager or Data Librarian. Where needed, the DMP should signpost that you will need to access such restricted data, and explain that the team understand the requirements and has a sound plan in place to meet them – for example, saying: 'I will need to access [named dataset] via the ONS Secure Research Service and will do so via the SafePod solution.' Many possible safe settings may be used to store commercially restricted or sensitive data. Which one is right for any particular dataset can be agreed in consultation with institutional IT teams and data suppliers.

In all such cases, researchers should seek out legal advice where possible, preferably from a university or lab legal team. Some data licence agreements specify strict penalties for breach of contract, or restrictive IP clauses that may constrain your ability to publish findings or share derived datasets. Preferably with assistance from an institutional legal team, researchers will need to establish with data suppliers exactly what data outputs can be shared and in what format. For example, qualitative researchers could request to share codebooks where including full-text responses may be too high risk. Or users of commercial quantitative datasets may agree with the data supplier to only share

summary tables. Many researchers will find that established data suppliers like the Office of National Statistics (ONS) will hold data outputs on a secure server and only release them upon internal review. This arrangement should also be flagged in the DMP as it demonstrates an in-depth understanding of all data outputs and due consideration of what can be shared.

Securing personal and sensitive data

In most countries data protection legislation comes with specific terminologies that refer to sensitive types of data. The EU's GDPR rules use the key terminology about 'personal data' and 'special category data', which is also used in the UK and in European Economic Area (EEA) countries (although in the UK it is followed under UK GDPR). Almost all social scientists are likely to be collecting data that relates to individuals, either in terms of survey responses or in terms of recording people's behaviours (like posts, tweets or social media activity, or their behaviours in markets or over healthcare). So every researcher really must have a basic understanding of the legal categories and concerns that relate to personal data collection.

According to the Information Commissioner's Office of the United Kingdom (ICO 2023b) '*personal data* is information that relates to an identified or identifiable individual'. It is important to recognize how broad this definition is. It clearly includes all direct personal identifiers like names and addresses and any forms or identification number. However, it can also cover indirect identifiers like people's ages, job titles, general locations, voice recordings and personal stories, etc. In the right context, any of these broader indirect identifiers could inadvertently reveal an individual's identity or allow a skilled interlocutor to combine even anonymized data so as to narrow down the scope of the people who may be being captured. In addition, *special categories data* is identified by the Information Commissioner's Office (ICO) (2023c) as 'personal data that needs more protection because it is sensitive'. This usually concerns information about inherent personal characteristics such as ethnicity, sexuality and biometrics data, and areas where people are entitled to strong privacy, such as their political and religious beliefs, personal health data and trade union membership.

Mishandling any kind of personal data can have serious consequences for all actors involved in the research, both participants who have chosen to help researchers by taking part, the research team (and their lab or university also) and the bodies funding research. For research participants or 'subjects' a data breach could result in increased risk of identity theft, 'doxxing', reputational damage, loss of employment and financial loss, or even running foul of the law or state authorities – all depending on the nature of the data breach, the subject matter of the research, and the countries or organizations where it takes place. For researchers (and their home university or lab), a data breach could result in substantial fines from the data protection agency of the relevant country, plus serious (and often enduring) reputational damage. Data breaches also risk damaging the wider research landscape because negative

news stories may discourage participants worried about future breaches, particularly participants from vulnerable communities, and affect trust in the wider research community adversely. If your work involves collecting *any* of these types of sensitive data, that must be recognized at the outset in the DMP and a plan outlined for responsibly and carefully handling all personal data, and sensitive data in particular. Demonstrating that the work will always use safe, secure and legally compliant data storage and security is key.

An additional useful tool, called the data protection impact assessment (DPIA), can be helpful in mitigating risks for any projects where the research topic or methodology are likely to result in higher risk to individuals. The Information Commissioner's Office (ICO) (2023a) gives a checklist of data types where using a DPIA may be advisable, and another one where it is mandatory to use this extra approach. No matter how large or small your project is, if you are collecting any personal data from research participants, you should consider the DPIA checklist for sensitive research and whether it is needed for your project. In any DMP document, mentioning that you have already carried out a DPIA check for sensitive research, or plan to do one prior to collecting data from participants, again demonstrates to reviewers that data management responsibilities have been taken seriously.

7.3 Identifying and assessing all your data

Over the lifetime of the project consider what formats different types of data will arrive in, how they may need to be transformed, and what end form they will be finalized as. In general, avoid importing uncommon data format types that cannot be accessed without highly specialist software packages. If you already have a repository for your data outputs in mind, check their guidance on recommended data file formats (see Annex 6A). Using accepted format types from the start will save you time when you eventually come to deposit your data (UK Data Service, no date a). Consider too how data formats might change over time. For example, an mp3 or mp4 file of a recorded interview might then be transcribed into a Word document and later possibly into tabular data.

Where projects involve large scale data collection, or extensive re-codings or transformations of data, with multiple new variables or computed indices defined, consider what final file sizes will be generated. Will they fit within your university's standard personal or project storage levels? Do they breach any limitations on data size in the analysis software to be used? For example, in late 2023 the file size limits in Excel were actually a fairly restrictive 256 columns, and 5,000,000 cells in total (implying a row limit with all columns used of 19,530). Other spreadsheet files loaded in Excel could not be larger than 20 MB, and needed to be under 400,000 cells with the same limit of 256 columns per sheet (implying a limit of just 1,563 rows with that many columns). Earlier software versions are typically even more restrictive. If storage size could be an issue for your project, contact your local IT services in advance to discuss options for expanding data storage or expanded hardware capacities.

Documenting and describing your data

Most DMPs are likely to have some questions that are designed to assess your planning in this area (see Chapter 6). Repositories and funders both see effective data documentation as preventing projects losing data through disorganization or shoddy data storage practices, and as vital if your data is to be read and understood by other users. So your answers should cover both how data labelling will be handled in the collection and analysis stages, and how the data will be documented at the end of the project. The two are linked because a clear plan for data documentation and description at the DMP stage will make it easier to prepare your data when you are ready to share your data outputs. In effect, most of the work will already have been completed.

In the DMP, describe any work you have undertaken to document and describe your data more thoroughly – such as planning for data catalogues or indices, generating interview guides and transcription templates, effective file organization strategies, anonymization logs, read-me files and details of file formats, etc. The plan is an overview document, so do not go into detail about how these will be utilized here, especially since some DMPs have word count limits. It will be enough to show reviewers that you are aware of documentation issues and plan sound strategies in your project.

Assuring high data quality can look different in the social sciences from other disciplines, and within the discipline group quality assurance measures can differ a lot, depending on the method of data collection (UK Data Service, no date c). Most commentary on data quality assurance is geared around data cleaning and clearing anomalies from large scale datasets (see page 143), and works well for quantitative researchers. Qualitative researchers might benefit from thinking about how to ensure higher data quality from the point of first collection. For example, prep and test audio recording software, run mock versions of focus groups to see what is capturable and what not from busy conversations and multiple interactions, or visit observation sites ahead of time to run checks on the environment. If more than one researcher is available to conduct, transcribe and analyse interview data, consider implementing inter-coder agreements where researchers share a sample of coded transcripts to test shared understandings, and develop a codebook to ensure consistency. Such measures improve the robustness of findings and boost data quality for sharing (with codebooks deposited alongside interview summaries, guides, or any allowable transcripts).

If the main (substantive) data collected in a project is too commercially, legally or ethically sensitive to share openly in a repository, it can still help other researchers to share the full methodological framework used in the study and to make data description tools like codebooks, indexes, transcription templates and interview guidelines openly available – all steps that demonstrate a commitment to open research principles to funders. At a minimum all data outputs should be accompanied by a data index or catalogue that describes each part of the data accessibly. Imagine yourself in the shoes of

researchers new to the project. Would they be able to understand the records for your dataset? Maybe also ask an uninvolved colleague from your university to take a look and give comments.

Some DMPs may ask you to attach any *supporting* documentation, such as relevant guidelines produced by your university on information governance policies, cyber security minimum standards, institutional password policies, IP policies, etc. Anything added here should show reviewers that the project team takes a well-informed approach to data management. With restrictive word counts it can help to say that all hardware and software to be used will be secured in line with university or lab recommendations, leaving the tech specifics to be covered in links to that guidance from this section.

Responsibilities within the research team

Individual or solo researchers will most often combine being the Principal Investigator and the Data Manager on their own projects, and so can skip this section. In small teams of two or three everyone may plausibly still assume an equal share of data management. However, for somewhat larger collaborative projects it is essential to clearly assign roles, and to do so early on – since the decision is integral to data security, data integrity and data sharing. In the DMP itself you do not need to provide full lists of individuals assigned to the project with roles and responsibilities, but just show that you have considered these issues. Word count limits often mean that you cannot elaborate much here. However, internally project members should discuss this in some detail.

For bigger teams, the DMP should include a project workflow listing everyone involved in data collection, writing code, developing databases, monitoring data standards, etc. Will a specialist Data Manager role be created, and with what responsibilities? For instance, will they make sure everybody on the team has or receives requisite training in data management and cyber security? Will they liaise with data librarians, Research Ethics and Data Protection Officers to ensure that institutional standards are met? And will they ensure that any shared file storage is compliant with agreed practice for file naming, versioning, etc.?

If ancillary work is being done outside the project team itself (e.g. by paid transcribers or translators) there should be a clear plan for how they collect data and share it with the project team. You may also need to consider having specialist data sharing agreements or non-disclosure agreements (NDAs) in place for external staff, particularly in cases where:

- data is being collected from firms and corporations (who will often require NDAs) or from government agencies in sensitive areas;
- the personal data collected may be sensitive; or
- any research is based outside the EU, EEA, or UK.

Writing a DMP within a restrictive word count

Some funders require DMPs to be as little as 500 words long, and here they may ask for a general data management statement rather than a full plan. In theory restrictive limits should reduce the time and resources burdens of writing funding applications, but shorter DMPs (or data management statements) can be harder to write and be sure what information to include and what can safely be left out. In the best-case scenario, a data management statement with a low word count would be supplemented later with a full DMP, once funding has been granted and the project is ready to begin in earnest.

Much of the information recommended for DMPs here may not fit within restrictive limits. Instead of providing a full picture of the data management lifecycle, try to drill down to the basics, focusing instead on what data will be collected, the formats involved, and how it will be shared long term. For example, begin by briefly outlining any key secondary datasets to be used, particularly if they are only available on a restricted basis due to sensitivity or confidentiality concerns. Also focus on secondary data that is most likely to be complex or high risk for your project and discuss the issues raised there to pre-empt reviewer concerns.

Next, move on to outlining the how and why of any new data to be collected, key methods and the likely data format after analysis. This is more challenging for projects with multiple workstreams or mixed methods, so perhaps use bullet points or tables to lay out the essential points and save words. Give only brief descriptions of data collection methods, and if a statement or short DMP is for a bigger funding bid proposal, refer to methodology descriptions in the main application, without duplicating materials there.

However small the word count allowed, reviewers need to be reassured that data will be stored safely and securely in at least a sentence or two – for example, a statement that 'all data will be stored on university servers which are backed up by IT services and compliant with GDPR'. With internal or external collaborators, consider what data they will need to access and how that will be achieved using accepted storage solutions. Check that the project will clearly meet any funder frameworks or specific reviewer expectations for data management, such as the ESRC Research Data Policy (2018). A line such as – 'in line with the expectations outlined in the ESRC research data policy' – can save your having to expand on common sense data management principles.

Finally, even in a brief data management statement, outline your broad plans for making data shareable and discoverable long term. Generally 'spend' more words here than in previous sections, outlining which of the datasets to be produced can be shared, and any restrictions or anticipated difficulties in doing so. Indicating commitment to FAIR principles (without elaborating each principle) can save words too, since in most cases reviewers will be familiar with them. Set out how you will share data in a recognized research repository, as with a full DMP.

7.4 Depositing completed data and documentation

When you have fully clarified your plans for what to collect, analyse and deposit, and have assembled (or determined how to assemble) the necessary files and documentation, some logistics and organizing issues at the deposit stage will still need decisions (Data Documentation Initiative 2025). They may seem unproblematic and quickly done, but depositing can in fact be more akin to the tasks involved in getting journal articles uploaded onto their sites. Filling in a given repository's deposit form may raise additional or new issues, since they each have somewhat different requirements.

Choosing a repository and setting an access level

Make sure to choose a dedicated research data repository to share your outputs. This ensures that your data receives a unique DOI number, and it provides a citation separate from any project publication DOIs. A good repository provides assurance that your data is stored sustainably for the future, and so is highly unlikely to be lost or corrupted. Using a repository is much better than storing data with journal publishers (who may not provide a dedicated DOI or make the data as discoverable as an indexed repository) or using a project website (which may quickly become unusable once the project is no longer actively staffed or underway). Good research repositories also offer guidance on best practice in depositing data, and make it straightforward to meet most FAIR criteria.

Many universities (and big labs that operate semi-permanently) have their own dedicated repositories, and using such a local institution repository may mean it is convenient to talk to staff who can offer active help in uploading data. Hopefully, universities in general endure for long periods (centuries in some famous cases), and in advanced industrial societies most have secured, long-term funding. But remember that their component departments and labs may change more quickly, and smaller institutions may merge or be absorbed by larger ones, just as with firms – with potentially adverse implications for any local repositories below university level or run by 'fragile' universities.

In the social sciences, however, a number of excellent and free-to-use repositories exist, giving plenty of options in first-rank solutions. At LSE we usually recommend:

- ReShare, the UK Data Service national repository;
- Zenodo, which was commissioned by the EC for sharing EC funded research;
- Harvard Dataverse, backed by Harvard university and open to all; or
- a national government data repository, especially one that specializes in the social sciences, as with the Finnish Social Science Data Archive (FSD) (Late and Kekäläinen 2020). This is especially appropriate where the research and dataset is in a language other than English, or one not widely spoken overseas.

Any of these well-known repositories, funded and run by major institutions, should help guarantee long-term preservation of your data.

Data depositors must choose an access level for people to reach their data. There are typically three tiers, although this can differ between repositories:

- *open* – data is made freely available to everybody. Users do not need to establish an account with the repository in order to access and download a dataset. This choice is the best one for maximum openness to any potential user worldwide.
- *safeguarded* – data is slightly or somewhat more sensitive. Consequently, users need an account with the repository and they may need to first agree to some terms and conditions on how they reuse the data.
- *closed/restricted* – data is too sensitive to be made freely available, so that permission for other users to access the data needs to be sought from the data owners who deposited it. Often this entails potential users following an application process where they need to prove scholarly intent or a public interest reason for why they want to access the data. Access via a trusted research environment (TRE) such as a SafePod or other safe setting may also be necessary.

Some repositories have an option for depositors to embargo their data for a (relatively short) period, often so that authors can restrict data access while they are pushing out publications. A typical embargo period could be anywhere from 3 to 12 months, but you will need to negotiate an embargo period with the repository and with any funders with a stake in the research. Funders may often disapprove of embargo periods over 12 months, without exceptional reasons to justify such restrictions.

Assigning a copyright licence to your data

Repositories also ask depositors to choose what kind of licence they want to use in order to make it discoverable and reusable by others. Usually this would be a form of the CC Creative Commons licence. Essentially this is a mechanism by which you 'license' users of the repository to freely reuse your data so long as they cite you as the creator. In Chapter 15 on publishing open access, Figure 15.2 shows the full set of CC licences with a brief explanation of what is permitted and what is disallowed under different variants. It is a good idea to check your university's or lab's IP policy as it relates to data depositing before you choose a particular type of creative commons licence. The best practice for OS is to choose a CC–BY licence that is as unrestricted as possible, only requiring that reusers of the data source attribute it to the original researchers – but leaving them free to modify or add to it. However, some institutions may require a CC-BY-NC (or non-commercial) clause on a licence, which means that the data can

only be freely reused for non-commercial purposes. If your dataset also has commercial value, the NC licence allows scholarly/scientific researchers free reuse, but not companies seeking to make money from the data. Researchers should be careful before applying NC clauses in creative commons licences. It is reasonable for academics and universities to have concerns about corporate or business users monetizing their data collected for research purposes without compensating them. However, there is some ambiguity over how far NC restrictions can extend. For example, it may also prevent users at private universities or paid speakers at conferences from reusing the data without a commercial agreement in place with the data owner.

As modern social science research methods become more participatory or use 'citizen social science' approaches more (see Chapter 13), so researchers have increasingly asked participants to produce data themselves – for instance, to document their past experiences or current daily life using videos, photographs, artwork or diaries. In any such case the original works produced by participants automatically become their original property. To use them researchers hence need to ask participants at the outset to assign joint copyright over any such works to you. Joint copyright is also essential if original outputs or creations are to be used in published work or shared with a data repository. The necessary permission is best accomplished by including a statement on the research consent form that participants sign – 'I agree that joint copyright of the [specific outputs] be assigned to [researcher name]'. It is vital, however, that this provision is explained well to participants in the project description or information sheet, and/or in person, so that they are clearly aware of what it means.

On large collaborative research projects it is particularly important that all of the researchers involved discuss together all data copyright and ownership issues at the outset of the research process. Finalizing a DMP is the ideal time to have this conversation. The ownership of data should be clearly assigned in every instance. Does it rest with the PI (Principal Investigator) alone, or with a group of senior authors of the projects' research outputs, or with individual data collectors? If sections of the overall project data are to be reserved for PhD or postdoc research projects, how will this work when most project data is deposited? And when might the reserved sections also be deposited or become open without restrictions? Every member of the team should be fully aware as early as possible that data outputs are going to be shared and that on deposit a CC licence will be applied that enables other scholars to reuse the data. In the past, misunderstandings within project teams about who owns which data, and what can be done with it, have posed stumbling blocks to OSS. So it is important to make sure that everyone is on the same page as early as possible. This issue links closely to wider issues around assigning credit for work done on larger projects, and for authorship (see

Chapter 14). Bear in mind that when using a data repository you can credit as many people as you like, and that OSS is best fostered by assigning some credit to all staff involved. In particular, some individuals who contributed to the research team's work but were perhaps not involved in the key project publications, can receive due credit for their work on deposited data.

Conclusions

From the earliest stage of a research project, a data management plan is a key tool to demonstrate to funders and reviewers that you are committed to the principles of OSS. Most funders now expect data outputs from publicly funded research projects to be shared in the public interest and to be available to contribute to knowledge within the wider research community. DMP template prompts ask you to consider what data outputs can be shared, on what access level and where they will be made discoverable. So a DMP is a key tool to demonstrate your own understanding of the importance of OSS and to outline how you plan to meet that requirement.

DMPs should not be understood as just appraisal or surveillance tools of funding councils, reviewers or regulators of research, although they do have some of these functions (Hall 2016). Rather, writing a DMP can also be seen as a key stage in capping and formalizing researchers' otherwise easily broken good resolutions to document research as it gets done fully and accurately, to critically assess and maintain data quality at each stage of the work, and to contribute as effectively as possible to the further development of their discipline, by making FAIR data open access soon after the end of the project.

For qualitative research the problems around data depositing will normally require more careful management. However:

> data repositories and academic libraries can partner with researchers to support three challenges associated with qualitative data sharing: (1) obtaining informed consent from participants for data sharing and scholarly reuse, (2) ensuring that qualitative data are legally and ethically shared, and (3) sharing data that cannot be deidentified (Mannheimer et al. 2019).

The chapters in Part III give more details on ways to enhance the openness of research conducted using one of the five main qualitative research methods.

Because of some metadata deficiencies around images in the humanities (Hansson and Dahlgren 2021), depositing them in ways that are easily reusable may present some extra difficulties. But since this is perhaps the most important use of repositories in the humanities, they are well worth addressing, as set out in Section 7.1.

Annex 7A – data repositories to consider

Name (URL)	About	Organization	Country	Metadata
Commons Open Repository Exchange (CORE)	CORE is part of Humanities Commons that was launched in 2016 to support a network for people researching and teaching in the humanities. In addition to the repository, the interface provides web publication opportunities, and teaching materials and tools. According to the website, in April 2021 there were 26,678 registered members, and almost 12,000 'Core deposits', comprising datasets and image files but also including articles and books https://hcommons.org/core/	Non-profit, the Modern Language Association	USA	
Harvard Dataverse	Harvard Dataverse is an open-source web application being developed at Harvard's Institute for Quantitative Social Science (IQSS), together with a global network of collaborators and partners (Dataverse, no date). The aim is to share and preserve data from any research discipline and to explore and analyse research data. In April 2021, the website reported that there were 108,436 datasets and 993,701 files in the system https://dataverse.harvard.edu/dataverse/harvard	Non-profit, Harvard University, Institute for Quantitative Social Sciences	USA	DataCite Metadata Schema, Dublin Core

Name (URL)	About	Organization	Country	Metadata
Figshare	Founded in 2011, Figshare's mission is to store, manage and freely disseminate any kind of research output. According to the website, Figshare provides research data infrastructure for Wiley, SpringerNature, Taylor & Francis and PLOS, among others, and collaborates with a large network of academic institutions to enable open research. In April 2021, Figshare counted 1,364,584 datasets and 1,876,879 figures in their system https://figshare.com	Commercial. Funded by Digital Science (part of the publishing company Holtzbrinck)	UK, USA, Romania	DataCite Metadata Schema, Dublin Core
Zenodo	Zenodo is an online repository for sharing publications and data. Launched in May 2013 as part of a European-wide research infrastructure, the Zenodo repository was specifically designed to enable a universal data repository for individual researchers providing data sharing in a wide variety of formats across all fields of research. In April 2021, Zenodo counted 77,453 datasets in their system and 596,877 image files, mainly from Nuclear Research, European Commission Horizon 2020, and FP7 programme projects. In 2024 Zenodo had a limit of 50GB per record, and a maximum of 100 files https:// zenodo.org	Non-profit, CERN, OpenAIRE, European Organization for Switzerland	EU	DataCite Metadata Schema, Dublin Core

Other helpful sources

- All European Academies (ALLEA): published a report in 2020 with recommendations for FAIR data in the humanities.
- DataCite: offers a search interface to find all kinds of registered research outputs, including unpublished reports and datasets.
- Digital Curation Centre: offers examples of DMPs, including those from researchers in the arts and humanities.
- Mantra: a short online course from the University of Edinburgh for all those who deal with digital data as part of a research project.
- Research Data Management and Sharing (Coursera): an in-depth online course about research data management, run by the University of Edinburgh and The University of North Carolina at Chapel Hill.
- The Turing Way – an influential collective project aimed at increasing the reproducibility and reusability of computational and scientific research.
- The UK Data Service Learning Hub offers free interactive tutorials on data skills.

Note

* For their patient support and invaluable help as reviewers I would like to thank both my current and former managers at LSE – Helen Porter previously Research Support Manager, and Rosie Higman, Open Research Services Manager at LSE. I also want to extend special thanks to Spyros Syrrakos of the LSE Law School, for his invaluable help in keeping my half of the LSE's data library service running while I concentrated on writing.

References

Association of Social Anthropologists (ASA) (2021) 'Ethical guidelines for good research practice', web document, accessed 1 June 2023. https://perma.cc/S9QC-JTUR

Bishoff, Carolyn and Johnston, Lisa (2015) 'Approaches to data sharing: an analysis of NSF data management plans from a large research university', *Journal of Librarianship and Scholarly Communication,* vol. 3, no. 2, article eP1231. https://doi.org/10.7710/2162-3309.1231

British Library (2021) *A Guide to Research Data Management,* June. https://doi.org/10.23636/0pbq-2r89

Corti, Louise (2018) 'Show me the data: research reproducibility in qualitative research', *Data Impact blog*, 18 September. https://perma.cc/XN4Z-W3BU

Data Documentation Initiative (DDI) (2025) 'Data documentation initiative', home page. https://perma.cc/75HF-EV36

ESRC (Economic and Social Research Council) (2018) 'ESRC research data policy', accessed 26 September 2024. https://perma.cc/FCK5-UPMG

go-fair (no date) 'FAIR principles', accessed 31 May 2023. https://perma.cc/AHZ3-S6AW

Hall, Gary (2016) *The Uberfication of the University*, Minneapolis and St Paul: University of Minnesota Press. $ https://doi.org/10.5749/9781452958439 Multiclicks form of OA: https://manifold.umn.edu/read/the-uberfication-of-the-university

Hansson, Karin and Dahlgren, Anna (2021) 'Open research data repositories: practices, norms, and metadata for a sharing images', *Journal of the Association for Information Science and Technology*, vol. 73, no. 2, pp. 303–317. https://doi.org/10.1002/asi.24571

Information Commissioner's Office (ICO) (2023a) 'Data protection impact assessments', accessed 18 May 2023. https://ico.org.uk/for-organisations/law-enforcement/guide-to-le-processing/accountability-and-governance/data-protection-impact-assessments/

Information Commissioner's Office (ICO) (2023b) 'What is personal information? A guide', accessed 18 May 2023. https://perma.cc/B73M-CXVH

Information Commissioner's Office (ICO) (2023c) 'Special category data', accessed 18 May 2023. https://perma.cc/PQ6U-Y9TZ

Late, Elina and Kekäläinen, Janna (2020) 'Use and users of a social science research data archive', *PLoS ONE*, vol. 15, no. 8, article e0233455. https://doi.org/10.1371/journal.pone.0233455

Mannheimer, Sara; Pienta, Amy; Kirilova, Dessilava; Elman, Colin; and Wutich, Amber (2019) 'Qualitative data sharing: data repositories and academic libraries as key partners in addressing challenges', *American Behavioral Scientist*, vol. 63, no. 5, pp. 643–664. https://doi.org/10.1177/0002764218784991

STM, DataCite and Cross Ref (no date) 'Joint statement on research data', Web page. https://perma.cc/6CZ8-834C

Wilkinson, Mark D; Dumontier, Michel; Aalbersberg, Ijsbrand; Appleton, Gabrielle et al. (2016) 'The FAIR guiding principles for scientific data management and stewardship', *Scientific Data*, vol. 3, article 160018. https://doi.org/10.1038/sdata.2016.18

Williams, Mary; Bagwell, Jacqueline; and Zozus, Mederith Nahm (2017) 'Data management plans: the missing perspective', *Journal of Biomedical Informatics*, vol. 71, pp. 130–142. https://doi.org/10.1016/j.jbi.2017.05.004

UK Data Service (No date a) 'Recommended formats', accessed 14 May 2023. https://perma.cc/EEK2-X9KC

UK Data Service (No date b) 'Costing', web document, accessed 29 May 2023. https://perma.cc/G2AC-JGY7

UK Data Service (No date c) 'Quality', accessed 30 May 2023. https://perma.cc/DVB6-D7GN

8. Open data for publications – replication archives and research methods annexes

The payoffs from 'documenting as you go' (covered in Chapter 4) are only partly realized by depositing all of a project's research data in a permanent repository at its very end (covered in the previous chapter). Long before this, at the key stage of submitting articles to journals, PhDs for examination, or book proposals to publishers, the presence or absence of good supplemental documentation can make a key difference to whether the research is regarded by reviewers as credible, worthwhile and publishable – or not. Supplementing *all* publications and final texts with additional data resources gives readers and reviewers higher levels of assurance about the quality and exact provenance of their exhibits – the key places for data presentation in the main text. Unlike comprehensive data depositing, underpinning publications requires a sharp focus just on the data files to support each chart, table or diagram presented, providing the information that readers need to understand and ideally reproduce any exhibit or attention point.*

For journal articles these additional materials are commonly called a 'replication archive' (actually not a very accurate label for the kind of extra information and data needed to make the output reproducible). Nonetheless, this is the term now widely used, and an increasing number of journals specify what such an archive should consist of. Most guidance is framed for articles reliant on quantitative data and code, and many journals now require authors to provide a finished replication archive before their article can be finally accepted and progressed into production – creating a strong incentive to get the extra materials delivered. However, what is asked for still varies greatly, even from one quantitative journal to another, as Section 8.1 discusses. Where PhDers pursue a 'papers model' PhD thesis or dissertation, replication archives may also be needed for each chapter in paper format.

How to cite this book chapter:

Dunleavy, Patrick and Monteath, Timothy (2026) 'Open data for publications – replication archives and research methods annexes', in Dunleavy, Patrick and Monteath, Timothy (2026) *Doing Open Social Science: A Guide for Researchers.* London: LSE Press, pp. 177–190. https://doi.org/10.31389/lsepress.dos.h

For research book authors, the authors of research-based chapters in an edited book and people writing a 'big book' thesis there are unfortunately few similar formalized expectations. Perhaps the best-developed guidelines here are the practices used for approving PhDs in internationally rated universities – even here though enforcement is mostly 'cultural' and reproducibility norms are rarely fully formalized or systematized. Thesis examiners and PhD committees for big book dissertations, commonly expect candidates to go beyond just a single Chapter 2 on methodology, and to also include sufficient annexes to give the examiners assurance in detail that the original data and information-gathering work was undertaken well, and that the analyses made were conducted appropriately. This expectation is sometimes extended to reporting for post-doctoral projects also. Here we argue that in the open social science (OSS) era all research book or edited book chapter authors should provide a research methods annex (RMA) showing in detail how their work was undertaken (Section 8.2). This might be a comprehensive account at the end of a thesis or published book, or a permanently available online annex in a reputable repository.

8.1 Replication archives for journal articles

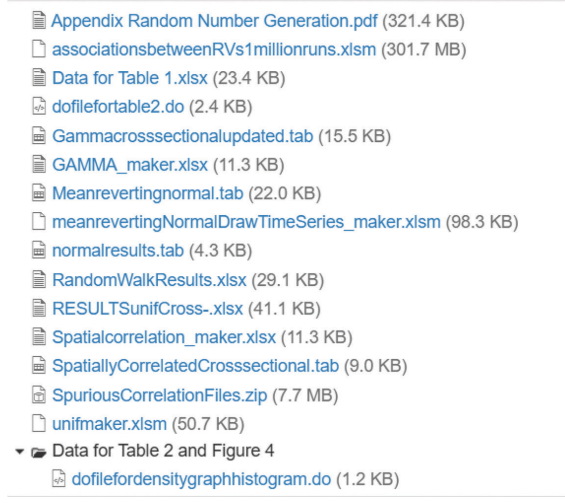
In journal articles, especially in the quantitative social sciences, the core messages of the paper are normally carried by the exhibits – that is, all the charts, graphs, tables, diagrams, images or photographs, and boxes that encapsulate the key findings. The paper's main text describes in broad terms how the information in the exhibits was collected and analysed, together with any limitations, and may discuss how the findings fit with other relevant literature, or not. The replication archives for articles normally:

- are focused tightly on underpinning each figure or table, arranged in the same order as they appear in the main text;
- are relatively short, serving only to establish how a given exhibit was produced in the form shown; and
- may consist of different materials from one exhibit to another.

Replication archives normally give readers and reviewers access to the underlying data used to produce an exhibit. Authors are often reluctant to do this before a project wraps up. However, the data revealed can be tightly confined to what is actually shown in the figure or table. Unveiling large amounts of underlying data may not necessarily be essential if the author says enough for a later reader or reviewer to reproduce what has been done and to check the exhibit's validity. This is a difficult line to fix but a couple of examples may help to show more clearly what is involved. Figure 8.1 shows the abstract for the replication archive deposited by Richman and Roberts to accompany a 2022 journal article, and its structure.

Figure 8.1: The 16 files listed in the Richman and Roberts replication archive for their paper, 'Assessing spurious correlations in big search data'

Article abstract, also Dataset Description
Big search data offers the opportunity to identify new and potentially real-time measures and predictors of important political, geographic, social, cultural, economic, and epidemiological phenomena, measures that might serve an important role as leading indicators in forecasts and now-casts. However, it also presents vast new risks that scientists or the public will identify meaningless and totally spurious 'relationships' between variables. This study is the first to quantify that risk in the context of search data. We find that spurious correlations arise at exceptionally high frequencies for variables following gamma and spatially auto-correlated distributions, and random walks. Quantifying these spurious correlations and their likely magnitude for various distributions has value for several reasons. First, analysts can make progress towards accurate inference. Second, they can avoid unwarranted credulity. Third, they can demand appropriate disclosure from study authors. (14 September 2022)



Source: Online deposited datasets: Richman (2022).
Original article: Richman and Roberts (2023a).
Supplementary Materials for article: Richman and Roberts (2023b).

The authors say that all the key steps included here:

> provide a description of the procedures and calculations used to generate each simulated random variable, followed by a description of the procedure used to input each variable into Google Correlate (Richman and Roberts, 2022, p. 1).

Yet the immediate usefulness of this information is restricted by their admission in the main paper that 'the specific tool used to collect the data used in this study (Google Correlate) has since been discontinued by Google' (Richman and Roberts 2023a, p. 293). However, the authors note in mitigation that 'other tools continue to offer the capacity to examine many searches'. (As an aside: this is an example of how relying on proprietary IT systems can seem to smooth the way for quicker or easier research, yet also restrict transparency, and may (as here) additionally rule out any future reproducibility by other authors.)

Our second example comes from a perhaps more intuitively understandable 2021 paper by Hooey, Peeters and Principe (2021) again with the abstract and replication archive structure shown in Figure 8.2 below. Here each figure or table has an accompanying data resource giving details of what is shown and how data was gathered or computed.

How should replication archives be designed and what should they contain? Some good practice tips include:

- ensure a Readme file is included in data deposited, guiding future users to the original paper and any online annex at the journal site, with full publication details, DOIs etc.
- data deposited should correspond to the exhibit (or interim analysis leading up to an exhibit) with no unexplained or undersupplied variables or data points, but also no excess or unused data.
- give readers guidance on how lengthy, difficult, or complex an analysis may be.
- check that the file types deposited are widely available so that analyses can easily be reproduced on other PCs or by other colleagues. Avoid having to respond to other people's difficulties by saying 'I'm surprised – it works for me.'
- for analyses based on random numbers, include the seed numbers used so that another user can reproduce the exact same numbers used in the exhibit.

Do journal requirements have an effect on how much supplemental data authors provide? A recent study evaluated the impacts of introducing mandatory reproducibility requirements in the journal *Management Science*, comparing 500 articles before and after the journal introducing a new Data and Code Disclosure policy in 2019 (Fišar et al. 2023). Before the new code came in only one in eight articles voluntarily provided replication materials and only 55% could be largely reproduced. After the policy change 95% of articles could be largely computationally reproduced, but in only 71% of cases was all of the dataset available to reviewers. Nonetheless, this was an important improvement in transparency (Fišar et al. 2023).

Figure 8.2: The replication archive for Hoey, Peeters and Principe's paper, 'The transfer system in European football'

Abstract	
We assess the proclaimed pro-competitive effects of the 'transfer system', the no-poaching agreement governing the European football (soccer) labor market. A major argument to legitimize this system is that transfer fees, which hiring clubs pay to release players from their current clubs, redistribute revenues from large market to small market clubs. This would strengthen small clubs' financial clout and their ability to compete in sporting terms. Player transfer fees represent over 10 billion Euros in asset value in the financial statements of the 202 clubs we analyze. Still, small market clubs rarely obtain substantial revenues from the transfer market. The main beneficiaries are clubs around the middle of the market size distribution. A select group of large market clubs makes significant transfer losses, but this does not undo their initial financial advantage. Overall, the transfer system therefore leads to a very minor reduction in revenue inequality.	
graphframe.csv	(24.2 kB)
graphframeage.csv	(28.52 kB)
sourcefile.csv	(140.67 kB)
Figure 1.pdf	(19.51 kB)
Figure 2.pdf	(29.26 kB)
Figure 3.pdf	(20.64 kB)
Figure 4.pdf	(52 kB)
Figure 5.pdf	(8.09 kB)
Figure 6. pdf	(6.02 kB)
Figure 8 CL.pdf	(6.73 kB)
Figure 9.pdf	(30.12 kB)
Table 1.csv	(0.6 kB)
Table 2.csv	(0.7 kB)
Table 3.csv	(0.35 kB)
Readme.pdf	(91.64 kB)
replication file. R	(46.7 kB)
Figure 8.pdf	(12.47 kB)
Table 4.doc	(2.3 kB)

Source: Online deposited datasets: Peeters, Principe and Hoey (2021).
Original article: Hoey, Peeters and Principle (2021).

8.2 Research methods annexes for books (and theses)

Just like any doctorate, academic research books must be based on sound methods, systematically applied. Critical readers of academic work must never be left in any doubt about exactly how findings and results have been generated, their precise provenance and their meaning. PhD examiners feel this imperative acutely because part of their task is to guard the portals of academia from the admission of faked or inadequate research. Yet the spiralling of 'replications crises' in many social science disciplines, and doubts about conventional peer review for journals, have also triggered a growing recognition that for research books publishers' closed peer review processes can rather easily let defective work through the net. As a result, it is increasingly common ground now to expect that research findings in books and book chapters should be supplemented by additional resources showing how data collection and analysis was done, in far more detail than is appropriate for main text chapters.

How can this requirement best be implemented? Traditionally research monographs or big book PhD theses mostly included a separate methodology chapter, usually placed immediately after the theory and literature review — so positioned at Chapter 2 if both these other functions are combined in the first chapter, or at Chapter 3 if these two functions get a chapter each. This conventional approach no longer works well in the OSS era for reasons set out below. By contrast, the RMA option lets the early presentation of methodology issues in the main text focus solely on theory-relevant fundamentals. And it can accommodate the ultra-detailed methods information needed for quality assurance and reproducibility far better.

What's wrong with the conventional 'Chapter 2 Methodology'?

An isolated chapter explaining methods in abstract terms can often be a dull and not very useful piece of academic text. For many books or PhDs, whose authors are using a fairly standard approach in their discipline, an early methodology chapter often says little more than: 'I used this perfectly well-known method, and I implemented it in the normal way, barring a couple of difficulties specific to my data or situation that needed some tweaks along the way.' An early methodology chapter also creates a common dilemma for authors. Once embarked on writing a whole chapter, authors may become acutely aware that the 2,000 or 3,000 words really needed will look 'too thin' on their own, and so often feel compelled to inflate the chapter to look more comparable to others. Even if they stay rather high-level and theoretical, penning a short chapter (of 5,000 words or less), it will be devoid of most of the detail needed for useful transparency or reproducibility.

On the other fork of the dilemma, many authors cram lots of highly specific information about their procedures into Chapter 2 alongside the major method themes. Conceivably, then, a good deal of methods information

is available all in one place – but often not in an understandable form. For instance, you may tell readers how you recoded variables *X* and *Y* in an abstract way, before they know anything about how the data on these variables was patterned. And you may defend the usefulness or insights flowing from analytical steps taken before readers have seen results of any kind. The progression of ideas and arguments in your thesis or book stalls in the mud of a hundred detailed battles which we have not even seen fought out yet. A lot more text then 'engages with the literature' to explain why the authors choices represent a perfectly legitimate approach, whose usefulness in this context no one may in fact be questioning. Cue long discussions of methodological 'isms' (and perhaps epistemological/philosophical macro-perspectives) that may well be light-years removed from the actual data and problems in hand.

Where authors are good at handling this kind of thing, the resulting text may be dull, but the book at least jumps through the conventionally expected hoops. However, for authors who are really empiricists, 'coming up for air' in an unfamiliar meta-level discussion, then it is all too easy to get out of their depth. The same syndrome may afflict PhD candidates trying to make an overly grandiose defence of their methods approach, a mistake that can then create potential oral exam (viva) time bombs as more neurotic PhD examiners try to put them right on any methods-theory misconceptions thus revealed.

The conventional Chapter 2 can also become dead weight text just because too much ground is covered prematurely, remote from readers actually getting to see any of the data, whether quantitative or qualitative. At the start of a book (and even more in a PhD thesis) readers normally only have your word for what you have actually accomplished. Your status and competence as a researcher remains to be established for them, and a conventional methodology chapter on its own does little to strengthen their confidence. After it, readers will know what methods choices you have made, but did these work out well or not? Without seeing the methods set in context against the data gathered, the analysis undertaken or the results achieved via these choices and procedures, readers just cannot tell at this stage of a big book text.

Dunleavy (2014)'s well-known 'build, blur, corrode' criteria stress that every part of your text must (ineluctably) do one of these three things in developing its overall argument. An overly detailed conventional methods Chapter 2 can all too easily become 'blur' materials, adding remarkably little to the book's exposition, because material is being handled in the wrong place and has to be repeated later on anyway, as readers get to see data, analyses and results in a concrete way they can better appreciate.

The other great, mildly corrosive theme of many a methodology chapter is the pre-emptive defensive move where authors argue: 'I did the best I feasibly could, given the limitations of the data available to me.' Even where this is perfectly true, banging on about it at length in Chapter 2 is dull and off-putting. For post docs trying to get a book from their monograph, retaining the conventional methodology chapter second in the running order is also a dead giveaway to publishers that the work started life as a PhD – helping to trigger a desk reject. Even if the book proposal starts well, with an interesting

theory chapter showing the relevance, novelty and importance of the work, it is a self-destructive move to slump straightway into 'what I did in my research' materials in Chapter 2.

Conventionally, then, Chapter 2 becomes a low-energy, low-value zone in your book (or thesis), dashing the readers' hopes built up in the motivating first chapter. Most professional readers familiar with your book's field will rarely encounter anything novel there. At PhD level a whole chapter solely covering methodology issues that are standard in the field simply prolongs the pages of text that examiners have to wade through before they see one original thing they did not know already. (A side tip here for doctoral folk doing a big book thesis: count the pages involved in your theory introduction, literature review and methods chapters combined. If it is more than 50 the examiners are bored. If it gets towards 100 pages they will be pretty impatient/sceptical by now.)

Why then does the conventional/expected pattern persist so stubbornly? Many research book (and PhD) authors believe that they have a 'duty' to inform colleagues of exactly what they did in a study. This is completely true, and nothing above should be taken to undermine the importance of thinking through clearly and presenting fully the epistemological and methodological issues that arose in research. But does your understanding of these issues have to be expressed at chapter length, in chapter style and located in the main run of chapters at this early stage of the argument? Can you honestly answer 'yes' to any of these questions:

- are your main methods innovative or controversial so that they need extended explanation for professional readers?
- do they solve a problem that has previously stymied science or scholarship?
- are your methodology changes, variations or innovations part of the *core value-added* of your research book (or PhD)?
- do readers really need a chapter-length exposition of your methodological choices in advance of seeing the data, so that a shorter exposition as part of your initial theoretical statement really would not suffice?

If you can answer yes to any one of these questions, then your book or thesis may indeed need a methods chapter early on. Stick to the standard approach, because for you methods innovation and originality is central and integral to the work done. However, if not, a different approach may be more useful.

The alternative – a brief methodology section early on, plus a full RMA at the end

If you are not a methods innovator, and have confined your methods changes to adapting and tweaking already well-known approaches in your field, a better alternative is to include a short methodology section within Chapter 1, your main theory chapter, and provide a comprehensive RMA at the end of

your book (or thesis). Of course, academic readers of any serious research work (and even more PhD examiners) must feel fully informed about what an author did and why. But in the current open research era doing without a separate methodology chapter in no way means leaving your methods obscure, but only presenting this essential information in a different way.

In the RMA route far greater transparency is achieved at three points:

(i) The opening of the book (or thesis) fully covers any major methods issues briefly (say, in 3,000 words), normally in the last section of the first, theory chapter. In academic work it is a natural progression to move from discussing major theory controversies or differences to considering the methodological approaches that are often integral to key debates. Here you make clear where you stand in methodology (as well as theory) terms, what broad approach you have followed and how you have tweaked or adapted techniques to better fit your problem. But do not go into exact details at this point.

(ii) More specific methods information (but still focusing on important issues) is then provided within subsequent substantive empirical chapters, as the data gathered is outlined and analysed. Concrete methods issues are best discussed fully and once only (following the 'say it once, say it right' maxim), and as close as possible to the points where detailed analyses are undertaken or key findings presented. Even expert readers (and certainly PhD examiners) cannot normally assess the viability or appropriateness of a methodological approach until they see what the data produced look like and what results have been achieved. Looking at the findings themselves provides essential cues that can allow in-depth reviewers to identify problems and assess the strong or weak methods choices made. Equally, armed with their substantive evidence, authors can now make a far more plausible case that their methods choices were correct or useful ones. For instance, they might show that they have undertaken sensitivity analyses without that affecting the results, or demonstrate that alternative ways of tackling data or analysis problems in fact produce convergent solutions. By contrast, authors can only assert such claims without proof in an early methodology chapter.

(iii) Finally, the RMA comes immediately at the end of the main chapter sequence, at Annex 1, and gives a comprehensive background on the methodological and epistemological choices made in your research, alongside an exact description of every aspect of how your research was undertaken. The discussion of issues here can be as extended and detailed as needed. If the full RMA is a very bulky and complex document, it may be deposited in a recognized repository online and permanently linked to from the book text via a shorter synoptic description of its most salient contents.

An RMA does not have to be written in the normal chapter style, although it normally mostly is. Researchers should use any style best suited to giving a detailed exposition of the work done. Online RMAs especially are not constrained in terms of word count length, the number or size of their exhibits or datasets, or the level of detail included. Instead, the annex aims to give detailed ('reproducibility') information about every exhibit to any other scholar or scientist following in the authors' footsteps or handling similar data. Both in combination with the datasets deposited in a repository, or without that backup, the RMA also provides exact guidance on the methods used and decisions made during analysis on how to proceed at each stage. A well-done annex will normally be far more specific and directly useful to other scholars than a discursive methods chapter placed early in the main sequence. And because it is located at the end of the analysis, when readers have already seen your data and findings discussed in earlier chapters, a detailed exposition of the dataset's provenance and the analysis behind exhibits strengthens the author's clear status as a provenly successful researcher (an important consideration at PhD level also).

The RMA approach recognizes also the powerful role played by 'tacit knowledge' in restricting the reproducibility of scientific and scholarly work. An RMA invites you to try and put yourself in the shoes of another skilled researcher and to set out all they need to know in order to replicate your approach in producing a chart or table. Of course, you cannot succeed in making *everything* explicit, as many philosophers of science have emphasized (Collins 2010). But by trying in this way you will get a lot further down a useful road than any other course.

Of course, the time needed for an RMA is a drawback. For an author finishing a 'big book' research book having to undertaking the RMA can take time that you begrudge in the last, rushed months of your project. Yet its payoffs are post-publication, in better book reviews, an improved status and reputation for the work done (helpful for promotion), and in due course more citations. For PhD students who plan to 'paperize' big book thesis chapters after submitting, an RMA can also generate the materials needed to help meet journal demands for a replication archive to back your papers (Section 8.1).

Including an RMA, either succinctly in print at the end of the volume, or more expansively in an online repository, is fundamental if books (and perhaps book chapters also) are to meet modern open research criteria. And of course, in the OSS era, supporting advice and datasets published as free-standing entities can attract scholarly credit in their own right. For original researchers seeking to ensure that their large datasets can voyage out alone in the world and be reused on their own, they need to be implemented in much the same ways discussed in Chapter 7 above, albeit focused more tightly here on underpinning exhibits in the published research. A post hoc re-version of the DMP document – covering how data was acquired and cleaned, how variables were coded and recoded, and how indices were constructed – can

also help the RMA to be crystal clear, as well as showing the provenance and supporting data for the multiple exhibits often included in the book, chapter or thesis.

Academic conventions, reviewers and examiners

Academic conservatism is the main block to the wider use of RMAs. PhD examiners reviewing research monographs are often ultraconventional, and a second chapter tackling methodology is sometimes even mandated in the reporting or progression forms that some departments use. Older supervisors did it this way in their youth when they did their doctorates, and for them not having a whole methodology chapter early on may seem like a heresy, something to be anathematized. Professional readers of research books are more flexible, but they may still expect what they are used to.

As always, the stickiness of academia is a factor to consider. We do not want to deny that expected forms can be important, especially in disciplines that adhere a lot to conventional or standard pro-forma macro-structures for the chapter sequence in PhD theses. This approach is common in more quantitative social sciences and many disciplines in science, technology, engineering, mathematics and medicine (STEMM). Our advice here is to discuss the RMA ideas in this section with colleagues, and gauge what they think of our suggested solution, namely:

- a methodology component in the main theory/scene-setting first chapter;
- detailed discussions of methods issues directly adjacent and related to the presentation of analysis approaches and results; and
- a detailed RMA that ensures your research is reproduceable and tackles methods choices in great detail.

However, across most of the humanities and the 'soft' social sciences it is important to recognize that past reliance on conventional formats, and academic resistance to doing creative writing better, has weakened greatly in the current century (Pinker 2014, 2015; Dunleavy 2003; Cayley 2023). Social scientists increasingly accept that research books (and ideally even big book PhDs) should be creatively designed, with subject-specific structures optimizing argumentative development, and written clearly and well for maximum academic impact (Dunleavy and Tinkler 2021, Part I).

When it comes to getting a publisher for a research monograph book, those authors who have already adopted an RMA approach have multiple advantages by junking any separate methodology chapter in favour of embedded methods discussions in the theory and the main data presentation chapters (as above). For most academic readers your book will likely be most successful if you keep preliminary epistemological or methods details to the minimum that readers need to know, but at the same time

provide a detailed RMA as part of the permanently available datasets and materials deposited in your university's repository, a national data archive, or a cross-national repository like Zenodo or the Harvard Dataverse (see Chapter 7).

For PhD students wanting to publish a version of their big book doctorate, the key advice is similar. Just junk from your thesis manuscript any conventional Chapter 2 on methods, only retaining key methodological or epistemological debates or points and your resolution of them in a somewhat longer opening theory chapter. Discuss detailed methods close to the data analysis (not ahead of that), and create a full online RMA, perhaps with a book annex to give readers the most salient study background information. You can also put in URL hyperlinks to the version in your university's online e-theses archive within the book at all later relevant points. Lastly, deposit the relevant datasets in a suitable enduring repository. Implementing these changes will allow your streamlined and 'bookified' version of the PhD text to focus consistently on its main arguments, maintaining a constant developmental tempo to your writing throughout, and minimizing the changes needed from the thesis text.

Conclusions

For social scientists, maximizing openness in research needs to go a lot further than (eventually) depositing the full datasets used at the end of the research process, or when funders require it (see Chapter 7). Instead, every research article, book or book chapter – that is, any stand-alone open research publication – should ideally lead to a focused form of supplementation that presents immediate additional information for readers about how the research was conducted, how variables were defined, what the data looked like, and how the specific analysis reported on was undertaken. For journal articles the norms and requirements for creating good replication archives across different disciplines are broadly understood already, but detailed practice is still evolving. For research books the development of coherent RMAs is at an earlier stage and authors may often need to innovate in ways specific to their particular projects, within the broad concepts set out here.

The gains to reproducibility that follow are primarily for the immediate readers and reusers of each published work. But the effort to document alongside publications as they begin to flow from the project can also have wider benefits for the original researchers. It is an effective way to learn how open data depositing works, and how users respond. Thus, it can grow an appreciation of how the project's final master dataset can be best deposited in ways that grow OSS. Completing replication archives and RMAs as you publish may also often lead to corrections, reversioning and iteratively improving plans for a comprehensive data deposit from a project.

Note

* This chapter draws on research and seminars forming part of the CIVICA Research programme, which was funded in 2022–2024 by the European Union's research and innovation programme Horizon 2020, under the chapter 'Science in Society'.

References

Cayley, Rachael (2023) *Thriving as a Graduate Writer – Principles, Strategies, and Habits for Effective Academic Writing*, Ann Arbor, MI: University of Michigan Press. **$** https://doi.org/10.3998/mpub.11664964

Collins, Harry (2010) *Tacit and Explicit Knowledge*, Chicago, IL: University of Chicago Press. **$** https://perma.cc/HXP9-FCMB

Dunleavy, Patrick (2003) *Authoring a PhD: How to Plan, Draft, Write and Finish a Doctoral Dissertation or Thesis*, London: Palgrave (now Bloomsbury), pp. 157–196. **$** https://www.bloomsbury.com/uk/authoring-a-phd-9781403905840/

Dunleavy, Patrick (2014) 'The "build, blur, corrode" test can help you edit new writing – every part of your text does one of these three things', *Writing for Research* blog, 27 January. https://perma.cc/W9V9-JS9A

Dunleavy, Patrick and Tinkler, Jane (2021) *Maximizing the Impacts of Academic Research: How to Grow the Recognition, Influence, Practical Application and Public Understanding of Science and Scholarship*, London: Palgrave (now Bloomsbury). **$** https://www.bloomsbury.com/uk/maximizing-the-impacts-of-academic-research-9780230377608/

Fišar, Miloš; Greiner, Ben; Huber, Christoph; Katok, Elena; Ozkes, Ali I; and Management Science Reproducibility Collaboration (2023) 'Reproducibility in management science', *Management Science*, vol. 70, no. 3, pp. 1343–1356. https://doi.org/10.1287/mnsc.2023.03556

Hoey, Sam; Peeters, Thomas; and Principe, Francesco (2021) 'The transfer system in European football: a pro-competitive no-poaching agreement?', *International Journal of Industrial Organization*, vol. 75, article 102695. **$** https://doi.org/10.1016/j.ijindorg.2020.102695 OA version at: https://perma.cc/VR4Y-FX2R

Peeters, Thomas; Principe, Francesco; and Hoey, Sam (2021) 'Replication archive', *International Journal of Industrial Organization.* Dataset at DataverseNL. https://doi.org/10.25397/eur.12901292 or https://doi.org/10.34894/N792OO

Pinker, Stephen (2014) 'Why academic writing stinks, and how to fix it', *The Chronicle of Higher Education*, 26 September. https://perma.cc/8FZM-UDAL

Pinker, Stephen (2015) *The Sense of Style: The Thinking Person's Guide to Writing in the 21st Century* (London: Penguin). $ https://perma.cc/CQ99-WBC4

Richman, Jesse (2022) 'Replication data for: measuring and answering the challenge of spurious correlations in big search Data', Harvard Dataverse, V3, UNF:6:dHnXWqr/uOdJcH/y9Aog1g== [fileUNF]. https://doi.org/10.7910/DVN/UW1UYR

Richman, Jesse T. and Roberts, Ryan J. (2023a) 'Assessing spurious correlations in big search data', *Forecasting*, vol. 5, pp. 285–296. https://doi.org/10.3390/forecast5010015

Richman, Jesse T. and Roberts, Ryan J. (2023b) 'Supplementary materials [for Richman and Roberts 2023a]: generation of random variables'. https://perma.cc/MM8N-3VCC

PART III

Achieving more open qualitative research

For empirical researchers using qualitative methods, many open social science (OSS) goals can seem to pose far greater difficulties than those faced by scholars using quantitative approaches, as with using pre-registration. Interviewing people and conducting case studies and fieldwork are essentially exploratory approaches. Project teams may set out some initial expectations and intentions in pre-registered documents before beginning their work, but they must always remain open and alert to discoveries, new meanings and unexpected developments that shift their research trajectories after initial contact with and learning from research subjects themselves. Nothing is more fatal for the reliability, integrity and accuracy of qualitative work than researchers who are so constrained by their prior understandings of what happened, what things mean, or how social processes work, that they approach research participants and experiences with closed minds. Interviews, case research and fieldwork also post problems for reproducibility because they are necessarily events that are unique in space, time and their social settings.

In systematic documentation work the existence of records that can be reinspected combats some reproducibility issues. Pre-registering the scope and intent of the planned research is somewhat easier. But again, the pathways that research takes must adapt to the gains in knowledge that researchers make from immersive reading and intensively analysing texts. In archival work recognition of this imperative has been strengthened amongst historians and social scientists by an acute awareness of the limits of what is recorded and how, which always necessitates critical interpretation of texts. Finally, on a quick look, some pre-registration and disclosure components of OSS approaches may seem to run partly at a tangent (or even counter) to

the growing shift towards 'citizen social science' (CSS), where research is co-produced with participants.

In this part we seek to show that these potential difficulties are tractable ones, and that OSS methods can already be applied well in adapted ways across all the major qualitative methods to greatly mitigate remaining issues. As OSS ideas spread and are applied widely and innovatively in qualitative work, so the range of feasible solutions to problems will also expand: we also trace some emergent trends in this respect. Finally, the effort needed to translate OSS practices into qualitative social science can also stimulate additional ideas for increasing its reproducibility and strengthening the assurance that scholars and the public can have in the reliability of our research.

We begin with interviewing, where only on-the-record interviews can be fully reproducible, while those that are non-attributable or off the record pose substantial problems for researchers seeking to demonstrate that their work is well-founded and reproducible (Chapter 9). Yet many practical steps can be taken to pre-register the main elements of interview work and to increase assurance about the content of more closed forms of interview. Ideas here include shifting interviewees towards more open formats, making better and fuller recordings of what gets said, opening up more information about non-attributable interviews in 'jittered' ways that still protect confidentiality, and creating quantitative metrics from interview data using innovative methods.

In case studies, different sets of issues arise around single or very small numbers of in-depth or 'diagnostic' case studies, medium numbers of intermediate cases and larger numbers of short 'apt illustrations' (Chapter 10). We show how readers and reusers can be given strengthened confidence in the work undertaken in each of these situations, especially where cases are anonymized, as they often must be. Being clear about the casing strategy being followed and giving enhanced contextual settings for cases can help. And newer methods focusing on qualitative comparative analysis, using cases within mixed methods studies, and developing comprehensive evidence-synthesizing approaches (such as 'systematic review' of cases in a given subfield) can all generate valuable extra insights and assurance. Similar steps can also increase the usefulness of fieldwork – for instance, strengthening the granularity of observations via computerized field journals and notebooks.

Turning to systematic documentation analysis, if the base records being analysed are easily available in full for (digital) reinspection by later project teams, then the potential for some forms of reproducibility or audit is greater. Researchers can do a good deal in 'show and tell' mode to help readers and reusers understand the character and limits of the original texts being analysed, and to demonstrate that their interpretation is accurate (Chapter 11). However, where reaccessing the base texts is not possible or practicable, as it often is not, then some similar issues with other qualitative work discussed above may recur. They can again be addressed by adapted forms of pre-registration; explaining exploratory adaptations to the planned scope or direction of research; curating a selection of key documents for later reusers

when feasible; counting within and across texts (even if the authors' main approach is immersive reading); coding text to make checks; and developing quantitative metrics to back up qualitative interpretations – for instance, via website analyses across multiple agencies or organizations.

Progress in developing the openness of archival research has been more limited, with no or scant practicable reproducibility for much archival evidence (Chapter 12). However, calls for improving on previously standard but limited historical scholarship methods have been made and 'digital history' approaches have diversified options. Although the potential for full digitization of older historical sources remains vast, social scientists and historians can help develop project-level openness for the materials they use, and help overcome legacy cultural constraints in moving archival work into more open practices.

Finally in this part, CSS and co-production approaches to social science have somewhat divergent aims from OSS. Some forms of CSS work have paid little heed to openness considerations, although the most participatory approaches have done so (Chapter 13). At root, however, both these streams of innovation can be well aligned in stressing the careful initial development of research proposals, and in ensuring that research participants' meanings and experiences are sympathetically documented and accurately interpreted.

9. Interview-based research

A powerful approach for social science researchers trying to understand different social situations and phenomena in depth is simply to talk to the people involved about what they experienced or did, and why. In-depth interviews draw on respondents' unique knowledge of their own lives, actions and situations, plus their perceptions of other actors, organizations and systems. Like other forms of in-depth qualitative research, interviews help researchers to develop a deeper and more contextualized understanding of actor's meanings and social phenomena. They contrast with the shallower understandings fostered by large *N* quantitative survey research, focusing most on the 'mass' perceptions and behaviours of many respondents, especially those that survive various tests of 'significance'.*

Qualitative research textbooks now provide huge amounts of guidance on how to guard against accepting biased, partial or inaccurate views from participants in many small *N*, talk-based methods, such as focus groups and participatory sessions as well as interviews. We make no effort to recap this literature here. Notwithstanding this progress, in open science (OS) terms a root problem of talking to people as a research method is that the back-and-forth talk between a questioner and a respondent is always specific to a given context, time and framing. We cannot fully assess the reproducibility of research interviews, because their original conditions will never occur again. Or, as Cratylus the sophist famously remarked (taking Herculaneus's dictum a logical step further), 'You cannot step in the same river once.' So how can interview-based social science become more open?

Our answer begins by recapping the essential features of research interviews, and other talk-based enquiry, focusing on the central importance of three different conditions governing how information was provided – off-the-record, non-attributably, or on the record (Section 9.1). In terms of openness, we briefly set out the most acute problems of interviewing as a research method, and the motivation for achieving open social science (OSS) improvements (Section 9.2). The practical steps feasible for boosting openness

How to cite this book chapter:

Dunleavy, Patrick and Monteath, Timothy (2026) 'Interview-based research', in Dunleavy, Patrick and Monteath, Timothy (2026) *Doing Open Social Science: A Guide for Researchers.* London: LSE Press, pp. 195–223. https://doi.org/10.31389/lsepress.dos.i

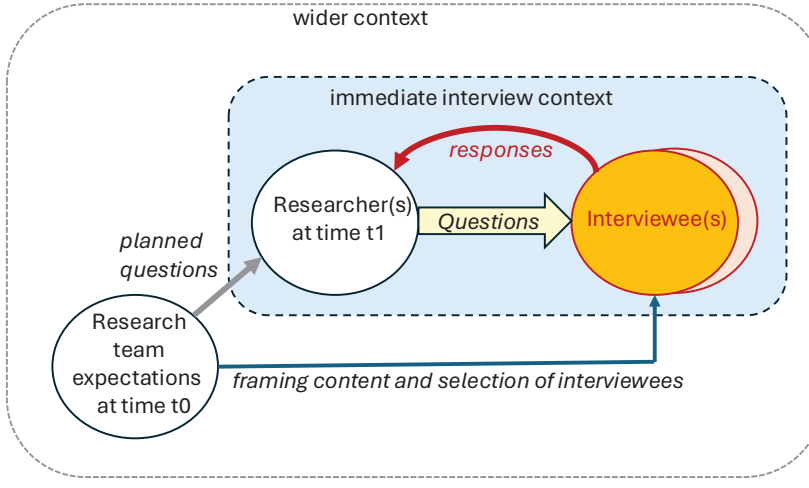
are considered next, beginning with the constraints of off-the-record interviews, the greater scope for innovation around non-attributable interviews, and how on-the-record interviews might also be made reusable (Section 9.3). Finally, deriving quantitative metrics from interviews and other conversational methods can help generate stronger measures of external assurance. We conclude that greater openness can boost the quality and impact of interview-based research.

9.1 Research interviews, dialogues and conversations

An interview is a guided conversation between researchers and participants. It is always framed by the project team's purposes and shaped by their selection of who to talk to, within a wider context (see Figure 9.1). More immediately, the outcomes of the interview are shaped by who is there on both sides of the conversation; the questions, follow-ups, prompts and agenda-setting used by the researcher(s) to guide the conversation; and the narratives, responses and feedback given by interviewee(s).

The immediate context can be shaped by the location, either with both (sets of) actors physically present in person in the same place; or increasingly now with remote but face-to-face interviews also on digital meeting software like Zoom, Teams, or Facetime; or without visual content in phone interviews (once frequent in academic work, but now declining, except amongst journalists).

Figure 9.1: Main factors shaping interview outcomes



Source: Authors.

There are three broad purposes for a research interview:

- *'Life world' interviews* seek to uncover the experiences of someone in a given social setting, and understand what guides their behaviour, meanings and perceptions. Here almost everything that the research subject believes, does or says may be of interest.
- *Causal influence interviews* aim to uncover what roles actors or systemic influences played in shaping a specific historical event or causal process, such as a government policy or an organizational decision. Researchers draw a strict boundary around the phenomena to be understood, and hence steer the dialogue away from 'out-of-scope' matters.
- *Exploratory descriptive interviews* are intermediate in scope between the two other categories. Here social researchers seek to make sense of why people or groups act as they do, right now, often in a 'first draft of history' mode. The project team may be interested in 'surfacing' diverse variables and inputs, especially in understanding new, complex or ongoing activities and choices. But they also want to focus on a specific topic.

The way in which an interview session is conducted shapes much of the outcomes, especially who is in the room asking or answering questions, what exact wording is used, and in what order. Although all interviews are 'structured' in some way, a conventional set of distinctions are often made between:

- (Fully) '*structured' interviews* where questions are asked and answered by all interviewees in a fixed order, historically to avoid introducing variations that might 'lead' respondents. (Sometimes now the order of items in a single question may be varied to try and control for sequence or wording effects.) Often only a portion of answers given are recorded, summarized or deemed relevant (a bit like a lie-detector test). Mostly employed now only to gather 'vox pop' extract quotes in quantitative surveys or market research, this approach has less use in modern qualitative research.
- *Semi-structured interviews* where questions more loosely follow a topic guide or set of core questions, but their order (and any follow-on questions and prompts) may be varied to achieve a more effective or 'organic' dialogue with the interviewee. Here the full dialogue is normally recorded, if this is possible, or reconstituted by the interviewer as far as possible if not. Most qualitative research interviews now are 'semi-structured', albeit in different ways – for instance, in lifeworld interviews tangents and deviations in the conversation onto unplanned topics may be acceptable for interviewers. In more focused interviews interviewers need to be alert to when respondents

are going well off-topic, and those where they actually are cuing up a connection that the project team was unaware of.

For elite and specialist interviews Dexter (2008) stresses that their inherently 'non-standard', 'exploratory' and even 'journalistic' nature requires a radically different approach from quantitative surveys or perhaps even much 'big qualitative' work. The researcher's job is not to move through a fixed list of questions, but to sustain a natural and unforced conversation with the interviewee, with a lot of latitude for them to explain how they saw situations, giving their own account of what went on, and explaining what they see as relevant. The researcher normally has an initial common spine of topic questions that get to be tackled in a sequence and depth shaped by the interviewees' answers, plus follow-on questions or prompts that may be asked somewhat differently in different situations – making reproducibility in such interviews hard to achieve. On personal, sensitive or confidential topics researchers now normally share the 'starter' (topic) questions with interviewees in advance, and elite interviewees may provide carefully written responses to the basic questions, making the follow-on stages around each main point even more important.

- *'Co-constructed' approaches*, where the interviewer uses neutral, encouraging prompts only to initiate and sustain a narrative (e.g. a video record) from the interviewee, unfolding in their own terms.

Multiple variations and combinations of these approaches are possible, and we cover some later on below.

Feminist researchers have argued that the power dynamics of research interviews may seem to thrust a lower status interviewee into an inherently subordinate role to their academic interlocutor, and have argued strongly that dialogue between the two needs to be even-handed and conducted in an atmosphere of care and respect (Oakley 1981; Oakley 2016; Herron 2023). However, it is important to note that quite obverse patterns may apply in other situations. A high status, difficult-to-access and elite interviewee may well have the 'upper hand' over a more junior researcher, as may an interviewee who controls a research team's access to an organization.

Paying experts for their time is common in business research conducted by corporations, especially where what is said has an immediate value or can be relayed to others, as when a consultancy may pass on paid-for expert views to their client firms. In general, however, interviewees are never paid in academic research, because payment would add a strong complicating kind of exchange relationship into the mix. Even paying online actors small fees to complete quantitative surveys, as with Mechanical Turk or other survey companies, raises additional issues (see Section 9.2).

Many other issues potentially surround the 'performative' nature of interviewing, including echoes of interviews in films or TV dramas (Denzin 2001), 'resistance' by interviewees to meeting researcher's wishes, and other

aspects. Carefully and self-consciously thinking through issues before and during the interview to frame questions appropriately, and dispassionately recalling and summarizing this immediately afterwards, may help to achieve a more 'reflexive' praxis (Burkette 2022; Beaunae, Wu and Koro-Ljungberg 2011).

Three main types and statuses of social science interviews

Any research interview entails a 'bargain' between the interviewee and the researcher over what use can be made of material that they talk about. As a general rule no ethical interviewing can take place without its status being made clear as one of the three options shown in Figure 9.2.[1] Interviewees or participants must always explicitly agree to one of these types by signing a disclosure form, and ideally being recorded as giving oral consent.

Off-the-record interviews are for 'background' only and cannot result in any direct quotation of what the respondent says. Knowing what interviewee X said about situation A allows researchers to hunt for other collateral sources to stand up this interpretation – like documentation, or statements made on the record by other people involved. But nothing covered in the conversation with X can be cited (even non-attributably) as a source. In business case studies companies often ask researchers to also sign non-disclosure agreements (NDAs), legally binding documents that prohibit them from disclosing any commercially confidential information. Normally, researchers must also keep secret whom they have interviewed, or which groups and communities co-operated with them. Rarely, it may be admissible to provide a fully anonymized list of people contacted, but only in situations where their identification would clearly be infeasible.

Off the record interviews pose the most problems for OS criteria, since there are no independent ways to corroborate who researchers have talked to, let alone what was asked or said in response. In addition, what gets recorded or remembered may be far less than is desirable for reproducibility. Elite interviewees in firms or public life may often refuse to allow audio or video recording of conversations with them, and may even restrict researchers' note-taking. Anthropological researchers working in groups and communities must respect limits that community members may impose on what is recorded and when.

In either of these situations, researchers need to make as many jottings as possible during an interview or conversation, and immediately after it finishes should compose a 'contemporaneous note' of everything asked and said in response. Having to rely chiefly on memories of the interview, when full transcripts are not feasible, inherently introduces vagueness into how researchers recall conversations. Provide a fully anonymized list in situations where identification would clearly be infeasible.

Non-attributable interviews preserve anonymity about who said what, but quotations can be made freely – subject always to the over-riding constraint

Figure 9.2: The three main types of research interviews in the social sciences

Purpose of the interview	Type of interview		
	Off the record – ***may not be cited or quoted at all***	**Non-attributable** – ***may be quoted verbatim but only in a fully anonymized way***	**On the record** – ***may be quoted verbatim and attributed to a named interviewee***
Understanding/ capturing the interviewee's 'life world' (or that of a group/ community)	Immersion into the perspectives and experiences of interviewees, for example anthropological studies of groups and communities	Deep contextualization of quotes used in research – especially in cultural or social anthropology	Fundamental in biography, contemporary history, ethnographic studies
Capturing one aspect or stream of an interviewee's experience	Informal, unofficial, or highly sensitive content from general interviews	Standard research interview, for example in medicine, health, social policy, consumer behaviour, business studies or public services	Official or professional capacity of interviewee
Understanding the role played by the interviewee in specific historical or policy events	Informal, unofficial, or very sensitive content/ perspective that is highly specific to the interviewee	Standard research interview adapted to dynamics of business and government elites	'Insider' understanding of actions taken by leaders, executives or professionals, especially history and business or policy case studies

that readers cannot identify who said what (even if they work away at doing so). This status nonetheless allows researchers to show readers (and data reusers) extensive materials to demonstrate that interviews indeed took place, and that quotations made come directly from interviewees. Instead of specific names, readers are told only some general characteristics of the person originating the quote.

Normally, non-attributable interviews are at least audio-recorded and in addition either a full transcript or a 'running log' of relevant comments is often available in text form. So researchers can quote what was said exactly and the precise words and phrasings used often vouch for the authenticity of the anonymized quote. Normally, notes and methods annexes also allow expert readers (like PhD examiners, journal reviewers or other researchers in the field) to check how often the author is relying on what just one or two people said, and whether the research design drew on a reasonable spread and diversity of interviewees. However, the data from non-attributable interviews may not be easy to deposit in a research repository, depending on multiple factors – such as how many interviews were conducted (in small *N* studies it may be much harder to conceal identities), and on how confidential and personally sensitive the information involved is (see ICPSR Institute for Social Research 2023).

On-the-record interviews allow the researcher to record what was said, and identify who said it, with the actor's permission. Sometimes, interviewees may go off the record for some sensitive questions, and so these (hopefully small) sections still have to be left out. But, in theory, on the record interviews available on audio or video are fully checkable (especially with text transcripts) and accessible to other researchers. Thus, they should also be straightforward to document and deposit in a repository.

In practice, things may be a bit more complex. Interviewees may get 'carried away' during the conversation and say things that they later regret, and perhaps subsequently try to deny or get deleted. In this situation, journalists normally take a hard line with elite interviewees, arguing that was said on the record automatically becomes public. However, academic researchers (especially early career folk) may decide to take a different tack and edit parts of the interview – especially if the comments made stray from the intended research topic, or could identify other interviewees or actors, or the research team feel that the interviewee may later come to regret what they said. It can be prudent to prevent situations where an interviewee later on objects to a whole transcript going live, because then they might 'row back' on their comments or demand to be able to 'edit' a transcript or recording before it is deposited. Researchers may also seek to avoid derailing their ongoing relationship with an interviewee, or their wider constituency or community – especially with elite interviewees where there is already a power imbalance that could weigh against the researcher. At the same time any such 'sanitizing' of interview texts should be noted in the final text as a redaction, and should not go beyond

what research 'fair play' considerations entail, nor create a misleading impression for later reusers.

Other talk-based data

Not all oral evidence used in social science research is collected from one-to-one interviews. In policy and business studies research, it is common for two or even three interviewees to be present, and they may contribute different insights and opinions, and often make their own recording or minutes to guard against being misquoted. This can create problems for researchers since respondents may restrict what they say to a particular public relations line or to 'safe' comments. Context also matters in how frank people are – even talking to one interviewee in settings where other people form an 'audience' can be problematic. Some similar issues may arise in other 'triadic or group' interviews with children (with parents or guardians present), or with couples, whole households or families, or patients and carers.

A further set of interview geometries may arise when there are multiple interviewers for a single interviewee, or several people on both sides of the conversation. These more 'group-like' interviews may be inevitable with professional or highly specialized elite interviewees (e.g. scientists, doctors, or specialist regulators) where their expertise can only be effectively tapped by having several research team members present to cover different aspects of the topic.

In focus groups the emphasis moves further apart from the researcher as the sole agenda setter. A group of participants (normally at least 6 people, up to a maximum of 10 or 12) are asked to discuss a set of prompts fed into a group discussion by a facilitator. Usually not one of the research team (so that they can be as neutral as possible), the facilitator's initial role is to invite comments on a topic or some initial 'stimulus' materials prepared by the research team. Thereafter facilitators concentrate on eliciting what different people present feel about the topic, for instance deliberately avoiding 'leading' anyone in the group by providing any information, and intervening to stop any speakers from being too dominant. Because the pace of discussion can be hectic, focus groups always need to be observed by two or more members of the research team and audio- and video-recorded, usually in multiple ways. The status of contributions here is always 'non-attribution'. So, normally, focus group transcripts can be deposited with anonymized names used. However, audio or video recordings cannot easily be deposited because it may be straightforward to work out exactly who said what from them.

In participatory research and deliberative discussions (considered in detail in Chapter 13) there may be far more 'priming' of group discussions with respondents. Both stimulus materials and extensive background information are often presented before discussion begins, and sometimes either academic experts or 'pro and con' speakers can present to participatory groups, starting off lengthy group discussion sessions (Willis et al. 2021).

9.2 Pressures for opening interview-based research

Interviews and talk-based data form one of the key tools in the qualitative social science toolkit. They can illuminate many aspects of social phenomena in ways that other approaches cannot. At the same time, they have some important limitations in generating precise information and easily sharable data. Accordingly, interviews have often been used in combination with other techniques – such as case studies (Chapter 10), systematic documentation analysis (Chapter 11) or archival data (Chapter 12). We consider the benefits and limitations of interviews through the lens of open research.

What interviews are good for

The meanings, understandings and emotional colouring of any social actor are always complex and manifold. Reduced to dots and dashes on computer disks by massing responses together in large *N* studies, understanding people's knowledge, values and purposes can pose huge problems of interpretation. But the data from a one-to-one interview normally has few such problems. After a researcher has spent 30–60 minutes with someone, discussing how they see the world, or their role in causing social change or events, they typically have a clear picture in their head of that interviewee's viewpoint and how it makes sense. In addition, their narrative may be crystallized by memorable details and quotes, especially those where the interviewee has thought through their position self-consciously, or has previously related a version of it.

Sometimes quoting interviewees' responses also reveals a lot about them as a person, or about the 'organizational culture' of the firms, agencies, non-governmental organizations (NGOs), etc. that they work for or the group or community they belong to. Quoting from interviewees' discourse and giving their reactions to questions or prompts can strongly underpin and improve the accuracy of the researcher's interpretation of their discourse. And for policy histories or explaining major social events it greatly helps surface different individuals' (or groups') perspectives and accounts of what happened. Interviews and accompanying anthropological/sociological approaches are often indispensable in seeking to understand the 'life world' of people whose social culture, life experiences and meanings could otherwise be difficult or impossible to appreciate. In all these areas, the precise discourse information generated from interviews can give a great deal of assurance to readers (or reusers of data) that the researchers' interpretation is correct.

Interviews are also an invaluable way of generating access to other forms of data. They are very effective at conveying the emotional, cultural and intellectual context in which decisions were made, or how respondents formulated their views of the world – exactly the kind of 'soft' data that can very quickly get lost as memories dim and only stripped-down documentation sources survive. They are often vital in achieving a 'first draft of history' after major events or policy shifts, by quickly recording how participants explained and

represented what they were doing. Interviews about elite causation involve achieving access to key people, and opening a trustful conversation with them, potentially opening the way to reach other potential interviewees. In 'cascade' or 'waterfall' interview techniques, researchers close out interviews by asking the current interviewee to suggest other relevant personnel. For life world or anthropological interviews, an interviewee from a given community may be crucial in securing contact with a wider group, and getting their acceptance of the research by others.

What interviews are bad at

There are numerous inherent limitations in talk-based information, so conducting interviews in ways that reveal interesting and relevant information for the research's purposes is not an easy task. One well-known limitation concerns anything adverse (such as a policy failure or an event or behaviours reflecting poorly on a group). Interviewees who were involved in a failure tend to give situational explanations of what went wrong, where external observers would give more dispositional accounts. When successes are involved, however, this difference is reversed – insider actors are quick to claim credit for things working well, in conditions where external observers ascribe more influence to situational factors or luck. Researchers can 'aim off' for such patterns, of course, but not remove their influence.

Conversation is best at surfacing people's values, memories, experiences and recollections, and capturing the inherently complex 'emotional' content of people's life worlds, life stories or involvement in specific decisions or events. Even here researchers will often need to engage an individual interviewee's interest and trust in the research project if they are to get past commonplace or first-order responses. For instance, members of a given community or group may reiterate only what they are 'expected' to say by other group members, or disclose information only if it serves the group's interests. This is particularly true of well-trained elite interviewees adept at repeating 'PR lines' developed by their party, agency or firm to explain what happened in any controversial case. However, interviewee 'resistance' can take many subtle forms, potentially misdirecting research enquiry (Beaunae, Wu and Koro-Ljungberg 2011).

Because researchers need to create trust with interviewees, and to sustain effective personal relations with them in the interview itself, it can also be complicated to keep interviewees on track in the conversation, so as to avoid subjects digressing at length into experiences that are off-topic. Common risks here include:

- irrelevance, with responses simply not addressing the questions asked;
- sanitized or lowest common denominator responses, a version of the PR answer where the interviewee only rehashes well-known facts or innocuous explanations, for instance, ascribing every aspect of how things operated to luck or chance; or denying any patterns in an

organization's or group's responses, to insist that everything was handled 'case-by-case' or *sui generis*; and
- over-extension or 'boosterism', where interviewees want to show their importance with other issues, decisions or groups with whom they were not in fact involved.

Of course, researchers can try to get an off-course interview back on track, but in doing so they also need to guard against asserting a 'power' relationship over interviewees, or even producing over-compliance, where an interviewee only seeks to tells them what they think researchers want to hear. For off the record or non-attributable interviews, ethical norms prohibit disclosing to any interviewee things that other people have said. Here researchers may only be able to probe in rather abstract ways about which versions of events they have been given is the most accurate.

On their own, interviews are also not a good means of surfacing detailed information on dates, or precise details of arguments or actions, or numerical levels of funding or regulation decisions. Respondents' recollections are often imprecise and rather vague. They may recall the emotional colouring surrounding events, but not the details of exactly what happened or who said what – unless the informant made a 'contemporaneous note' of things or kept a diary. Interviews become much more valuable if they lead to someone offering previously unseen documentation to back up their account. In addition, many matters covered in interviews (even off the record) can also be backed up by material in public documentation, such as official minutes of discussions, press or media commentary (especially around leaks), or social media posts. Even if the interview evidence is just for background, or can only be quoted non-attributably, it gives valuable extra assurance to researchers that their interpretation of information in the public domain is correct.

In conducting interviews, analysing or quoting from them, and depositing records of them, researchers must behave ethically, which generally means safeguarding the interests of interviewees far more than (say) happens with press or media interviews. A common mistake for students conducting their first research interviews is to emulate 'journalistic' approaches they have seen portrayed in films or on TV as a model. Interviewees may stray into social controversies, creating risks for themselves if their remarks were to be quoted (e.g. in the context of ethnic group conflicts). Or in elite interviews they may inadvertently defame others – potentially making themselves liable for lawsuits and damages. In such circumstances, minimal careful editing of transcripts or recordings for deposit to safeguard core interviewee interests may be inevitable. A small but significant and widely applicable issue here is that interviews often generate somewhat confused text. People start but do not finish phrases or sentences intelligibly, cutting across from one topic to another, or going back and forth in covering a single topic, and referring in complex ways to different aspects in the same sentence. Listening firsthand to confused conversations or reading raw transcripts of these sections of

conversation can make things appear a lot worse to readers. Interviewees may seem to be more confused or incoherent in their answers than was actually the case. In any quotations used in article or book texts, researchers have an obligation to represent interviewees in an acceptable light, without allowing technical or grammar confusions to obscure or predominate over the sense of what was said.

Key uncertainties around interview-based studies

In terms of creating reproduceable knowledge interview-based studies have many key drawbacks. For off-the-record interviews how can we be sure how many interviews took place and with whom? For non-attributable interviews we also cannot check with any named persons that an interview cited as happening did in fact take place. Even if assurance was greater in either case here, some key questions that other researchers might have are often difficult to get answers to – such as: who did you speak to? Would speaking to different people have surfaced different views? Especially with group interviews about an event, or elite interviewing around the causation for decisions, were those interviewed in a position to credibly answer the questions asked? Were they in the right locations, in the relevant roles and organizations, and at the right level of seniority to speak authoritatively about the time periods to be explained?

Other critical questions focus on the research instrument used, the core question set used in semi-structured interviews. Did the questions asked, or the way that the researcher sequenced topics over the conversation, tend to 'lead' respondents to answer in a particular (desired) way? Even if the core prompts used were unbiased ones, what about the interviewer's follow-up prompts and behaviour (such as encouraging or disapproving commentary or non-verbal signals)? Finally, given what an interviewee actually said (and in response to what prompts), how can readers of an interview-based study be sure that the analysis or interpretation made is a legitimate or credible one? Would the interviewee themselves accept the analysis made or would they reject it as fanciful or biased?

Of course, part of the answers here lie with the training in qualitative research provided by university social scientists, and the apparatus of rules, norms and integrity institutions which police research ethics (especially for early career researchers). They answer these concerns at the most basic levels, while the peer apparatus for refereeing books and journal articles addresses them in published work. Yet periodically all of these reproducible knowledge concerns are still regularly resurfaced by cases where apparently soundly conducted interview-based work turns out to be fraudulent.

Two examples of such scandals illustrate the risks involved. In 2014 a rare social science article appeared in the leading journal, *Science*, dealing with the subject of gay marriage. Co-written by a PhD student from Yale, and a top Yale political scientist, it claimed rare evidence that across successive rounds

of interviews with the same respondents it was possible to persuade interviewees critical of gay marriages to become more tolerant in their attitudes (Konnikova 2015). This rare experimental mitigation effect (sought but not found in many previous studies) was claimed to be especially strong where the interviewer revealed to respondents their own gay sexuality. The senior academic offered to help write up the study after seeing results with high survey response rates. The PhD student claimed to have raised hundreds of thousands of dollars in grant money to pay people who responded to the survey. After surviving *Science*'s peer review processes, the apparently highly original work was duly published. However, doubts about the paper later emerged when other researchers sought to emulate it (Brockman and Kalla 2016). A leading critic was puzzled because the scanty details of the methods included in the *Science* text suggested that the estimated costs of the interview fieldwork said to be involved were far too large a sum for a doctoral student to have at their disposal. Accordingly, they repeatedly wrote to the PhD student first author for any transcript evidence of interviews to back up the claimed quantitative coded evidence, but no such materials were ever supplied. The senior author's role in the article also apparently had not extended to any checks that these interview materials existed or any due diligence about the alleged research ever having taken place. Subsequently the article was retracted by the journal and disowned by the senior author.

In a wider-application case, one of the key recent developments in large *N* interview-based research has been Amazon's 'Mechanical Turk' system for cheap spot-contracting between survey sponsors and 'gig' interviewees who sign up online to answer questions for small fees (Ahler, Roush and Sood 2020). This approach is used a lot for customer research by businesses (who may not care overmuch about the type of people they are reaching) and social scientists (who really should). To reach the right kind of people for their topic, researchers can ask potential interviewees before hiring them about their demographic characteristics, behaviours and even their views, and the Amazon system plus internet evidence has provided some basic cross-checks that the answers given on characteristics were accurate. However, recent evidence has shown that respondents can use virtual private networks (VPNs) to shield themselves from any investigation of their locations, roles and qualifications – allowing them to fake their suitability for many different projects – and therefore garner more work through Mechanical Turk (Ahler, Roush and Sood 2020). Of course, researchers can employ improved checking algorithms to push back against fakery (which is often quite tricky to sustain in a coherent way). Yet the core uncertainty about exactly who the gig folk are in Mechanical Turk studies has remained.

Both these examples come from interview-based but semi-quantitative studies where only very small, disembodied 'snippet' samples of text taken from interviews were ever made available to readers. We argue that in-depth interview methods generate far richer data – such as giving longer discourse quotations, describing interviewee responses in much greater detail, and far

better contextualizing and situating the research – making qualitative studies actually far harder to fake. The two examples above unfortunately raised unwarranted doubts over in-depth interview methods, where in practice it is often easy for other researchers in the field to spot fakes or made-up materials, even where the interview status is non-attributable.

9.3 Practical steps towards opening different types of interview data

Some important technical factors matter a lot in making available information collected from the three types of interviews discussed above. Just how an interview was recorded, and the quantity, comprehensiveness, quality and precision of data available from it can vary sharply. Figure 9.3 shows how different recording technologies normally capture or fail to capture different types of information, and how the record made can result in difficulties for the openness of different statuses of interview. Assurance that an interview took place, and that its contents were fully and accurately captured, is greatest with audio or video records. But for non-attributable interviews these recordings cannot be deposited without completely betraying confidentiality about who was involved. Even depositing full interview typescripts may allow readers or reusers to cross-check different parts of the response and narrow down who the person interviewed was, which is an ethical breach for non-attributable interviews.

By contrast, the value of on the record interviews for making knowledge reproduceable emerges strongly in Figure 9.2 – yet this approach has traditionally been far less used in the social sciences than non-attribution. However, the growing use of remote interviewing software packages (like Zoom, Teams and FaceTime), with easily used, built-in recording and transcribing capabilities, has radically transformed the possibilities for good practice. In particular, it should make it easier for researchers to get interviewees to agree to full recording, and to the whole recording being 'on the record'. Similarly, the development of podcasting and videocasting has greatly increased the chances of elite interviewees agreeing to be interviewed 'on air', with the understanding that the conversation will be broadcast publicly from the get-go.

Why do elite personnel agree to be interviewed by academics about the causal patterns shaping decisions and policy choices in firms or government agencies? What motivates them? One key concern is that people want to 'get their side of the story on the record' – especially perhaps recently retired people, who are often the most interesting and frankest informants about firms, agencies and policy changes. So, persuading respondents to shift the status of non-attributable fully recorded interviews to an on the record status can be a great step forward in openness terms. So too can asking people to use formats that are more open from the outset formats – like podcasts or videocasts – rather than interviews in private. These are important steps that

Figure 9.3: The major gains and limits of different recording modes

Mode of recording	Information lost	Loss or compression	Practical issues
Written notes, made in (or near) interview	Any full or reliable record of discourse	Very high	– Taking notes while listening/talking is very difficult to do – Recalled notes may omit things, or be biased by the researcher's reactions, even if made contemporaneously – 'Field notes' comprehensively documenting all the researcher's experiences (e.g. settings, time context) as they occur may be useful – for example, for focus groups, participant observation
Typescript of audio	Character of the interviewee's voice and subtle audio cues in the conversation	Medium	– Time-consuming to produce, but using an agency or automatic transcribing can cut the burden – Easy to deposit for on the record interviews and *without any related audio files* for non-attributable interviews – Whole-interview transcripts may include much irrelevant materials – compared with, for example, keeping a 'running log' (see below) – *A full transcript of whole interviews (or comparable sections)* is essential for any quantitative coding and analysis, across all types of interviews
Full audio file	Facial expressions, and body language. Visual cues to the interview context	Low	– Retained audio files obviate the need for transcripts, although a 'running log' of topics discussed must still be kept, and quotations contextualized verbatim – High level of assurance from depositing complete audio files – *It is only feasible to deposit audio or video files for on-the- record interviews, or podcasts/ videocasts (below)*

(*continued*)

Figure 9.3: (continued)

Mode of recording	Information lost	Loss or compression	Practical issues
Video	The interviewee is continuously observable but the interviewer is normally not in shot	Low	– If more than one person is being interviewed, only the main speaker at any one time may be in shot – *Depositing is feasible only for on the record interviews*
Zoom, Teams online conversations	Not much is lost. Transcripts are also generated automatically	Very low	– Few drawbacks – easy to activate (with permissions). Recording works well, even with multiple people talking – High level of assurance from retained files – *Depositing is only feasible with agreement from all participants – normally easy to secure initially*
Audio podcast	Facial expressions, and body language. (The interview context is usually some form of studio)	Low. Edits made may not be visible	Pod or video casts can be more complex to set up. Editing to meet production quality standards may be needed, but poses risks of the record becoming incomplete – *Depositing is feasible for on the record interviews only*
Videocast	Both interviewee and interviewer are covered	Very low. Any edits made may not be visible	

modern interviewing approaches make feasible for almost all social science fields. We turn next to consider ways of improving the openness of the three types of interviews.

Opening more information about off the record interviews

Given the restrictive conditions around off the record interviews, most observers have concluded that making more information openly available about them is simply infeasible. Yet in practice it is still valuable to make some core OS steps, while completely preserving the anonymity of informants and without quoting at all from what gets said in interviews. In Chapter 4 we outlined pre-registration approaches that are relevant for qualitative work. So before interviews begin in the field it is valuable to pre-register the core question frames that will guide them, and also any follow-on prompts that might elicit documentation. Knowing these elements can be very helpful for other researchers. No extra work is needed since modern good practice is to send all interviewees an advance copy of the core questions and to use the same list on the day itself (albeit introducing questions and prompts into the conversation in different sequences to sustain a constructive dialogue). This 'no surprises' step is a basic courtesy to respondents. It lets them reflect on the questions before the meeting, and perhaps check relevant documentation, dates or other details so as to be as accurate as possible in responding. Of course, prompts going beyond the pre-registered initial frame questions normally cannot be covered because they arise from different interviewee's replies and comments. Yet some advance details of likely further questions may still be helpful, both for researchers themselves, and for other academics following their path.

Other useful things to pre-register (even if just in a date-stamped private document) are the planned strategy for selecting and gaining access to interviewees. Often initial plans for who to gain access to do not work out well because of low response rates, so new approaches must be tried or alternative interviewees sought. The initial strategy note should also be updated in a clear narrative way as interviewing unfolds – for example, by recording how 'cascade' interviewing links developed. This record of changes provides a useful hedge against criticisms that external observers (such as PhD examiners, journal/book reviewers or post-publication critics) might later make about 'biases' in the final set of interviewees accessed.

Of course, in interview-based work the numbers of interviews needed for a given piece of research to be soundly based are always *sui generis*, and may be quite small (say 20–30). However, in modern sociological work in the US funded by government grant-giving institutions like the National Institutes of Health (NIH), the average in recent years has crept up to nearly 90, with some studies covering over 300 interviews (Deterding and Waters 2021). Pre-registration is additionally important for sociologists and ethnographers because of the controversial concept of reaching 'saturation' in the number

of interviews. Deriving from grounded theory, it denotes the point where all viewpoints or perspectives about a given situation have been covered and additional interviews would not generate new insights (Tight 2024; Chitac 2022; Labobe, Morgan and Hoffman 2021; LaDonna, Artino and Balmer 2021). Rigorous research will be at or somewhat above this hard-to-decide-on level, but some justification in terms of it may be helpful for transparency.

Given that getting interviews at all is often a difficult job, it can be particularly useful to try and collate tips or advice at the planning stage in a blog or on social media before interviewing or field research gets under way in a big way. This might cover the main focus of the study questions and the rationale for contacting the anticipated set of interviewees. Inviting comments and suggestions from other researchers, and even from potential interviewees themselves or other actors active in the field being researched, can be helpful in combatting researchers' blind spots (there are always some) and in suggesting relevant factors to explore. Publishing a general outline of the work can also get an early fix on possible lines of criticism of the research.

A final step in making the best use of salient information gained from off the record interviews is to systematically try and 'stand up' any contested or controversial points made by interviewees. It is useful to cross-check with what other interviewees say, but they too may still be confidential. So the best assurance that an interviewee's comment is accurate comes from any supporting evidence in public documentation sources. Where different methods concur on the same account or explanation, the additional level of assurance gained can also be mentioned (in carefully general terms) within the main text of the article or book, without breaching an interview's confidentiality.

Still the best way to make knowledge more reproducible is to try and persuade interviewees as much as possible to move the status of conversations away from being off-the-record, and instead on to at least non-attributable status. Early career researchers may feel intimidated by elite interviewees (and may not even ask for non-attribution). But stressing non-attribution's advantages (in terms of your making a fully accurate record of what people say) can often shift their early insistence on being off the record. Using autorecorded and remote software interviews makes this easier still, and researchers can promise interviewees a copy of the recording and of the autotranscript at almost no extra cost in time or resources – although this step may also have risks. Assure any worried interviewees that if they wish to go off record temporarily on a particularly sensitive matter, this is easily accomplished – but remember to quickly resume recording when that topic is finished.

Opening up non-attributable interviews more

The precise wording of your non-attributability guarantee and planned anonymization safeguards should be notified to all interviewees in advance, along with the plans for how data will ultimately be deposited or stored. In addition to getting respondents' signed consent forms, it can also be helpful to record them on audio or video agreeing to non-attribution status at the start

of the interview. Completely respecting the agreed terms of the interview is a fundamental aspect of research ethics and should never be in doubt.

Many of the same pre-registration and archiving steps discussed above (including pre-registering questions and prompts to be used, and how interviewees were chosen or added) are equally applicable to quotable but anonymized interviews. Because quotes from interviews will be broadly attributed to types of respondents – for instance, 'a senior policy maker', or 'a senior executive involved' – it is wise to record 'cascade' links and who recommended whom. Monitoring potential biases or one-sidedness as interviewees accumulate can also help researchers to seek balancing views in good time. To maintain some diversity, it can help to ask respondents to suggest people who might take a different perspective from them.

A well-thought-out strategy for quoting people can also boost openness for content that is non-attributable. Many researchers use only short or mundane/unexceptional quotes (of fewer than 10 words or so), but they may not give readers much assurance that interviews took place, or that a good range of people were interviewed, because the quotes are too undistinctive, or so short as to give little sense of an interviewee's context or personality. Reproducibility is best enhanced by using medium to long quotations (of 10–50 words) that better convey interviewee's complex meanings and distinctive perspectives. Used cumulatively, longer quotations also give readers insights into the discourses and organizational culture in which actors operated. Where potentially breaching anonymization prevents this approach (as with small *N* datasets), using multiple short but distinctive quotations may be an alternative approach.

Other important gains in assurance and reproducibility can be made by designing the interview frame from the outset to help generate counts of information that yield quantitative data (see Section 9.4). Making creative use of text analysis software (such as N-Vivo or similar packages) to count interviewees' use of vocabulary, short phrases and fragmented memes can illuminate their discourse and organizational or social situations, but cannot be linked to particular individuals. For instance, instead of longer quote examples that demonstrate how a given concept operated in a firm or agency's organizational culture, show how frequently it is used across multiple interviewees, and list 'partner words' or phrasings that regularly accompany it.

Finally, wherever feasible, researchers can try to move interviews off non-attributability and onto the record (with exceptions for sensitive matters) – for example, by offering them an initial choice at interview. Alternatively, perhaps interviewees might be persuaded to go public with some mini-podcasts or videocasts taken from the non-sensitive parts of unattributable interviews. Academics sometimes fear that asking potential interviewees for too much disclosure will make them refuse to be interviewed at all (and sometimes with good reasons), or will be off-putting at interviews, or increase actors' reserve or self-censorship in conversation.

However, two key reasons why people agree to be interviewed in the first place are first to 'set the record straight', and/or to document their own involvements and achievements in personal or professional life. So where

talking about people's experiences is a motivation, then being quoted as an interviewee in an expert study, or speaking publicly at length in a podcast or videocast, may not be such a disincentive. Of course, there will remain many circumstances where non-attribution must be maintained as a key safeguard for interviewees' security and wellbeing. And sometimes in sensitive situations researchers may legitimately be more cautious than interviewees themselves about preserving non-attribution (see Chapter 13).

Openness for on the record interviews

In addition to taking the pre-registration and archiving steps discussed above, the main steps needed to maximize the openness of on the record interviews involve recording everything that is said (both as questions and as responses) in the most systematic, comprehensive and digitally archivable form possible, as in Figure 9.3. Consider:

- Why just audio record when you might as easily video record?
- Does the extra quality of using specialist recording equipment outweigh the difficulty or 'friction' involved in setting it up?
- Does the ease of recording for interviews in Teams, Zoom or other software actually make remote interviews preferable to in-person ones (Marzi 2021; Lobe, Morgan and Hoffman 2020)? For example, maybe you can cover *X* number of interviews in-person, but twice or three times that in Zoom?
- Are the Teams or Zoom autotranscription features sufficient or helpful for your needs? Their outputs normally require some editing and cleaning, but usually still offer huge time savings over transcribing by hand.

In addition, recollections of interviews quickly grow vaguer. So, researchers must always write up a full or note record of their subjective impressions of an interviewee's motives, meanings and reaction to questions as soon as possible. These often prove important for unpicking hidden clues or cues later in the analysis. These confidential records should be carefully anonymized and discreetly phrased – it may be possible for respondents to request data about them under data protection regulations. If anyone, such as another researcher (or an interviewee), raises any issues with your interpretations it can be useful to have contemporaneous notes to respond from. The confidential record can also be carefully edited to give public background notes that help analysis during the project, and can be deposited to aid any follow-on researchers.

Complete sets of on the record interviews should be deposited as a full discourse record in a permanent repository or archive with good documentation, and wherever possible with links to the depository's location in open access articles, books and the university repository. Some ethical issues may arise – for example, an interviewee wanting to withhold permission for on the record depositing; or if the research team fears that an interviewee has perhaps said

some things in too rash or unguarded a way that might backfire on them. Here it may be necessary for some interview material to be non-attributable.

To safeguard against such risks, remember that moving respondents towards podcasting or video-casting all interviews from the start is a win–win action, and in itself may also generate some immediate open commentary or debate that helps the research analysis. Interviewees can get their point into the public domain, in a direct way that they control; other relevant actors may chime in; and recording and depositing the discourse are both radically simplified, since the content is already out there.

9.4 Data from interviews – creating quantitative metrics

When interviews are quoted in research books or articles only a tiny percentage of the content from underlying interviews ever gets to be seen, as in museums that can only display a tiny proportion of the artefacts held in their storehouses. Readers naturally wonder about all the material that never gets cited – how far do the passages or text quoted occur in the data as a whole? What was the overall range of opinion amongst interviewees and how diverse or 'saturated' was the coverage? How accurate is the research author's own summary or synopsis of what got said about different topics? Using qualitative interviews to also generate quantitative data from discourse and text addresses these concerns by providing numerical summaries of salient responses across *all* the interviews conducted. There are three main approaches, two familiar ones with relatively limited potential, and the third less well-known method.

(a) *Coding short 'open text' text comments* from open-ended questions included in standard ('mass') quantitative surveys has been used for many years. Here the interviewers' survey forms allow them either to write down short, verbatim, free-text responses to an open-ended question, or respondents can add text to short write-in boxes with online or paper questionnaires. This is especially useful where respondents may have different phenomena or meanings in mind in responding to some questions, or as a 'sweeper' tactic to identify factors that researchers might have missed at the piloting and pre-coding of questions stage.

Illustrative qualitative quotes from quantitative surveys can give hints about what respondents were thinking about in replying to questions, And with large *N* data, they can sometimes shed valuable light on intriguing aspects – for example, how little respondents are aware of issues or concepts that elite political insiders dwell on; or how people may use some concepts that elites and researchers take for granted (like the left–right divide in politics or 'class' in social research). Normally, however, the amount of free text recorded is very small, prompted and enigmatic. And unless respondents wrote it themselves online or on a paper questionnaire, academic researchers may depend on survey firms' interviewers coding the 'open' oral responses accurately and in the same ways: so cross-interviewer reliability may not be high.

(b) *Counting concepts, vocabularies, memes or phrasings in extended interview texts.* With only small numbers of interviews (say, 20 or less) researchers

may find that listening to all their interviews and 'hand coding' people's complex views or orientations is feasible. Manual coding is defended as more responsive to individual variations, idiosyncrasies and emphases. But even in small *N* studies researchers could find it useful to check that their hand-coded judgements complement computer-coded ones in expected ways.

With any larger number of interviews using text analysis software becomes inevitable. Note that either *all whole interview transcripts* must be transcribed or digitized or *the same relevant sections* of all interviews (where researchers are using a 'running log' approach). For any qualitative research (not just interview data) there are three key 'moments' of coding – making codes, then organizing around them and making patterns from the coded data (Locke, Feldman and Golden-Biddle 2022; Reyes, Bogumil and Welch 2024). NVivo text analysis software moves inductively from respondents' own words to produce counts across whole interviews (or comparable sections) of how often participants used concepts or vocabularies, and with what 'partner words', especially phrases indicative of participants' organizational cultures, world views or personal ideology in some aspect. Alternatively, more 'thematic' analyses go in a more deductive or 'top down' way from researchers' analytic indexing of topics or codes about what any form of text signals to summary numbers (Jowsey, Deng and Weller 2021; Williams and Moser 2019). Flexible coding using qualitative data analysis software is recommended by some authors as a '21st century' sociological approach, especially for larger interview datasets (Deterding and Waters 2021; Locke, Feldman and Golden-Biddle 2022).

Critics argue that text analysis packages may only address the most 'surface' features of discourses, and thus only draw on a tiny fraction of the information and diverse meanings generated by qualitative interviews compared with intensive manual analysis (Mattimoe et al. 2021). However, they are useful in supplementing readers' assurance that researchers have considered relevant text across all the conversations held (or across particular questions asked), and are unavoidable when interview numbers grow much beyond 30 (for a given purpose). But over-relying on text analysis to sum up what interviewees said could create a rather impoverished qualitative analysis if not accompanied by enough 'deep reading' of the discourses.

(c) *Using a data-generation framework filled out by in-depth qualitative interviews* is a less well-known approach, but one with great potential for giving readers assurance that every aspect of what respondents have said has been considered before careful and checkable post-codings of what they said were made. Obviously the most reliable and checkable codes are the basically binary yes/no ones of the kind recommended for documentary analysis below (Chapter 11) – for example, counts of whether respondents did or did not use different concepts or phrases of interest in their replies. Text analysis software is good at such counting. However, binary codes rarely capture the richness of information that qualitative interviewing can generate.

So, instead it is best to think of a qualitative interview that can generate quantitative measures as a means of getting answers to a largish number of

carefully chosen core variable codings that are designed into how questions are framed from the outset. There might be 20–30 of these items at most and they might be scored along a spectrum of meanings with two end categories and one intermediate category, giving three boxes, as shown below.

On a spectrum of possible responses the interviewee's answer was:

Wholly or clearly at the left end of the spectrum	More in the middle than clearly towards either end	Wholly or clearly at the right end of the spectrum
Code = 1	Code = 3	Code = 5

The questions asked of respondents do not disclose the codes for the spectrum end categories, but are designed to elicit answers that allow the interviewer to make the coding. For instance, a study of hospitals in the UK and other countries (Bloom and van Reenan 2013; Bloom, Sadun and van Reenan 2014) used an interviewer to telephone one senior manager from each hospital and have an hour long conversation with them designed to generate codes on a range of items such as the following.

Physical organization of the hospital?

Basically unmodernized in terms of smooth patient throughflow (e.g. because of historic buildings altered in ad hoc ways)	Partially modernized and improved but with significant throughflow problems remaining	Fully conforms to modern design in terms of a smooth patient throughflow
Code = 1	Code = 3	Code = 5

Can a problematic member of the hospital staff be fired or moved?

A problematic staff member mostly cannot be removed nor moved or demoted, but must just be worked around (e.g., because of trade union or professional restrictions, or weak HR procedures)	A staff member whose behaviour causes problems or concerns can be moved to less salient roles, but can rarely be dismissed or demoted	HR procedures allow for easily moving or firing a staff member whose behaviour has become problematic
Code = 1	Code = 3	Code = 5

For each item the Bloom et al. (2013) interviewer had several minutes of open-ended conversation with the interviewee (which was recorded), seeking information about their hospital designed to firm up a post-coding judgement of the kind above on each item. Interviewers were asked to score how they perceived the reliability of respondents (which then became a variable in the quantitative data analysis, to check for any systemic biases). The overall coding across hospitals was then checked for consistency and reliability. In other applications, the same approach was used to gauge the management culture of firms and their handling of innovations, again using remote interviews with a single interviewer making the codings, but with checks (Bloom and van Reenan 2006, 2010; Bloom et al. 2016). This same approach has also been applied productively in public administration and development economics, sometimes with codings being made by two separate people, and then checked and reconciled (Williams and Yecalo-Tecle 2020; Rasul et al. 2020, Appendix A; Rasul et al. 2018).

Modern Zoom and Teams remote interviews strongly enhance the viability of these approaches (since videos of respondents are also available), and autorecording and transcriptions may greatly reduce their costs (Archibald et al. 2019; Falter et al. 2022; Gray et al. 2020; Saarija¨rvi, Maarkus and Bratt 2021). The Covid-19 pandemic emphasized their usefulness where face-to-face contact is infeasible (Tremblay et al. 2021). In this way it is possible to move strongly beyond the eclectic and disparate things true of each hospital or firm to nonetheless tap the most salient aspects of their managerial setup in a sophisticated quantitative way, based on a well-informed qualitative engagement with appropriately senior and relevant interviewees. Bloom et al. used sophisticated econometric analyses to understand the otherwise elusive issue of management productivity, regressing the variables derived as above against multiple other data characteristics of hospitals or firms (Bloom and van Reenan 2006, 2010, 2013; Bloom, Sadun and van Reenan 2014).

If the full underlying interview recordings or transcripts (for non-attributable interviews) are deposited alongside the codings made for each answer from a reasonable *N* of interviewees, then later researchers can cross-check them in detail. Assuming the coding decisions pass muster as correct, reusers of the dataset can have very strong assurance in the evidence presented and the original researchers' judgements made about it.

Conclusions

Pre-registering the maximum amounts of disclosable information about an interview-based study at its outset (especially the planned main question frame) is a key initial step. A narrative should also document how the interview process was planned to go and then actually went, how the selection of interviewees happened and changed, and how the researcher's understanding

of phenomena, behaviours, outcomes and meanings altered as interviews went on. Consider early on which parts of your research data can be deposited at the project' s end – in anonymized form for non-attributable interviews and in attributable full text for those on the record. Be brave in asking interviewees to be somewhat more open – for example, to allow their views to be deposited, and perhaps to go public in blog, podcast or other formats earlier on during research.

These steps can yield real benefits for advancing any given piece of research. Readers gain immediate extra assurance about the provenance of the work and accuracy of the analysis. Other researchers can check the accuracy and validity of your interpretation against a fuller context of what questions were asked and what interviewees said that can be disclosed (positive gains on the Turing Way *reproducibility* criterion). If they have documented interviews well, researchers can respond positively and quickly to other researchers' enquiries, requests for information and advice-seeking on questions to ask. Follow-on researchers can also apply different methods to analyse the interview database (e.g. using textual analysis to confirm a previous, purely qualitative judgement). If they can reach similar conclusions, the *reproducibility* of the analysis is confirmed (in Turing Way terms). *Reputational gains* should follow, because analyses supported by OSS evidence will be better regarded in the discipline and by outside readers.

If other researchers can learn from or use the same methods (e.g. initial questions, concepts/vocabularies and follow-on prompts) with other similar sets of interviewees, and obtain the same results, then *robustness* is strengthened. By varying the interview approaches or methods, across different sets of interviewees (e.g. across different spatial contexts and perhaps time periods), other researchers can find if only different results emerge, or if instead very similar results occur, enhancing *generalizability* (in Turing Way terms). Researchers across the field as a whole can make *learning gains from positive results*. Replicated results and greater generalizability can boost both sub-field consensus and learning. Understanding the reasons for non-replication may also boost *learning gains from negative results* – for example, firming up the reasons why an association of variable *X* with variable *Y* fails to hold in other sets of interviews.

Notes

* This chapter was developed from an original online seminar held on 1 December 2022, forming part of the LSE's input into the Civica project. The project was funded by the European Union's research and innovation programme Horizon 2020, under the chapter 'Science in Society'. The PowerPoints from the seminar and a video of the presentation and subsequent discussions are permanently free to download from the

Civica Zenodo pages at https://zenodo.org/records/7630777 We thank all the participants at this session for their questions, comments and inputs.

[1] However, there are some specialized contexts and occasions (like studying criminal behaviour or in war or disorder zones) where 'covert' research may be permissible. In other situations, unobtrusive observation of public behaviours (e.g. postings on social media, or patterns of foot traffic around a park or a gallery) and well-managed ethnographic observation may be legitimate without participants being aware of it. Ethical approval by a suitable review body is normally essential.

References

Ahler, Douglas J.; Roush, Carolyn E.; and Sood, Gaurav (2020) 'The micro-task market for lemons: data quality on Amazon's Mechanical Turk', *Political Science Research and Methods*, vol. 13, no 1, pp. 1–20. $ https://doi.org/10.1017/psrm.2021.57 OA at ResearchGate. https://perma.cc/8T9T-NFRQ

Archibald, Mandy M.; Ambagtsheer, Rachel C.; Casey, Mavourneen G.; and Lawless, Michael (2019) 'Using Zoom videoconferencing for qualitative data collection: perceptions and experiences of researchers and participants', *International Journal of Qualitative Methods*, vol. 18, pp. 1–9. https://doi.org/10.1177/1609406919874596

Beaunae, Catherine; Wu, Chiu-Hui; and Koro-Ljungberg, Mirka (2011), 'Exploring performativity and resistance in qualitative research interviews: a play in four acts', *Qualitative Inquiry*, vol. 17, no. 5, pp. 412–421. https://doi.org/10.1177/1077800411405166

Bloom, Nick and Van Reenen, John (2006) *Measuring and Explaining Management Practices Across Firms and Countries*, London: London School of Economics, CEP Discussion Paper No. 716. https://perma.cc/CBR5-B2Z6

Bloom, Nicholas and Van Reenen, John 2010 *New Approaches to Measuring Management and Firm Organization*, London: London School of Economics, CEP Discussion Paper dp0969. https://perma.cc/9ZAS-DP74

Bloom, Nicholas and Van Reenen, John (2013) 'Does Management Matter in Healthcare?', London School of Economics, online paper. https://perma.cc/GXU4-GEZN

Bloom, Nicholas; Sadun, Raffaella; and John Van Reenen, John (2014) 'Hospital Organization and Productivity', *Health Affairs*, vol. 33, no 5, and vol. 33,no. 7 (May and July), and 2013). $ https://www.nber.org/books-and-chapters/hospital-organization-and-productivity/does-management-matter-healthcare

Bloom, Nicholas; Sadun, Raffaella; and Van Reenen, John (2016) *Management as a Technology?*, London: London School of Economics, CEP Discussion Papers dp1433. https://perma.cc/VH9E-B9V7

Broockman, David and Kalla, Joshua. (2016) 'Durably reducing transphobia: a field experiment on door-to-door canvassing', *Science*, vol. 352, no. 6282, pp. 220–224. https://perma.cc/BP25-84X3

Burkette, Jay (2021) 'The research interview: a performative reinterpretation', *Qualitative Inquiry*, vol. 28, nos 3–4, pp. 300–311. https://doi.org/10.1177/10778004211051060

Chitac, Iuliana M. (2022) 'The rationale for saturation in qualitative research: when practice informs theory', *Cross-Cultural Management Journal,* vol. 24, no. 1. https://perma.cc/7H24-NZQM

Collins, Harry (2010) *Tacit and Explicit Knowledge*, Chicago, IL: Chicago University Press. **$** https://perma.cc/HXP9-FCMB

Deterding, Nicole M and Waters, Mary C. (2021) 'Flexible coding of in-depth interviews: a twenty-first-century approach', *Sociological Methods & Research*, vol. 50, no. 2, pp. 708–739. https://doi.org/10.1177/0049124118799377

Dexter, Lewis Alexander (2008) *Elite and Specialized Interviewing*, Brussels: ECPR Press. Originally published 1969. **$** https://perma.cc/3QS7-VS9R

Denzin, Norman K. (2001) 'The reflexive interview and a performative social science', *Qualitative Research*, vol. 1, no. 1, pp. 23–46. https://doi.org/10.1177/146879410100100102

Falter, Michelle M.; Arenas, Aaron A.; Maples, Gordon W.; Smith, Chelsea T et al. (2022) 'Making room for Zoom in focus group methods: opportunities and challenges for novice researchers (during and beyond COVID-19)', *Forum: Qualitative Social Research*, vol. 23, no. 1, article 21. https://doi.org/10.17169/fqs-23.1.3768

Gray, Lisa M.; Wong-Wylie, Gina; Rempel, Gina; and Cook, Karen. (2020) 'Expanding qualitative research interviewing strategies: Zoom video communications', *The Qualitative Report*, vol. 25, no. 5, article 9. http://doi.org/10.46743/2160-3715/2020.4212

Herron, Brigette A. (2023) '40 years of qualitative feminist interviewing: conceptual moments and cultivating ecosystems of care', *Qualitative Inquiry*, vol. 29, no. 6, pp. 659–669. https://doi.org/10.1177/10778004221139611

ICPSR Institute for Social Research (Michigan University) (2023) *Guide for Sharing Qualitative Data at ICPSR*, 3 October. Web page. https://perma.cc/MQ62-U7ZX

Jowsey, Tanisha; Deng, Carolyn; and Weller, Jennifer (2021) 'General-purpose thematic analysis: a useful qualitative method for anaesthesia research', *British Journal of Anaesthesia (BJA Education)*, vol. 21, no. 12, article 472e478 https://doi.org/10.1016/j.bjae.2021.07.006

Konnikova, Maria (2015), 'How a gay-marriage study went wrong', *New Yorker*, 22 May. https://perma.cc/YJ22-JTZ3

LaDonna, Kori A.; Artino, Anthony R.; and Balmer, Doreen F. (2021) 'Beyond the guise of saturation: rigor and qualitative interview data', *Journal of Graduate Medical Education*, 15 October, vol. 13, no. 5, pp. 607–611. http://dx.doi.org/10.4300/JGME-D-21-00752.1

Lobe, Bojana; Morgan, David; and Hoffman, Kim A. (2020) 'Qualitative data collection in an era of social distancing', *International Journal of Qualitative Methods*, vol. 19, pp. 1–9. https://doi.org/10.1177/1609406920937875

Locke, Karen; Feldman Martha; and Golden-Biddle, Karen (2022) 'Coding practices and iterativity: beyond templates for analyzing qualitative data', *Organizational Research Methods*, vol. 25, no. 2, pp. 262–284. https://doi.org/10.1177\1094428120948600

Marzi, Sonja (2021) 'Participatory video from a distance: co-producing knowledge during the COVID-19 pandemic using smartphones', *Qualitative Research*, pp. 1–17. https://doi.org/10.1177/14687941211038171

Mattimoe, Ruth B.; Hayden, Michael T.; Murphy, Brid; and Ballantine, Joan (2021) 'Approaches to analysis of qualitative research data: a reflection on the manual and technological approaches', *Accounting, Finance, and Governance Review*, vol. 27, no. 1. https://doi.org/10.52399/001c.22026

Oakley, Anne (1981) 'Interviewing women: a contradiction of terms?' In Roberts, H. (ed.), *Doing Feminist Research,* London: Routledge, pp. 30–61.

Oakley, Anne (2016) 'Interviewing women again: power, time and the gift', *Sociology*, vol. 50, no. 1, pp. 195–213. $ https://doi.org/10.1177/0038038515580253

Rasul, Imran; Rogger, Daniel; and Williams, Martin J. (2020) 'Management, organizational performance, and task clarity: evidence from Ghana's civil service', University College, London, website. Online Paper, August. https://perma.cc/MA2X-SPT8

Rasul, Imran; Rogger, Daniel; and Williams, Martin J. (2018) 'Management and Bureaucratic Effectiveness Evidence from the Ghanaian Civil Service', World Bank, Policy Research Working Paper 8595, Development

Economics Development Research Group, September 2018 WPS8595. https://perma.cc/87PT-C2PE

Reyes, Victoria; Bogumil, Elizabeth; and Welch, Levin E. (2024) The living codebook: documenting the process of qualitative data analysis, *Sociological Methods & Research*, vol. 53, no. 1, pp. 89–120. https://doi.org/10.1177/0049124120986185

Saarija¨rvi, Maarkus and Bratt, Ewa-Lena (2021) 'When face-to-face interviews are not possible: tips and tricks for video, telephone, online chat, and email interviews in qualitative research', *European Journal of Cardiovascular Nursing*, vol. 20, pp. 392–396. https://doi.org/10.1093/eurjcn/zvab038

Tight, Macolm. (2024) 'Saturation: an overworked and misunderstood concept?', *Qualitative Inquiry*, vol. 30, pp. 577–583. https://doi.org/10.1177/10778004231183948

Tremblay, Stephanie, Castiglione, Sonia; Audet, Li-Anne; Desmarais, Michele; Horace, Minnie; and Pelaez, Sandra (2021) 'Conducting qualitative research to respond to COVID-19 challenges: reflections for the present and beyond', *International Journal of Qualitative Methods*, vol. 20, pp. 1–9. https://doi.org/10.1177/16094069211009679

Williams, Martin J and Yecalo-Tecle, Liah (2020) 'Innovation, voice, and hierarchy in the public sector: evidence from Ghana's civil service', *Governance*, vol. 333, no. 4, pp. 789–907. An earlier OA version (from 2019) is here: https://perma.cc/B2GU-JSB5

Williams, Michael and Moser, Tami (2019) 'The art of coding and thematic exploration in qualitative research', International Management Review, vol. 15, no. 1, pp. 45–55. https://perma.cc/48CG-6955

Willis, Rebecca; Yuille, Andy; Bryant, Peter; McLaren, Duncan; and Markusson, Nils (2021) 'Taking deliberative research online: lessons from four case studies', *Qualitative Research*, vol. 23, no. 4, pp. 921–939. https://doi.org/10.1177/14687941211063483

10. Case studies and fieldwork

In creating and analysing a case study researchers seek to separate out and examine intensively a set of multiple causal processes made more tractable for analysis because they are situated in a specific spatial location, social sphere or time period. They want to understand how the causes and phenomena studied interact with each other, and what they meant to the people involved. Many case studies are essentially conducted using either interviews (covered in Chapter 9) or systematic documentation (covered in Chapter 11) to gather evidence. Field research similarly eschews the isolated, artificial experimental testing feasible only in laboratories, instead examining the multiway interaction of causal processes in complex 'real life' settings (that are also carefully bounded to allow intensive study).*

We begin by briefly reviewing the main rationales for case studies that explain their continuing important role in empirical social science (Section 10.1). Three solely qualitative case study formats are also reviewed – in-depth 'diagnostic' studies, shorter 'apt illustrations' and 'intermediate' cases lying between the other two. These 'classic' approaches are the hardest to adapt for open social science (OSS) ways of working (Field et al. 2021). But using pre-registration, self-conscious reflection and good documentation can again help researchers enhance the reproducibility of their work. In recent decades three more quantitively orientated uses of case research have developed strongly – mixed methods approaches, 'fuzzy set' qualitative comparative analysis (QCA) studies with small numbers of cases, and medium *N* studies using QCA methods. All three are somewhat easier to 'open up' in ways that boost readers' assurance in the research and its reproducibility (Section 10.2). Transparency for case research has also been increased by the growth of systematic review and other meta-methods for synthesizing results across very large numbers of studies, even where there are variations in methods or uneven quality of research (Section 10.3). The last section briefly considers other forms of field research where similar lessons about enhancing openness seem feasible (Section 10.4).

How to cite this book chapter:

Dunleavy, Patrick and Monteath, Timothy (2026) 'Case studies and fieldwork', in Dunleavy, Patrick and Monteath, Timothy (2026) *Doing Open Social Science: A Guide for Researchers.* London: LSE Press, pp. 225–254. https://doi.org/10.31389/lsepress.dos.j

10.1 Case study rationales and three classic types

A case study explains how multiple interacting social factors and processes worked in a holistic way within a specific time, geographical space or social situation, when dozens or hundreds of variables were in play at once, yet within a manageable analytic frame (the case).

> According to Ragin (1992, pp. 217–218) 'a case is a spatially and temporally defined entity created by researchers via a process referred to as "casing" … which is "a research tactic" whereby researchers "concoct" cases as a way to manage complexity. Researchers do not so much find cases as they define, "delimit" and "declare" them' (Sandelowski 2011, p. 154).

So 'casing' (defining case studies for analysis) is an instrumental activity, shaped by the researcher's purposes and the need to maintain what Sandelowski (2011) terms 'empirical intimacy' with the case – that is, an ability to analyse relevant phenomena in a reasonably deep and holistic way. Case research seeks to surface, uncover and make visible multiple factors operating in a specific social situation, and to illuminate their relative salience. It may be able to distinguish necessary causes (pre-conditions for X to happen, but insufficient to cause X alone) and sufficient causes (whose occurrence alone ensures that X occurs), and multiple possible combinations of causal factors. Case research can also be used:

- to compare across a small number of social situations while holding some factors constant (using matched cases);
- to capture the maximum range of variation in the operations of a well-defined process (using limiting cases); and
- to complement insightful but limited understandings of social processes and meanings obtained using large *N* quantitative methods. Case studies can instance in a concrete fashion the mechanisms by which causal processes operate under specific types of conditions.

Case study research is tractable because researchers seek to demarcate only a deliberately limited terrain, aspect or time period of study, chosen on theoretically informed grounds. And because case research is fairly easy to set up, it has historically played a vital role in exploratory research and exploratory data analysis methods (see Chapter 4) – as where researchers are:

- creating a rich and detailed, holistic and intuitively meaningful account or narrative of how a defined set of social processes work and change;
- developing sense-making explanations of societal conditions that are distinctively new or were previously unstudied from within a given theory perspective;

- describing and first-analysing rapidly developing histories and changing local social situations;
- building out new knowledge and theory development in ways that require a high level of 'grounding' in the meanings and understandings of actors;
- identifying, surfacing and then process-tracing complex causal mechanisms observed or tracked in action within a given setting;
- generating preliminary research narratives that can inform more systemic quantitative or qualitative studies conducted in exploratory ways; and
- providing researchers designing confirmatory research with essential underpinning for large *N* confirmatory and quantitative studies.

In many of these circumstances case research cannot be easily matched by other approaches.

Every case study and piece of field research is *sui generis*, specific to the spatial location, social setting and time period when it is undertaken – and herein lies one central problem of opening up the approach. As with interviews in Chapter 9, no exact reproduction of case study research is usually feasible, since either time has passed from an original study and/or the processes studied may be similar to previous research but in a different location. This core non-reproducibility can often be compounded by the conditions under which case research must be conducted, including having to anonymize cases as a condition for access, and the inherent 'atypicality' of any given setting for other settings.

In the inter-war social sciences some U.S. researchers searched for isolated cities and 'communities' to stand in for the whole of U.S. society as 'representative' locales for studying social power or culture with the pre-behavioural sociological methods of that era. In fact, they only studied free-standing cities or small town settings that were distinctive in being isolated – but in that respect also completely 'unrepresentative' of communities in the east coast U.S. megalopolis or major Californian or southern U.S. metropolitan areas. Since that period, researchers have occasionally sought for ways of describing in microcosm sets of social processes that are otherwise intractable to close analysis (or too macro-scaled). But the ill-fated search for 'isolated' cases or 'typical' cases has retreated into social science history. Situating case studies against a wider background of variations is always important and worthwhile, but it never qualifies their uniqueness. So how can scientists have greater confidence that case study and field research insights are well grounded in replicable and reliable methods, so as to derive the maximum value from them?

One somewhat countervailing response has had some effect in enhancing reproducibility, namely a reliance in some disciplines on 'model cases' or examples that have come to partly define 'the canon' there (Krause 2021; Linke 2022). Krause argues that one or a few especially salient case studies can become 'privileged material objects' within a discipline (in a similar way to the differential focus on a few organisms like the fruit fly in biology or genetics).

These privileged material objects become multiply cited or contested and reinterpreted, sometimes over and over again, in ways that foster (good and bad) forms of knowledge cumulation in social research. For instance, examples or analysis from western countries are still far more salient in social sciences than those undertaken elsewhere (Krause 2016). The central or 'model' cases (e.g. the French revolution in the study of revolutions) may work to boost consensus. Or they may be contested in depth by critics partly because they are taken to have wider implications for the field of study. Yet because these cases are particularly reused by other researchers, some greater cumulation of knowledge may occur and the transparency of the methods and analysis used by the original authors of key cases are explored with atypical depth and thoroughness.

In modern social science, some of the limitation of case studies can be controlled or mitigated by 'triangulating' their results with other kinds of evidence (Rothbauer 2008; Olsen 2004), especially:

- quantitative data, gathered from large *N* survey interviews;
- quantitative analyses of behavioural data in the field or on social media; and
- systematic documentation analysis of written, official or other sources.

The limitations of each approach are here offset by cross-checking with different sources, researched in different ways. For instance, the nuanced views and social meanings of research participants in one case study can be set against more basic survey data or large-scale behavioural information to gauge how commonly held such views might be. And documentary sources may either back up case interviewees' accounts of how events happened, or suggest questions about them. Triangulation often also means that analysts encounter several different epistemological and philosophical positions (Olsen 2004). Greater methodological pluralism strengthens the diversity of explanations and phenomena considered and should boost reproducibility.

There have been three main types of solely qualitative case studies, shown in Figure 10.1, and we discuss them in the order shown.

Figure 10.1: The three best known forms of case study

Type of case method	One or several case(s) only, deep holistic analysis, grounded theory, etc.	Few cases (3–10), longer with some sustained analysis	Multiple (10+), short cases, brief summary account
Solely qualitative interpretation	**1. Diagnostic case** shows unique causal pathway in complex, multicausal situations. Strong process-tracing and mechanism-demonstrating	**3. Intermediate cases** showing a variety of holistic interactions in narrative or 'guidebook' form	**2. 'Apt illustration' cases** giving brief narratives or exemplification of diverse processes across contexts or situations

Source: Authors.

Diagnostic cases

The most celebrated (and perhaps most mythologized) form of case is a diagnostic study where the analysis of a single example (or perhaps two or three examples) is so inclusive and compelling that the processes involved can be exhaustively understood – somewhat like a postmortem analysis conducted in pathology, where an account of causal processes can be tracked in enough detail that each step appears necessary, and the sequence cumulates in a tightly observed manner without unobserved factors. Similarly, social scientist case researchers argue that looking in depth at a single instance can nonetheless identify causal factors in ways that are logically compelling: 'A case study is an intensive study of a single unit with an aim to generalize across a larger set of (similar) units' (Gerring 2004). Whereas any statistical association (taken on its own) remains a surface correlation, a deep case generates an exhaustive understanding of the analysed social phenomena, one that some theorists and historians have argued implies a 'covering law' generalization about all other (sufficiently similar) cases (Little 2020, Section 3.1).

A wide range of social scientists and most historians have used different deep case approaches, but all of them are well theorized and 'grounded' in a close reading of empirical phenomena considered holistically, normally also drawing out the social meanings and understandings of the actors involved themselves. Given that researchers have to 'immerse' themselves in the empirics being studied, it is common for a journal article to cover only a single 'depth' case, while research books or long-form PhDs might include two, three or four.

From an open science (OS) perspective, in-depth case studies raise some key worries. Extended analyses take a long time to do, and so are hard to redo. Efforts to counter-interpret evidence cited may be feasible in secondary analyses, but often cannot be based on re-examining the primary materials accessed by the original researchers. These authors are also likely to argue that interpretations 'ungrounded' in the original theory have limited relevance and can only be based on superficial reinterpretations. The more esoteric the methodology and analysis approaches taken by the original researchers, the more likely it is that they will be inaccessible to non-cognoscenti. In addition, because all casing is instrumentalist, with studies defined to sustain key purposes, reproducibility may be limited by researchers' having (perhaps subconsciously) chosen cases that lend themselves to their preferred interpretation. Finally, critics often complain that the original authors left unclear what level of analysis warrants claims that a single case (or a small number of cases) can identify either necessary or sufficient causes of social phenomena.

What potential remedies can enhance the openness of deep case studies then? Again authors pre-registering hypotheses at the outset of case work is a valuable step, setting out the rationale for why casing has been done in a particular way, then the reasons for the case choices made, and also the initial expectations that they think they are likely to uncover. Pre-dated in a fixed way, such a document can provide researchers with valuable evidence that

their case analysis has not been 'retrofitted' to accord with their theoretical preconceptions or emergent findings. A 'living' version of this document can then give a brief narrative of the developing analysis, showing both how their chosen perspective crystallized, and how they explored and evaluated alternative views as the research deepened.

Can later researchers test or partly redo elements of in-depth cases? Researchers can enhance the openness of in-depth case studies by depositing all the materials that can be shared in an online repository. Because case analysis stresses holistic analysis, the original authors always need to show that all or most potential influences were considered. Given a lot of in-depth information on the case, critics or reviewers with even limited access to case information themselves should be able to detect fairly easily when other potentially relevant causes have been ignored or neglected. Many in-depth cases rely heavily on interviews as primary source texts, and so all the restrictions on openness and the safeguarding of interviewees' confidentiality discussed in Chapter 9 also apply here. The recommendations suggested there may mean that even where cases rely heavily on non-attributable or off the record interview data, some useful disclosures can still be made, even if in summary form. Depositing the full transcripts or recordings of interviews is the most useful step for improving the reproducibility of case materials.

In addition, however, social science case studies often also depend on systematic documentation analyses (discussed in Chapter 11). Here original researchers can enhance the reproducibility of in-depth case studies by depositing for other researchers' inspection any collateral documentary information gathered in case building. Most useful would be non-copyright full text sources that allow useful inspection checks. A judgement call is inevitably involved here since researchers can rarely provide more than a (curated) selection of documentation to back up their case analysis. And in some situations, case researchers (like interviewers) may have only conditional, confidential or non-disclosure agreement (NDA) forms of access to documentation from a firm or agency studied. Finally, in-depth case studies are often lengthy and so word limits are often under pressure from journals or book publishers, making it vital for transparency that researchers accessibly explain where they have employed specialist or distinctive analysis methods in a journal replication archive or a book/thesis research methods annex (RMA).

Providing collateral evidence deposits to back up in-depth cases is hardest where interviewee- or documentation-access constraints compel the original researchers to anonymize the whole case under a pseudonym. Not being able to identify the case location, time period or other boundaries naturally raises the greatest concerns about the research involved being non-reproducible – but it is quite a common issue with research books. It is even more prevalent at PhD level, but here the apparatus of university supervision should normally provide 'under way' assurance that the work claimed to have been done was actually undertaken. The various forms of independent PhD examination processes across Anglosphere and most European countries also provide safeguards. There are

strong reasons for PhDers to make very careful case selections that can be well justified. Where article or book authors rely on anonymized in-depth cases the need for them to pre-register and to deposit as much as feasible is at its greatest. They should also be sure to briefly but explicitly evaluate countervailing explanations to the one they prefer – in particular, identifying critical 'gates' (e.g. problems, events, choices) that were used in filtering out rival accounts.

Apt illustration cases

In the rightmost column of Figure 10.1 are multiple short case studies, where only limited research is undertaken for each one, and each narrative given is not meant to accomplish a detailed process trace but only to give an appropriate illustration of the diversity of processes in a given area, time period or sphere of social life. The cases serve to demonstrate heterogeneity via examples, but not to systematically document or analyse the causal processes behind outcomes. Case selections may also need only briefer justification.

Here some different kinds of OS worries arise. Given that the range of exemplifications is important, was the case selection 'stacked' in a way that supports the researchers' preferred interpretation? Since analysis of multiple cases is also necessarily quicker and shallower, how was the restricted analysis included undertaken? Methods descriptions for short cases are often sketchier or inexplicit, unless closely similar methods were applied to all cases and can be described synoptically. And do the case narratives suggest that some plausibly important variables or situation dimensions may have been omitted, or just not spotted once work was under way, potentially leading to the misascribing of causal influences?

In general, however, apt illustration cases pose far fewer openness problems for scientific reproducibility. Pre-registering key overarching hypotheses or expectations should normally be feasible, depending on the *N* and types of cases. The analysis methods in short cases are also typically simpler and more descriptive, and can often be grasped without first investing heavily in understanding the original researcher's specific preferred theory. Follow-on researchers can more easily rerun the original analysis of any problematic or contested case by using publicly available information, less intensive checks and conventional secondary analysis methods. Even ordinary readers may be able to assess the intuitive plausibility of a non-anonymized case, and to look up collaterally relevant materials that can help them to critically assess a briefly sketched case study. Suggesting counter-explanations or pointing to potentially relevant cases left out is normally relatively easy for researchers in the same field, or professional actors in the relevant sub-field of business, government or civil society.

These factors normally make it easier for critics to understand the original interpretation of short cases, to construct counter-explanations and often to dispute the case selection made. But this change also makes it even more useful to pre-register a cases rationale and initial expectations (perhaps hypotheses) relating to them before detailed work is undertaken, again with later

tracking for changes in interpretation, unexpected developments or and explicit alterations made in the process for the cases included. It also places a premium on justifying case selection in something of the same matter needed for intermediate cases, to which we now turn.

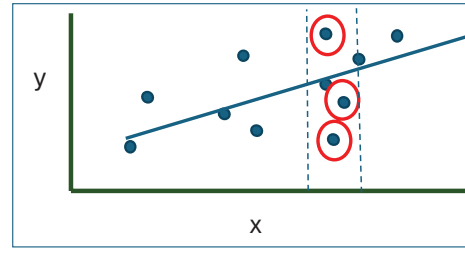
Intermediate comparative cases

The middle column in Figure 10.1 shows that three to six medium-length case studies are often used for comparisons. Each case is studied in more depth than apt illustrations but lesser depth than diagnostic cases. A small number of intermediate cases may be studied in enough depth to realistically show holistic interactions (especially with matched cases, discussed below). The selection criteria for cases are usually theoretically based, and informed by an overall view of the empirical distribution of time or space situations in the relevant research field.

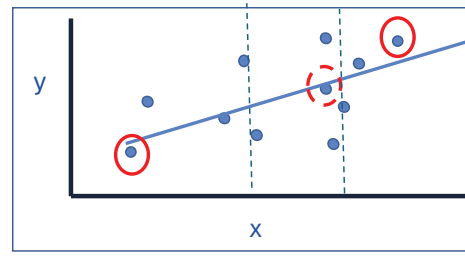
Carefully comparing across cases means that more attention must be paid to formulating a clear rationale for the case selection, and understanding the overall distribution or landscape within which their potential cases sit (Figure 10.2). In a 'matched cases' strategy researchers seek out cases where many likely explanatory variables are as similar as possible, for example because characteristics and processes are shared or operate in similar ways.

Figure 10.2: Selecting cases against a known population distribution – matched cases and limiting cases

(a) Matched cases



(b) Limiting cases



Source: Authors.

Matched cases are explicitly located in (roughly) the same place along the main horizontal causal axis in Figure 10.2a.

For instance, analysing how 'core executive' governance operates in four Anglosphere countries (e.g. the UK, Canada, Australia and New Zealand) already controls not just for numerous, historical, social and cultural similarities, but also for continuing links between the connected nations and for multiple 'Westminster system' institutional and party system features that the three countries have in common. Sometimes, too, intermediate studies can be used not for comparisons but to achieve a spread of conditions more suitable for model building. The hope then would be that any resulting differences in executive governance across the countries might be accurately fine-traced to the remaining, non-controlled differences between countries – such as the effects of federalism in two countries but not in two others, or differences in the sizes or geostrategic settings of the four states.

Sometimes, though, matched cases may be used to simply broaden the evidence base for model building. For instance, researchers used in-depth interviews with multiple healthcare providers across four qualitative case study areas in Accra to build a wider analysis of how a third of urban poor women in Ghana covertly used contraception (despite many risks from familial and community pressures not to do so), particularly highlighting the salience of providers maintaining clients' confidentiality (Kushitor et al. 2022). Closely matching cases may allow such complex causal problems to be better decoded.

By contrast, a 'limiting cases' approach studies examples from different ends of the relevant population distribution, plus (often) a carefully chosen mid-distribution case (or cases), so as to span the full available scope of variation (Figure 10.2b). The effort here goes beyond just narratively interesting (perhaps even random) exemplification, to instead achieve a form of calibrated differentiation that hopefully makes the influence of key variables or dimensions governing the distribution of cases maximally evident (so reducing complexity). In between the end-limit examples, one or more 'on the regression trendline' cases may also help the analysis further, perhaps showing intermediate effects. The aim is a case selection where comparative insights do a lot of work, allowing informed judgements about the extent to which outcomes differences across cases reflect (in part) their location within the main and subsidiary distributions.

In terms of openness issues, intermediate cases lie somewhere between in-depth cases and apt illustrations. Thanks to the more-detailed background knowledge needed to pick a well-justified case selection, researchers can usually achieve enough detailing and holistic 'fitting' to each case studied. They may often be able to show that this approach also generates insights that outperform both in-depth analyses of a single case on one hand, and the kind of 'surface correlations' obtainable from quantitative studies on the other. However, potential critics may still worry either that the methodology for comparison is not well worked up, or that the case selection perhaps 'oversupports' the researchers' preferred view. Covering multiple cases might

make researchers overconfident in attributing causation, when their necessarily fairly limited case analyses may easily miss out some causal influences or misascribe the influence of others. Every comparative case approach also confronts a potential *reductio* argument – that each of several cases is *sui generis* in its own distinctive way, so that when looking across cases no viable 'covering law' can emerge.

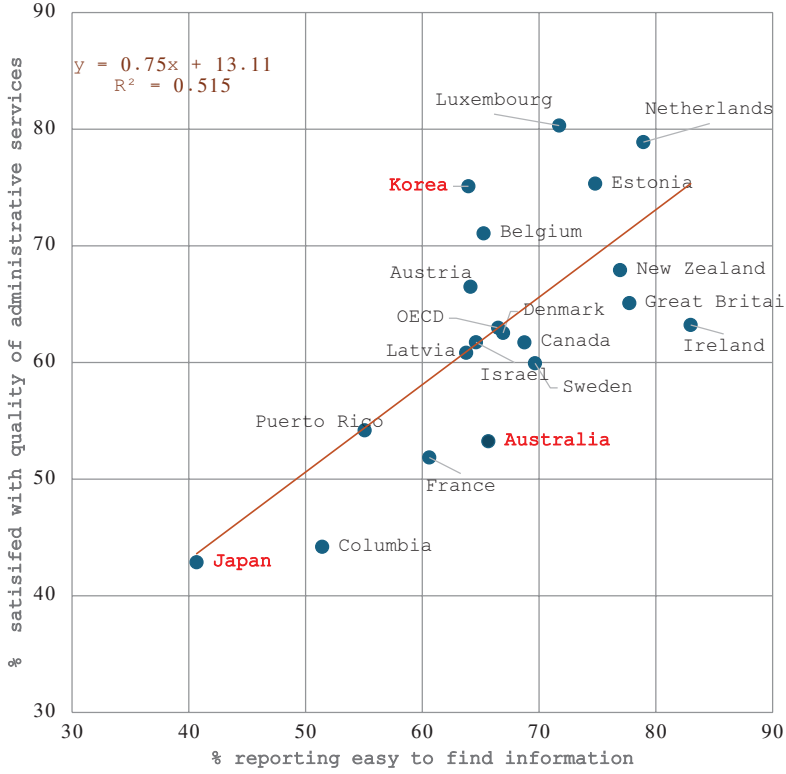
Again pre-registration can offer researchers an important defence against potential criticisms of the case selection criteria used and show that they were adhered to throughout the project, or altered only in defensible ways. Expert readers or researchers experienced with a particular case may also be able to assess intermediate cases relatively easily, for example in checking whether the original research excluded relevant variables. However, criticisms made by scholars of a single case are frequently rebutted as 'special pleading' or overly myopic arguments by comparative researchers. There may be good reasons for this stance also, unless a very *sui generis* factor has been linked to the causal phenomena being explained in a wider body of research. General (non-expert) readers may find checking the reproducibility of the analysis in comparative cases too difficult or costly in terms of time or effort. Researchers can help openness here by systematically depositing the case selection materials alongside their case analyses, and including in deposited data the fullest feasible documentation of the case comparison data and any associated literature reviews.

10.2 Case studies with quantitative aspects

Beyond pre-registration and documenting self-conscious case selection, one approach to increasing the openness of case studies involves bringing them into closer association with 'large number analysis' (Lieberman 2005). Here the links between more open and checkable quantitative data and potentially less open case materials can enhance assurance in the qualitative research undertaken. The recent trend in the social sciences towards 'methodological pragmatism' or mixed-methods research design has increased the impetus here (Olsen 2022). It stresses using the right tool for different research analysis tasks (whether quantitative or qualitative), and combining both approaches in complementary ways, either within single major studies or in the planned broader overall development of a research field.

Even small steps towards the quantitative contextualization of cases can make a difference to how open and reproducible or easily contested cases are. Justifying the choice of even two or three intermediate cases within a wider quantitative picture can help in showing what the relevant field of research looks like overall. For instance, Figure 10.3 shows a contextualization of a study looking at the digital development of government IT systems in three countries (Japan, the Republic of Korea and Australia), taken from an OECD 2021 survey of how citizens in advanced liberal democracies view public administration information and service quality. Australia and the Republic

Figure 10.3: Example – situating three case study countries against a wider set of comparable advanced liberal democracies



Note: A study of government IT development in three countries (Australia, Japan and the Republic of Korea, shown in red above) situates them against other liberal democracies, and against each other, using quantitative data from a contemporaneous OECD survey of public opinion reactions to standard questions across its member states. Japan respondents scored their country poorly on both aspects asked about, how easy it was to find government information online, and how satisfied they were with the quality of administrative services. Australian respondents reported good (but not world-leading) access to information online, but very low satisfaction with the quality of administrative services nonetheless. On the same level of good access to information, Korean respondents were far more satisfied with services (+20% compared to Australia).
Abbreviations for comparator countries here: AUT, Austria; BEL, Belgium; CAN, Canada; DNK, Denmark; FRA, France; GBR, Great Britain; IRL, Ireland; ISL, Iceland; LUX, Luxemburg; LVA, Latvia; NLD, Netherlands; NZL, New Zealand; SWE, Sweden.
Source: Computed from OECD (2022). We are very grateful to Hajime Isozaki for the opportunity to discuss this data with him.

of Korea might be selected as offering insights here because they have similar good but not world-leading scores on how citizens perceive the ease of availability of government information online, but quite different levels of citizen satisfaction with government. Japan might be chosen here because its score

Figure 10.4: Three case approaches (4, 5 and 6) with quantitative elements (compared to classic forms)

Type of case method	One or several case(s) only, deep holistic analysis, grounded theory, etc.	Few cases (3–10), longer with some sustained analysis	Multiple (10+), short cases, brief summary account
Solely qualitative interpretation	**1. Diagnostic case** shows unique causal pathway in complex, multicausal situations. Strong process-tracing and mechanism-demonstrating	**3. Intermediate cases** showing a variety of holistic interactions in narrative or 'guidebook' form	**2. 'Apt illustration' cases** giving brief narratives or exemplification of diverse processes across contexts or situations
Some categorization or comparison of cases		**5.** Small *N* study using **'fuzzy' QCA/set theory** categories (e.g. 5 or 7 point set inclusion scale)	**4. 'Mixed methods'** use of case studies to closely support quantitative analysis
Cases analysed in quantitative ways, as well as qualitative description			**6. Medium *N* study** using QCA/set theory software – numeric QCA scaling

Source: Authors.

here on both indices is amongst the worst among liberal democracies (despite sharing many cultural, scientific, industrial and institutional features with Korea). Including other (matched, rich, liberal democratic) countries in the chart also contextualizes the three countries as a set, suggesting other potential insights into how they compare.

Some authors link greater quantitative contextualization to simply doing more cases, as recommended by King, Keohane and Verba (1994) to qualitative researchers. In addition, more developed strategies are feasible for interrelating case studies with large *N* quantitative analysis. Three promising ways of increasing the openness of case research are shown in the two bottom rows of Figure 10.4, which also shows how they differ from classic cases studies.

Mixed methods studies

In mixed method studies the quantitative categorization and comparison of cases in the intermediate cases form is taken further. In particular, nested analysis seeks to situate case studies within overall study designs that are primarily shaped by quantitative analyses. For Lieberman (2005):

> there is a synergistic value to a nested research design – for example statistical analyses can guide case-selection for in-depth research, provide direction for more focused case studies and comparisons, and be used to provide additional tests of hypotheses generated from small-N research. Small-N analysis can be used to assess the plausibility of observed statistics relationships between variables, to

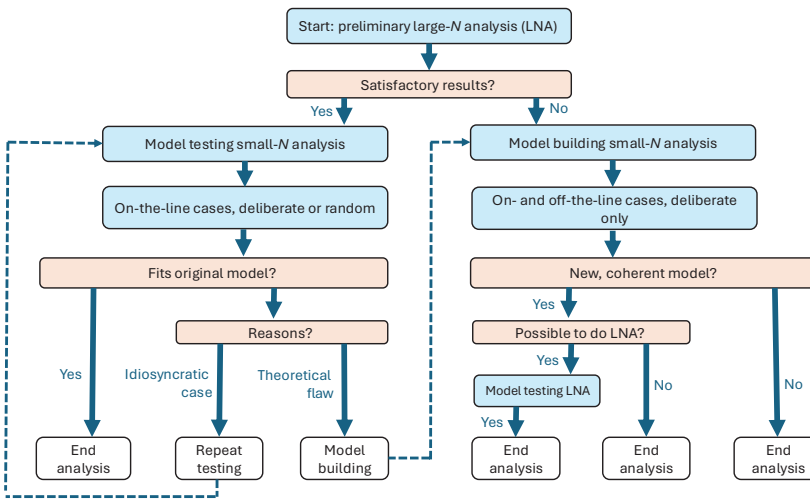
> generate theoretical insights from outlier and other cases, and to develop better measurement strategies (p. 435).

Figure 10.5 shows a reversioned picture of his 'nested analysis' model, which is relatively elaborate, and so only the largest and well-funded projects are likely to fully follow this design.

For most case researchers, most of the time, their task will likely be to situate case studies in a more integrated way with already known quantitative results accessed via secondary analyses of already existing primary research. However, well-funded studies may perhaps also include one original piece of large number analysis. Open science advocates argue that this is potentially fertile ground for boosting transparency because many qualitative researchers can be so embedded in their research sub-field or so committed to their existing detailed knowledge that the background reasons for case selection can be assumed or glossed over. In such circumstances, setting new detailed case knowledge within a properly described, known background can increase transparency and reproducibility.

Most mixed methods studies select cases in data-informed ways, situating cases explicitly against pre-specified and registered 'spread' criteria. Critics or reusers of the original research can thus assess the effects of substituting alternative (especially neighbouring) cases. Cases are quantitatively described in multiple ways, and they do not stand alone, nor just by comparison with each other. Instead they derive a good part of their force from their interactive

Figure 10.5: An ambitious 'nested analysis' model, after Lieberman (2005)



Source: Adapted with permission by the authors from Lieberman (2005).

relationship with large *N* studies. Some authors have even recommended radical strategies for 'casing', such as using random selection to choose intermediate-style cases. Some cases may be chosen as 'on the line' (close to the trend line from a regression analysis) – that is, they are well-behaved instances spread across the distribution where the implied quantitative model from the large number analysis works well. Other cases may be 'off the line' from a regression analysis – that is, instances where the implied model from the large number analysis does not work well and something extra makes them conform poorly with the variables that shape the bulk of observations.

In mixed method studies process-tracing or demonstrating mechanisms consistent with quantitatively identified possible causal links is often the key rationale for undertaking case studies. They help to:

- drive sense-making by analysts amidst very complex, multiway interactions;
- 'surface' a diversity of precise and observable mechanisms that may be involved in transmitting causal influences from one part of a social situation, time period or spatial system to another;
- demonstrate that a given mechanism operates in the manner needed to attribute causal influence to it; and
- (hopefully) confirm that the scale and efficacy of a given transmission mechanism is commensurate with the effects attributed to it.

Process tracing is especially key where quantitative analysis suggests complex interaction effects involving several variables. To be able to plausibly link such statistical effects to a relevant mechanism charted in detail is a key step in model development. Similarly, showing that changes in the timings of quantitatively identified effects fit with an empirically ascertained sequence of operations and exchanges has considerable value.

Mixed methods case studies pose fewer problems for openness than the classic methods above. Researchers are more likely to deposit case materials in a way as similar as feasible to how they deposit quantitative data where they can, especially by pre-registering and later depositing in a repository their 'frame' designs and contextual data searches undertaken when choosing cases, and their initial expectations for qualitative research. Adding a narrative of how the case research and its relation to the quantitative work developed over time boosts transparency further. These elements can greatly strengthen the provenance of the work, for instance in clarifying that the case selection was legitimately based and not 'rigged' to produce given results.

Of course, interview-based cases will still be subject to the openness issues considered in Chapter 9, and cases based on documentation analysis to the questions tackled in Chapter 11. Apart from these elements, the chief openness issue will be whether readers can easily check that the researchers have made a legitimately accurate and nuanced interpretation of each case. Such assessments are much easier with extensive contextualization and an

explicit linkaging between case analyses and well-aligned and supportive large *N* findings.

Fuzzy set (QCA) cases

An important effort to tackle the remaining interpretative uncertainty around case analysis in a much more open and transparent way is made by fuzzy set or qualitative comparative analysis (QCA) (Ragin 2000, 2008), shown as the middle cell in Figure 10.4. QCA is an approach to analysis that seeks to avoid making only qualitative or text-based *sui generis* interpretations of each case. Rather than only trying to fit a narrative or interpretative sleeve closely to each case, each one is instead categorized by being grouped with others as included in or excluded from a short listing of specific analysis sets (Bastow et al. 2013). The method aims to make only a series of very much simplified judgements, fundamentally of a yes/no type, in response to a small number of well-specified and single-issue questions of the form: 'Does case A belong in the set of social phenomena with characteristic X, or does it not?' A key alternative to dichotomizing responses uses a simple five-point scale for outcomes:

1	The case falls wholly within set X in a clear-cut way
0.75	The case falls partially within set X, more than it does outside it
0.5	It is impossible to determine whether the case falls within set X or not
0.25	The case falls partially outside set X, more than it does within it
0	The case falls wholly outside set X in a clear-cut way.

Further developments can carry the scaling/classification into more detailed (but less intuitively explainable) approaches.

Small *N* studies, with (say) between four and ten cases, can often offer improved clarity of analysis when compared with classic narrative deep or intermediate cases (Bastow, Dunleavy and Tinkler 2014, Ch. 4; Dunleavy et al. 2006, Chapters 4 and 5). As in the conventional approaches, the first step is to qualitatively describe in a richly detailed way the nature of each case, focusing on the social meanings of the people involved and often on a narrative of what happened, when. However, the point of these 'thick descriptions' is not so much to end up with any exhaustive characterization of any case, but instead to help the analyst decide how a case matches up or not against a question that defines a set. Multiple questions about different sets are normally constructed, defined in theory-relevant terms by the purpose of the study. Where an aspect of the cases lends itself to comparative quantification the analysis will not operate directly with the quantitative values themselves so much as look for cut-off points that are socially or temporally significant across the cases. A cut-off level defines the boundary of the set, and cases inside the set with different values on the relevant variables are nonetheless treated as having the same value answer, as are the different group of all cases outside the set.

The criteria for being included in a given set of cases, or not, are always based on the judgement of the analyst having conducted the core case research and reviewed the results for cases comparatively. Each set-defining question is not something stated in advance, but rather a judgement-based decision about what is most relevant or apparently crucial, drawing on the rich information available from all cases covered. The number of set-defining questions used in a study depends on how many cases there are, how complex they are and how fuzzily they can be characterized. With (say) six cases there could be no more than four or five relevant sets. If there were ten cases maybe eight or nine sets could be defined. The analyst must then construct a table cross-tabulating cases in rows down the vertical axis and set-inclusion judgements (or spectrum scores) in the columns on the horizontal axis. Each set-inclusion question much be explicitly justified (including the use of any cut-off points or spectrum choices). Each attribution of a case to a cell must be explicitly recorded and justified by the original analyst in a way that can be easily checked and potentially disputed by readers or later secondary analysis.

The logic of Ragin's analysis is that the social world is highly complex and can only be partially and often inadequately grasped by taking a variable-orientated view of it. In particular, variable-focused methodologies often lead to a kind of regression tournament between rival singular theoretical models, where the criterion of success is to explain more of the observed variation in the dependent variable(s) than rival models. Underlying this approach is often a vague meta-assumption that a single determinate causal process exists that shapes variations in outcomes. But the Ragin/QCA approach entertains the possibility that multiple different causal chains are involved in equi-causing the same outputs in complex social situations. For instance, some richly analysed cases may show a causal patterning of a type I outcome where they simultaneously fall into sets A, C and E (three necessary conditions) – whereas other cases may also show a type I outcome but fall into sets D and F (two different necessary conditions), or perhaps lie just in set B alone (a single sufficient condition).

> Qualitative comparative analysis (QCA) opens up two new forms of knowledge: (1) knowing about alternative pathways to one outcome (equifinality) and (2) perceiving nuances of necessary cause and sufficient cause (Olsen 2014).

Combining and comparing cases across set categories can generate additional insights that would not be available from just looking at the same cases in a text-based or narrative manner. In particular, QCA gives a lot of attention to unions of sets (e.g. all cases falling into *either* set A or set E) and overlaps or intersections of sets (e.g. just those cases falling simultaneously into *both* sets A and E). These additional fuzzy categorizations are also made explicitly and so are contestable by readers or later analysts.

QCA or fuzzy set methods resolve many of the openness problems around conventional case approaches. Collecting rich qualitative information forms the initial basis for the analyst to formulate a small number of set-allocation questions, each of which needs explicit justification in the methodology description, and an initial rationale for selecting the cases covered. The analysis should also record potential variations on set-defining questions considered, and for sets where inclusion or not can be quantitatively specified, the impacts of changing cut-off score boundaries. However, because QCA is an exploratory method, pre-registration of specific hypotheses is not a relevant step here (beyond the initial rational for cases included). The QCA case data tables and all the analyses made should be straightforward to deposit once the research is completed and published. Such high-level data categories should not normally raise any issues about breaching sensitive data constraints or anonymization promises to case research participants. These advantages are considerable compared with other small *N* or only lightly categorized qualitative case studies. But they also continue into the medium *N* qualitative studies that are now feasible using QCA software packages developed in recent years.

Medium *N* studies using QCA methods

With 10 or more cases qualitative comparative methods studies begin to cross over into the realms of quantitative analysis. Hand-crafted comparisons of cases become hard to organize and to compare union and intersection sets for; the number of explanatory set questions can expand, allowing more nuanced categorization of cases. Multiple QCA software packages now offer many more modes of categorization, including the identification of apparently important cutoff values (Melo 2021). And numerical codes for gradations of set inclusion replace categorical codes in ways that foster more numbers-based forms of analysis. Of course, the 'case' element in the overall analysis tends to decline as the number of units of observation increases (in a range from 20–50 cases). Yet even with a medium number of cases, the QCA approach retains a case-orientated, more holistic focus than conventional variable-orientated analyses (which tends to dissolve units of observation simply into congeries of variable values). This difference makes cross-referring between individual cases and overall patterns easier and of course stresses the concept of multicausation – where inclusion in multiple differently structured set combinations may nonetheless produce similar outcomes for different reasons.

In terms of openness, it should be easy to reproduce QCA case-based analysis with medium *N* cases. The questions used to define sets should always be registered as early as feasible, along with a later narrative about how systematic classification of the cases was undertaken. At the end of the research, data deposited should show the classification of every case on the finally used criteria, so that each case judgement involved is immediately visible and contestable. Numerical scoring systems are relatively restricted and their QCA

rationales are well-covered in the methodological literature. There may be good reasons for keeping them as simple and as intuitively explicable as possible in order to aid readers and reusers of the QCA coded data. The impacts of adding or excluding cases or changing scoring systems should always be open to exploration by later analysts. However, QCA studies with larger *N* (50 +) may run into some danger that codings and numerical scorings move a long way beyond readers' intuitive understandings of cases. This might potentially recreate the problem of inaccessible 'dots and dashes on computer databases' found with large *N* quantitative analyses that are unsupported by apt illustrations or insightful case or qualitative materials. Normally, however, even very numeric and technical QCA studies give more information on cases selected and how they were set classified.

Cases, systematic review and meta-analysis

In addition to the problem that every case is unique in some aspect, so that none is 'typical' (discussed earlier), and notwithstanding some reliance on 'model cases' in some disciplines, comparing across multiple, differently conducted small or medium *N* case studies has also been hard to do in the past. Cases that might seem to bear on each other at first sight have in practice often been undertaken in different spatial settings or times, and by differently trained authors (using concepts and methods of analysis that are discipline-specific). Even with academic case studies, the methodological quality and research provenance of studies can vary a great deal, because time or funding constraints and practical feasibilities (like timing) limit important checks and balances. For example, in randomized control trials (RCTs) exceptional levels of care are normally taken to produce reproducible research. Yet RCT studies often vary sharply in how effectively they 'blind' researchers (and sometimes even participants) about which people or cases were included in different treatment groups or in a control group. And the specificity of 'treatments' against external factors may not really be guaranteed (leaving alternative explanations of 'treatment' effects still possible). Some 'randomista' studies measure only relatively 'instant' treatment effects, while others resurvey after several months to find out if any initial effects endured, proved evanescent or eroded away more slowly (Leigh 2018). Durable effects matter most for researchers giving policy advice.

Beyond academia in most areas of social research huge numbers of potentially worthwhile enquiries are also undertaken by actors such as firms, market analysts, consultancies, genuine think tanks (as opposed to 'junk tanks' or lobby/front organizations), international organizations (like the World Bank, International Monetary Fund (IMF), European Union (EU) or Organisation for Economic Co-operation and Development (OECD)), government agencies, government inquiries, public service professionals and non-governmental organizations (NGOs). Many are external studies completed

for the commissioning body by social scientists as contract research. These studies should normally follow rigorous social scientific processes and may be better funded than pure academic work, and be subject to something close to peer review appraisal. Other studies are conducted by internal teams as part of organizations' normal operations and they may be far more variable in methods quality than academic analyses. In some cases, as with many consultancy and all 'junk tank' reports, it may not be clear what the 'research' has amounted to. Yet with these two exceptions, the rest of the organizations listed above often produce serious and purposeful research seeking to understanding applied but important issues, such as:

- identifying new trends in societal behaviours;
- working out what went right or wrong in a given project or operational initiative;
- establishing how a policy fiasco or disaster occurred;
- ascertaining what problem factors responded to a previous organizational treatment and for how long; and
- why some treatments worked in some places and not in others.

In past eras, when external research was far less professional than now, academic analysts tended to focus solely on studies produced in journals with peer review. The persistence of this stance has often meant that academic researchers have ignored or unjustifiably neglected valuable information in non-academic studies (or even just studies by academics from other disciplines), sidelining them because of differences or defects in 'othered' or practitioners' methodologies.

The potential contamination of non-academic studies' results by the funders' organizational biases or culturally induced judgements may also be seen as an extra risk factor with some external organizations, as more rigorous declarations of interest have developed in many disciplines. However, the costs of blanket exclusion of non-academic case studies can often be high in many fields. For example, in medicine and health studies numerous medical audits or small research studies are conducted annually, covering decentralized innovations in professional treatments. When cumulated they might generate extra layers of insights about some very specific issues – such as whether acupuncture treatments have any impact on lower back pain in different cultures and groups of people. And in development studies many penetrating assessments linked to funding or project interventions have been conducted by NGOs or aid bodies on local projects or unique interventions – for example, examining whether development assistance funding to counter child malnutrition works better by boosting direct nutritional or income assistance to households, or by giving more children meals at schools. Bear in mind also that in modern conditions any sustained case analysis done by firms or agencies within advanced economies will either have been contracted out to university or research consultancy units, or have been carried out internally by

professionally qualified research social science staff with postgraduate social science qualifications – especially in public services, finance industries, communications, etc. (Bastow, Dunleavy and Tinkler 2014, Chapter 9). In science- and tech-related industries a good deal of 'submerged social science' is also conducted by university or consultancy teams in a science, technology, engineering, mathematics and medicine (STEMM) academic discipline close to the commissioning firms, though with varying levels of awareness of social science issues. Both external and internal case projects are commissioned and reported to demanding line managers and funders, many of whom are well-educated in social science methods and demand high-quality research feedback.

The recent development of *evidence synthesizing methods* in the social sciences (following earlier progress in STEMM disciplines), together with modern digital databases resources, should mean that the cumulation of case study knowledge can now be far more effectively done than in earlier periods (Haddaway et al. 2021). In principle, both academic research and non-academic case studies can now be considered together using well-defined case-assessment and case comparison approaches. We focus here just on 'systematic review' (Pettigrew and Roberts 2006; Chapman 2021), but it is important to note that other essentially similar meta-review processes exist and may be useful for particular types of case literature. (In addition, meta-analysis is a well-developed approach to getting a pooled estimate of effects across multiple quantitative studies.) However, the integrative possibilities of systematic review in applied areas are often undermined in practice by review teams arbitrarily deciding only to look at refereed journal articles. In many qualitative social sciences, and across most humanities subjects, any such initial decision is doubly problematic, since books and book chapters are excluded, inevitably leading to results that are partial or biased.

The core idea of systematic review for case studies research is that the project team surfaces and then filters for quality the most inclusive set of documented relevant studies possible. Key stages are:

- The scope of the review and its limits must be precisely defined because the coverage within the boundaries chosen must be comprehensive and cover all relevant studies. So the criteria for including or excluding cases that are 'borderline' for relevance must be well-based in theory and methods considerations, well-justified and exact. For openness they should also be pre-registered.
- The evaluation criteria to be used to grade and assess studies must be fully defined and described before analysis starts, and again pre-registered. Some refocusing of boundaries and evaluation standards may occur when empirical analysis starts and researchers' knowledge widens. But in these cases a living narrative is needed showing that the changes made were not ad hoc modifications that biased the original scope.

- The largest possible pool of items is then surfaced by comprehensively scanning all the main relevant digital databases for journal articles, and research books and chapters. Because journals can be ranked and indexed via Google Scholar Metrics, Web of Knowledge, Scopus and other bibliometric systems there has been a disturbing tendency for many systematic reviews in the social sciences to only cover articles in high-prestige journals (a pattern too in the medical sciences). For a given research question, it will be vital to consider whether this is an admissible step or not.

  Wherever there is any significant literature in research books or book chapters, or in firm or government reports and 'grey literature', then only looking at articles is an arbitrary step, a move made for convenience that immediately wrecks any claims to comprehensiveness. Book databases can be scanned for significantly cited items within clearly specified boundary constraints. And high-value grey literature can be easily and systematically accessed via search engines or now large language models (LLMs). Any difficult decisions about excluded borderline studies are preferably assessed by two or three readers separately, comparing against the already set and fixed relevance boundary criteria.
- All the admissible studies are then individually read by researchers and their key results distilled. For large sets of studies (in the hundreds or more) scanning for relevant content and straightforward duplications via automated searches and newer artificial intelligence (AI) methods can be used to 'triage' large masses of work for a salient 'read-in-depth' subset.
- Within the final (stable) pool of relevant studies the researchers then apply the first (usually the most basic) exclusion criterion based on the methods used. For instance, a set of relevant literature numbering in the hundreds of items might be restricted in a way that reduces it to one numbered in tens of studies.
- The now-filtered and reduced set of studies is then passed through the second, third, fourth, fifth, and so on, exclusion criteria, administered in a sequence set by their increasingly demanding and specific nature. Where large tranches of studies are being set aside the reasons are noted, and any common trends in their conclusions are noted and retained for potential reanalysis and comparisons at the analysis-testing stage. For late-on exclusions more details are recorded and kept in mind than for those sidelined earlier on.
- In qualitatively based research areas, this funnel-down operation will typically aim to end up by focusing most attention on (say) 10–20 well-conducted studies that are the most reliable. Ending up with many fewer than 10 studies may make it too hard to draw comparative insights, suggesting that the exclusion criteria are too difficult to meet. At the same time trying to handle and keep in view key lessons from

many more than 20 to 30 high-quality studies will be complex and may be too difficult to do.

- Researchers need to decide when enough filtering has been done, and adjust the stringency of the criteria being used to end up with the 'right' numbers of cases – and to be able to justify their decisions here.
- For qualitative researchers, it will be helpful to try to identify (in QCA style) the variables appearing to be most important in the best-conducted case studies, looking for common variables and for missing expected variables, and covering any distinctive case factors appearing to be important in at least one instance. Analysts should be clear if they are seeking a single inclusive causal 'model', or instead supposing that multicausation processes may operate simultaneously.
- Researchers can next assess the sensitivity of their conclusions to the method choices made. For instance, they can ask how far their results would vary if the case selection to admit or exclude marginal cases criteria were to be changed; or if different criteria were applied in rating the quality of cases; or if the criteria were relaxed and applied less stringently.
- Finally, once a preferred 'synthesis story' of multicausation across qualitative studies has been defined, then researchers will look back to the mass of studies filtered out earlier for any strong or strikingly alternative results patterns amongst them – ones that might suggest possible contra-indications there for their view.

If the studies being compared in a systematic review have quantitative results (e.g. *Z*% of patients who were administered acupuncture for back pain reported an improvement, compared with a lower *Y*% in the control group) then the precise effect or other outcome levels found in the best studies may be pooled with a view to deriving a consensus effect estimate, while still explaining any divergences (perhaps arising from method differences).

Analysts can also try to specify any uncertainties remaining in summing up quantitative works, especially asking if results visible in less methodologically sophisticated studies provide extra assurance, or suggest an alternative hypothesis. They can also try to show readers what level of assurance they could have in any pooled effect estimates (e.g. 'a mean of 10.5% more patients reported pain alleviation from acupuncture treatment, plus or minus 5.1 percentage points').

In terms of opening up social scientists' use of cases studies (and of very localized or relatively small or medium *N* field research or even RCTs), systematic review (SR) is an important possible step to consider before adding another original study to the pile of existing work. The SR method can be more labour-intensive and time-consuming to conduct than conventional literature reviews, especially if boundary limits are not tightly constrained at the outset or are allowed to change in mid-flow. However, SR approaches normally generate more worthwhile cumulation of knowledge. They may especially assist

reproducibility researchers in identifying potentially anomalous previous work to revisit, while also increasing readers' assurance about well-founded studies' findings. Using SR methods can also identify where to undertake new salient or even 'capstone' cases – for example, giving more focused attention to apparently non-fitting or disconfirming cases may show how they may be reconciled better with wider findings or earlier results. The same steps that help improve new casing choices will also allow follow-on researchers to situate cases more accurately, better appreciating what motivated previous studies and how they were designed.

10.4 Field research

By definition, most social science field research takes place in non-experimental conditions where information gathering often cannot be completely or even tightly controlled (O'Donnell Goldensher 2023), and in the conditions of a specific temporal period that inherently cannot be repeated. Researchers go outside their research laboratories, or libraries or offices, to collect primary, firsthand or 'raw' data in specific settings that bring them close up to their research subjects and locales. Most but not all social science field research relies on case studies, but may differ in some other dimensions – such as the prolonged time spent observing groups in ethnography, or the collection of detailed STEMM or other quantitative data within the research site and period in health studies. Often these may also be settings where large flows of relevant and likely irrelevant empirical information will be gathered quickly and simultaneously, usually via less formal or observational methods. They will then need separating out and later, off-line processing to fully understand them. And much fieldwork will collect sensitive information, posing a problem for openness (Butnaru 2023).

In STEMM most field research shares these features along with a concern for holistic research. Yet nonetheless science and tech fieldwork has a driving focus on making highly precise information records, often entailing multiple measurements of natural phenomena or recovered artefacts (like archaeological remains) that generate a wealth of data for subsequent quantitative analyses. In the social sciences some field research can have this character too. However, most field research is more qualitative, focusing on understanding individuals and their life histories or world views, or wider groups, or whole communities of people involved in specific organizations or sub-systems (van de Ven and Poole 2017). Both STEMM and social science fieldwork stresses placing a thoroughly trained researcher (someone saturated in knowledge about a given topic) into an appropriate concrete setting where they are likely to be able to use that expertise to see things or identify phenomena that a common sense or amateur observer would likely fail to recognize as distinctive. For instance, a trained and experienced geologist will be far better at recognizing a rock that is potentially fossil-bearing or otherwise distinctive

despite its weathering. And a culturally expert observer will detect far more detailed patterns of significant social behaviours and non-linguistic modes of communication in community events or settings.

For these observations to be maximally useful the researcher will ideally need to record interesting material (and perhaps geo-locate and date the observation precisely). Both these tasks are far quicker and easier to do with modern tech than in the past. In more culturally constrained settings, if videos, photographs and a full firsthand audio recording are not feasible, then making a quick voice note into a phone or recorder, or jotting down notes using jotting software or lab notebook equivalents will be important (McSweeney and Winkler-Prins 2020). Recording quick, 'first impression' field notes about actors' behaviours and mores can be vital, along with ideas, theories or speculations – before any contemporaneous notes and memories get 'crowded out' (as they almost inevitably must) by later occurrences in the project.

After intensive field experiences researchers need to gain some space or better conditions for more reflective work as quickly as possible, so that they can also document well their highly evanescent personal reactions to things seen. Many authors suggest that this is an important time to begin journalling, so as to develop (possibly rival) hypotheses or tentative assessments about meanings and connections. Speedily capturing and not losing a wealth of potentially fleeting experiences, and quickly 'tagging' phenomena with potential meanings or codes that can be later collated and analysed, may be especially productive for sense-making, narrative-building or even later new theory construction. Some fieldwork lends itself to the development of narrative accounts that are inductively close to the researchers' original experiences, but easy to communicate to readers in accessible ways. More thematic analysis approaches, by contrast, can resemble the coding of text and quantitative analysis of interviews using text analysis software (discussed in Chapter 9), although the units of analysis here are usually relatively complex behaviours, meanings, norms or rules observed by researchers.

Fieldwork sites will need to be just as carefully chosen as case studies are, using developed and ideally pre-registered criteria. The choices made will vary depending on:

- how sophisticated the study being undertaken is (e.g. an RCT trial or a choice experiment rather than just a descriptive qualitative case study);
- how complex the setting is where the field research is located; and
- how that fieldwork informs the wider design of large *N* studies (as in many medicine, economics and psychology studies).

For instance, seeking to develop RCT or survey-based studies of how effective western-style vaccines or contraceptive advice are in the face of social resistance and strong religious sensitivities in areas like the northern regions of Nigeria (with polarized Christian and Muslim communities) would typically require that researchers have an immersive understanding of the situation on

the ground before framing survey questions or designing treatment and control group approaches.

The key steps towards making fieldwork more transparent and reproducible focus on the materials and methods used to collect information – especially research notebooks or journals that would be valuable for follow-on researchers. Again these decisions need to be made early on in a project's life when the means for recording field experiences is first planned. If field materials can be deposited, then ideally they should be easily machine re-readable, or be intelligently summarized or sampled in planned ways. This is much easier where researchers are used to documenting as they go (see Chapter 5) and reach the field well-equipped with computerized lab notebooks, journalling software that converts handwritten scrawls to clearly readable fonts, and powerful phone or other video and audio recorders for lodging data securely in real time.

Preparing such materials later on for deposit may require a certain amount of editing, cleaning and perhaps selection of records – all factors that may also condition archives' interest in and willingness to accept such materials. As Chapter 5 stresses, postponing this post-analysis editing stage to later on in the research process may ensure that it is never done and any data depositing effort also lapses. The closing stages of projects are rushed for research teams, and the passage of time or movement of staff to new work often create barriers to reunderstanding the field materials later on. These factors place a premium on researchers making maximally reuseable field notes and analyses first time around. This step should not be difficult to make, though, because it is also vital in enhancing the accuracy and depth of any field researcher's primary analysis.

Conclusions

Moving to more open practices in how case study research and fieldwork are carried out, especially via improved preregistration and fuller depositing of materials in repositories, has four immediate benefits for research teams and lone scholars alike (Humphreys et al. 2021). Readers can far better assess the *provenance* of case studies, gaining assurance from the pre-registration of cases' selection criteria and rationales. Follow-on researchers or potential critics can see that studies were undertaken for explicit and thought-through theory reasons and implemented in methodologically sound ways. Giving enhanced (especially quantitative) contextualization for small or medium *N* cases may also expand their status and the resonance of the case development narrative. The more developed and explicit a case study strategy is, the more the *reproducibility* of research (the most basic Turing Way criterion) increases. It becomes much easier for other researchers to test your interpretation against what they know of the cases covered. Where case methods are explicit and well-specified, follow-on research teams can also try applying a different analysis model to the same cases, strengthening the *robustness* of research (in the Turing Way categorization) if they reach similar conclusions, and highlighting issues for

further resolution and exploration if not. Well-documented case studies, clearly situated in a relevant 'universe' of related situations and supported with more open data evidence, will also be better regarded in the discipline and by outside readers, strengthening their *reputation* and citations.

Looking at the wider benefits for the broad field of research, greater openness in case study and fieldwork design and documentation can clearly enhance *replicability* (in Turing Way terms), increasing the cumulation of knowledge. Other researchers can use the same qualitative approach (or one of the methods discussed in Section 10.2) with other comparable or parallel cases, to check whether similar narratives work there, hopefully generating cumulative insights or sustaining systematic reviews. Where case materials are open then applying the same case methods in different societal contexts or time periods in ways that are comparable in depth can help other researchers to establish if the same results emerge, if so strengthening the *robustness* of knowledge (on the most demanding Turing Way criterion). In all these aspects, if case analyses converge on the same conclusions, then replicated results and case cumulation boost sub-field learning and give extra assurance on macro-issues that may otherwise be very hard to address – for example, assurance about when decentralization and development work to foster participation and policy responsiveness in democracies, but may have diametrically opposite effects under flawed democracies and autocracies or dictatorships (Faguet and Pal 2023). Even where case studies fail to cumulate, apparently yielding only divergent or *sui generis* explanations across different instances, cumulating across cases that can be systematically compared using stand-back methods like systematic review may well allow a deeper recognition of underlying patterns.

Note

* This chapter was developed from an original online seminar held on 19 January 2023, forming part of the LSE's input into the Civica project. The project was funded by the European Union's research and innovation programme Horizon 2020, under the chapter 'Science in Society'. The PowerPoint slides from the seminar and a video record of the presentation and subsequent discussions are permanently free to download from the Civica project's Zenodo pages: https://zenodo.org/records/7630418 We thank all the participants at this session for their questions, comments and inputs.

References

Bastow, Simon; Dunleavy, Patrick; and Tinkler, Jane (2014) *The Impact of the Social Sciences: How Academics and their Research Make a Difference*, London: Sage. **$** https://doi.org/10.4135/9781473921511

Bastow, Simon; Tinkler, Jane, Dunleavy, Patrick; and Williams, Sierra (2013) 'Making better use of case studies in studying academic insights', Paper to Political Studies Association Conference, Cardiff University, 25 March. https://perma.cc/SWZ6-7CZP

Butnaru, Denisa (2023) 'Being loyal to fieldwork: on building the "contract of silence"', *Journal of Data Mining and Digital Humanities*, https://jdmdh.episciences.org/12480/pdf

Chapman, Karen (2021) 'Characteristics of systematic reviews in the social sciences', *Journal of Academic Librarianship*, vol. 47, no. 1, article 102396. https://doi.org/10.1016/j.acalib.2021.102396

Dunleavy, Patrick; Margetts, Helen; Bastow, Simon; and Tinkler, Jane (2006) *Digital Era Governance*, Oxford: Oxford University Press. **$** https://perma.cc/4LEH-UJL7

Faguet, Jean Paul and Pal, Sarmistha (eds) (2023) *Decentralised Governance – Crafting Effective Democracies Around the World*, London: LSE Press. https://doi.org/10.31389/lsepress.dlg

Gerring, John (2004) 'What Is a Case Study and What Is It Good For?', *American Political Science Review*, vol. 98, pp. 341–354. **$** https://www.jstor.org/stable/4145316

King, Gary; Keohane, Robert O; and Verba, Sidney (1994) *Designing Social Inquiry: Scientific Inference in Qualitative Research*, Princeton, NJ: Princeton University Press. **$** https://press.princeton.edu/books/hardcover/9780691224633/designing-social-inquiry?srsltid=AfmBOoovgjdTnCOd4M7Xt74BO--yJpLYt9glezq_K1PVeSlVX30RVtKl

Krause, Monica (2016) 'Western hegemony in the social sciences: fields and model systems', *Sociological Review*, vol. 64, no. 2, pp. 194–211. **$** https://doi.org/10.1111/2059-7932.12008

Krause, Monica (2021) *Model Cases: On Canonical Research Objects and Sites*, Chicago, IL: University of Chicago Press. **$** https://perma.cc/3DYS-FF5R

Kushitor, Mawuli; Henry, Elizabeth G; Obeng-Dwamen, A.D.; Agyekum, Martin W. et al. (2022) 'Covert contraceptive use amongst the urban poor in Accra, Ghana; experiences of health providers', *Reproductive Health*, vol. 19, article 205. https://doi.org/10.1186/s12978-022-01516-5

Leigh, Andrew (2018) *Randomistas: How Radical Researchers Are Changing Our World*, New Haven: Yale University Press. **$** https://perma.cc/R9L2-YMVV

Lieberman, Evan S. (2005) 'Nested analysis as a mixed-method strategy for comparative research', *American Political Science Review*, vol. 910, no. 3, pp. 435–452. **$** https://doi.org/10.1017/S0003055405051762

Linke, Vara (2022) 'Book Review: "Model Cases: On Canonical Research Objects and Sites" by Monika Krause', *LSE USAPP* blog, 22 January https://blogs.lse.ac.uk/usappblog/2022/01/16/book-review-model-cases-on-canonical-research-objects-and-sites-by-monika-krause/

Little, Daniel (2020) 'Philosophy of History', *The Stanford Encyclopaedia of Philosophy* (Winter 2020 Edition), Zalta, Edward N. (ed.). https://perma.cc/VA3R-LDJU

McSweeney, Kendra and Winkler-Prins, Antoinette (2020) 'Introduction to the special issue: fieldwork in the 21st century', *Geographical Review*, vol. 110, nos 1–2, pp. 1–7. https://doi.org/10.1080/00167428.20110.1689361

Melo, Patrick A. (2021) *Qualitative Comparative Analysis: An Introduction to Research Design and Application*, Washington DC: Georgetown University Press. (Up to date on large *N* studies). **$** https://press.georgetown.edu/Book/qualitative-comparative-analysis

O'Donnell Goldensher, Liora (2023) 'Problems of knowledge, problems of order: the open science field site', *Frontiers in Sociology*, vol. 8, article 1149073. https://doi.org/10.3389/fsoc.2023.1149073

OECD (2022) *Building Trust to Reinforce Democracy*, Paris: Organisation for Economic Co-operation and Development. https://doi.org/10.1787/b407f99c-en

Olsen, Wendy (2004) 'Triangulation in social research: qualitative and quantitative methods can really be mixed', in Olsen, Wendy K.; Haralambos, Michael; and Holborn, Martin (eds) *Developments in Sociology* (Vol. 20), London: Macmillan. OA at: https://perma.cc/LSE2-H6QE

Olsen, Wendy (2014) 'Comment: the usefulness of QCA under realist assumptions', *Sociological Methodology*, vol. 44, no. 1), pp. 101–107. **$** https://doi.org/10.1177/0081175014542080 OA at: https://perma.cc/7BZN-JGHY

Olsen, Wendy (2022) *Systematic Mixed-Methods Research for Social Scientists*, London: Palgrave (now Springer). **$** https://perma.cc/TE8K-VKUM

Pettigrew, Martin and Roberts, Helen (2006) *Systematic Reviews in the Social Sciences*, London: Wiley. **$** https://doi.org/10.1002/9780470754887

Ragin, Charles C. (1992) ' "Casing" and the process of social inquiry', in Ragin, C. C. and Becker, H. S. (eds), *What Is a Case? Exploring the Foundations of Social Inquiry*, Cambridge: Cambridge University Press, pp. 217–226. **$** https://www.cambridge.org/gb/universitypress/subjects/sociology/research-methods-sociology-and-criminology/what-case-exploring-foundations-social-inquiry?format=PB&isbn=9780521421881

Ragin, Charles C. (2000) *Fuzzy Set Social Science*, Chicago, IL: University of Chicago Press. **$** https://perma.cc/S3WX-4549

Ragin, Charles C. (2008) *Redesigning Social Inquiry – Fuzzy Sets and Beyond*, Chicago IL: University of Chicago Press. **$** https://perma.cc/E4D9-GLC9

Rothbauer, Paulette (2008) 'Triangulation', in Given, Lisa M. (ed.), *The SAGE Encyclopedia of Qualitative Research Methods*, London: Sage, pp. 892–894. $ https://perma.cc/U9YD-TKBA

Sandelowski, Margarete (2011) ' "Casing" the research Case Study', *Research in Nursing and Health*, vol. 34, pp. 153–159. **$** https://doi.org/10.1002/nur.20421

van de Ven, Andrew H and Poole, Marshall Scott (2017) 'Field research methods' in Baum, Joel A.C. (ed) *The Blackwell Companion to Organizations*, Oxford: Blackwell, pp. 867–888. **$** https://doi.org/10.1002/9781405164061.ch38 OA at: https://perma.cc/9L8G-9BLV

11. Systematic documentation analysis

Analysing documents in a comprehensive and systematic way is a critically important qualitative research approach across most of the social sciences, especially in fields focusing on organizations and public policy decisions or implementation (Lamba and Madhasudhan 2022). Key evidence sources used include:*

1. *Contemporary (modern period) records* that are easily available – for example, government national archives (especially digital online and Freedom of Information/open government sources), materials on government websites and public documents required to be published by government departments and agencies, law courts and regulatory bodies. In addition, this category includes the public regulatory filings made by firms and businesses, along with their annual accounts and annual reports, risk registers, project reports to investors, and market and stock market statements and performance data – often both published independently and gathered together in large-scale databases. Most non-governmental organizations (NGOs) and civil society bodies publish similar information.
2. *Contemporary documentary records for which special access and usually permission from the owning organization is needed*, such as the non-public internal documents or organizational archives of businesses, firms and government agencies.
3. The *full text application programming interface (API) databases* of major social media and Silicon Valley platform corporations, and large-scale text data gathered from these or similar sources in any other way (e.g. via web scraping).
4. Other *digital open access versions of contemporary history sources* that are comprehensively archived.

In this chapter we exclude research using historical documents, dating back far longer and stored in national or scholarly archives, where readers need

How to cite this book chapter:

Dunleavy, Patrick and Monteath, Timothy (2026) 'Systematic documentation analysis', in Dunleavy, Patrick and Monteath, Timothy (2026) *Doing Open Social Science: A Guide for Researchers*. London: LSE Press, pp. 255–284. https://doi.org/10.31389/lsepress.dos.k

specific training. Because of their specific qualities and issues for openness we discuss the issues here separately in Chapter 12.

We begin by very briefly summarizing the many plus points of using documentary sources in qualitative social research, and some substantial potential problems (Section 11.1). Document-based analysis has also seemed to pose some important issues for the openness of research, and we discuss the reasons for this next (Section 11.2). However, the normal main steps towards open social science (OSS) can also be applied here, especially pre-registration and depositing methods and (where possible) full text data (Section 11.3). And, as with case studies, adding quantitative metrics to qualitative documentary analysis can strongly boost readers' assurance about the provenance of research and its reproducibility (Section 11.4).

11.1 The advantages and limits of studying documents

Documentary sources have many advantages for qualitative researchers (Bowen 2006; Gorsky and Mold 2019). The texts accessed are normally carefully prepared by multiple actors in organizations, and checked for accuracy – as with firms outlining their plans and projects for investors, filing documentation with regulators, or publishing annual reports for markets and market analysts. Similarly, in government agencies even internal reports and emails are often very carefully constructed and edited for accuracy by numerous officials. Published reports often integrate what organizations say that they were doing in staging major events or making policy shifts into a coherent narrative. Documents of this kind may thus appear to have high reliability or finality as statements of who did what. Accessing internal documents allows public narratives to be quickly fleshed out, firming up the sequence of events over time. And it produces numbers and dates that can be cross-checked exactly. Thus, documents are a type of evidence that cannot lightly be denied or gainsaid – especially with 'official' accounts.

Internal documents are additionally valuable in helping researchers to understand and illustrate in detail (parts of) organizational processes, perceptions and discourses. Researchers can also use interviews with actors, and their diaries or personal notes to capture more elusive emotions, understandings and causes shaping events. Full text or verbatim records of discussions are often harder to find, but where available and written up in informative ways they can help researchers grasp people's complex meanings, understandings, emotions, colouring and framing effects. Of course, such texts are also subject to the constraints of the setting in which the discourse was voiced or produced. For instance, professional public agencies generate extensive verbatim documents wherever trials are held, crimes investigated, medical complaints made, service-provision cases re-examined for professional mistakes or misconduct, or lengthy public inquiries undertaken (e.g. with witnesses

questioned about a policy disaster). Verbatim documents are often seen as the 'gold standard' for evidence, since every word that the people involved say is recorded just as it was said, and in addition the individuals involved are often speaking under special legal requirements to tell the truth.

At the same time, using documentation sources requires qualitative researchers to be alert to a daunting array of potential biases and familiar problems, shown in Figure 11.1. Most documents are purposefully produced – to argue a business case, make a 'political' case, create an audit trail, represent the viewpoints of responsible organizations or personnel, or explain why failures occurred, and so on. Consequently, they always give only one version of what happened that excludes some information, selects what is to be included ('filtering'), and edits completed documents before release – all actions likely to be strongly affected by the interests and viewpoints of the actors producing the text.

The effects of the common biases in Figure 11.1 can be controlled by well-known research strategies like:

- immersive but critical reading to create 'deep' familiarity with a narrative and spot issues or problems;
- constructing a careful chronology of events and actions against which the texts and actors can be located;
- triangulating between multiple documentary sources (e.g. using media, social media, diaries, biographies, other published books and academic articles, and other civil society sources);
- triangulating with non-documentary sources (especially interviewing the actors involved in a documented episode, and seeking external judgements from observers or experts about what went on; or checking evidence of photographs, videos, social media, group cultural events etc.); and
- considering 'reflexivity', being aware of the potential for framing effects from prior experience, or over-commitment and self-deception by researchers unconsciously reflecting cultural biases or self-interest (Berger 2015).

Academics generally accept that source biases can rarely be fully appreciated without iterative comparisons being made between theory questions and expectations, documents claimed to be empirically sound and triangulated evidence. In general, it is harder to counteract filtering and the *non*-inclusion of events or viewpoints in a documentary record, especially where they follow impartially applied institutional practices. It is often easier to spot disinformation elements, such as biases, deliberate falsifications, hype, and so on – because they often do not align with other elements included elsewhere in the same documentation – as when an organization's leadership describes their achievements in overdone terms that performance tables in annual reports show was simply hype.

Figure 11.1: Fifteen key problems in reading or qualitatively appreciating any documentation source

Problem or bias	Issue raised for research
Filtering	Some relevant materials are left out of the source through general (impartial) official policies or archiving processes
Sanitizing	A kind of filtering where the texts retained and stored specifically exclude any materials judged unethical, or that refer to unwelcome events or conflicts between organization actors. For instance, the official minutes produced in many organizations often contain no record of different courses of action considered, but instead project an image of unanimous agreement on the actions finally decided upon
Censorship	A kind of filtering where documents or records judged to be politically or organizationally sensitive or unacceptable for some other reason are recurringly eliminated from retained data
Self-serving bias	A kind of filtering or censorship focusing on excluding from the record any other discourses, views, actions or documents potentially reflecting badly on the producing organization
Pro-ethics bias	Representation of actions taken in terms of respectable moral principles or motivations, and non-mentioning of non-ethical or unrespectable ones – usually self-serving, but not always
Hype	Systematically recording promises, events, actions or developments favourable to the organization in ways that suggest far more (of significance) has occurred than actually has. Includes exaggerations, multiple announcements of the same policies or pledges, public relations-only content and over-optimism or Panglossian narratives ('everything is for the best in the best of all possible worlds')
Political or ideological bias	Doctoring of accounts and evidence to foster a particular political or ideological world view
Implicit knowledge	Unspoken assumptions or contextual knowledge, often about multiple small practices for treating information, or micro-institutions and administrative procedures

Figure 11.1: (continued)

Problem or bias	Issue raised for research
Cultural biases	Structuring of what is recorded in terms of organizational cultures or group meanings, values and world views, often inaccessible to outsiders and requiring 'grounded' decoding
Technical complexity	Pervasive use of esoteric, professional or area-specific concepts, vocabulary, formulae, jargon, acronyms and data in documentation
Authenticity/status	Was the originator of a text in a position to record accurately what is described? Were they in the right geographical or organizational locations, at the right rank, etc.?
Provenance	Are there clear records showing how (and by whom) the text was originally produced, and how it has been edited or handled (and by whom) since?
Falsification of evidence	Does a text actively or deliberately misrepresent events and actions, or show signs of clearly intentional lying or disinformation?
Mistakes in record- keeping	Inadvertent losses of evidence or misconstruction of texts by later handlers (e.g. mislabelling, miscategorization or faulty metadata making relevant materials hard to find)
Comprehensiveness, representativeness and selectiveness	Is the finished dataset being used by researchers sufficiently inclusive of all relevant documentation produced? If only a selection of available documents is being analysed, how were these selected (e.g. by randomized selection of text passages across a whole corpus otherwise too large to handle)?

11.2 Why documentation analysis may not seem 'open'

Access to the documentation sources used in research work is often (in theory) open to all. So in principle such sources are checkable by other (well-resourced) academic researchers, and perhaps by readers in general. For instance, most national official archives are open to anyone who applies for a reader's ticket – essentially a library card. This may seem to create a high degree of reproducibility, so long as researchers document exactly what they did. Researchers who are new to a given public source may nonetheless be

willing to consider a checking visit or even a replication study, especially if the process by which access can be obtained is well-described. Even where highly confidential information is involved, as with medical audit studies, other qualified researchers may be able to redo previous analyses by gaining access to the back-records of a health organization – for example, to investigate the effectiveness of hospital surgeons conducting operations, or of community doctors' (GPs') practices in diagnosing patients. So in principle it should be more straightforward to establish the provenance of most systematically accessed documentation than with other qualitative methods.

However, access to non-digital public sources may be made costly or infeasible by the need to travel to a single location to view offline documents, especially if the holding locations are distant or overseas. Specialist expertise may be needed to use many public policy and private firm text stores, along with detailed knowledge of the organization and its history for anyone else to make sense of them. For actors not already expert in such areas, evidence checking would demand a big, upfront time investment.

Many non-public documents may be closed by commercial confidentiality, or the need to secure difficult organizational clearances or permissions. Some internal business or government documentation may only ever be accessed by one original research team, granted unique access by power holders. In addition, some critically important government documents are comprehensively embargoed for long periods (e.g. 30 years for Cabinet minutes in the UK). In their absence researchers may have to rely on other, weaker or more flawed documentary sources (like ministerial diaries or media reports in the UK Cabinet case).

The government, firm and NGO documents most used by social science researchers are normally produced for internal or public relations purposes – not just to impartially describe a set of events. All official and most public documentation has multiple vague areas, deliberate silences, euphemisms, damage-limitation strategies and rhetorical rules that are complex to convey (see Figure 11.1 above). Getting past unacknowledged, withheld, or publicly recused passages, or heavily sanitized 'public relations' text is difficult – with reproducibility also limited by the passage of time, shifting patterns of causal relations, and environmental changes that can quickly alter the research 'field' or context conditions radically.

Consequently, the quality of academic work depends on how far researchers maintain an independent or appropriately critical stance about text sources they access. A weak form of Stockholm syndrome effects may operate here. Just as interviewers may feel sympathy for and even identify with the actors, groups or communities who grant them access and talk with them in person, so researchers committing hours and days to an organization's in-house archives (or even online records) may come to over-accept as valid or useful the narratives and sources uncovered there. After reading multiple filleted documents (on whose novelty and credibility their own work and career depends), researchers may tend to internalize some of the concepts, vocabulary and

world view of the actors or organizations being studied. This 'empathetic' risk increases where researchers depend on continuing permission access from the organization being studied. Firms, agencies, trustees of estates, or other bodies are likely to choose only researchers judged to be a 'safe pair of hands'. Hence critics frequently argue that research based on documentary analysis can tend to 'whitewash' or explain away what is covered.

This risk is accentuated because holistic group or organizational processes are often hard to capture or describe qualitatively. There is a plausible argument that there are only really four basic possible structures for writing about any social phenomena (Dunleavy 2003, pp. 68–74; 1986, pp. 86–97):

- *Descriptive accounts* (D) include detailed narratives and chronologies where one thing follows another, basically in date order following the David Hume principle that a cause must occur before any event that it causes. Built up on a large scale, as in many official inquiries, 'big book' research projects, PhDs and historical accounts, a massed descriptive account can easily seem almost theoryless. Their authors seek to appear highly objective by giving highly detailed explanations. Yet such massive narratives may also appear hostile to drawing any analysis or value-judgement conclusions. They may yield only complex and overly inclusive 'explanations', suggesting that everything happened for complex reasons and cumulative causes. Such accounts, beloved by conventional historians and implemented at massive length by official inquiries (e.g. Chillcot 2016), can be hard for readers to follow for any genuine causal argument, still less critique or contest their diffuse arguments.
- *Analytic accounts* (A), by contrast, use a set of abstract headings or categories to split up the social processes or phenomena to be covered into different narrative threads or separated explanatory chains (e.g. economic, political or cultural causes). The headings show how different 'factors' came to interact in ways that account for the phenomena involved. These grouped categories do not exist externally in the world, but solely in the minds of the researchers, however much some may be presented as 'common sense' (i.e. 'conventional wisdom') categories. Analytic headings are necessarily 'subjective' in what they highlight and what they assign less attention to, or even draw a veil over. If researchers give only sketchy qualitative explanations to justify their choices, analytic headings more easily seem ungrounded or selected solely to sustain a pre-defined argument.
- *Argumentative structures* (R) accentuate the role of researchers' subjectivity even more visibly, by setting out well-articulated 'pro' and 'con' positions for a preferred interpretation of the findings, reporting a battle of competing macro-explanations with accompanying evidence. Authors here normally conclude and may justify why one of the available lines of argument is most compelling, albeit perhaps mentioning

potential limits or amendments drawn from other perspectives. Again the researcher's subjective choices in choosing which preferred argument to support (and which counter-arguments to rebut) are prominent subjective choices in this approach.

- *Matrix patterns* may overcome some of these problems. They combine two of the three approaches above (rarely more) in a simple grid table sequence. One approach defines what happens in moving down each of the columns in turn, and the other approach sets what goes on in moving left to right across the rows. (In case this sounds too simple and restrictive, there are 12 possible binary combinations of argumentative forms here. With rows as the primary organizers (descriptive components) first, and columns holding second-order factors, these are DA, DR, AD, AR, RD and RA; plus the same pairings with columns as the primary (macro-argument) organizers and rows as secondary. For instance, the diagrams here show a matrix with four descriptive components (Di, Dii, Diii and Div) versus an analytic take (with two categories Ai and Aii).

	Ai	Aii
Di	1	2
Dii	3	4
Diii	5	6
Div	7	8

	Ai	Aii
Di	1	5
Dii	2	6
Diii	3	7
Div	4	8

In the left picture, rows are the primary organizer and the analytic column categories are secondary. On the right, the columns are primary and rows secondary. If a matrix pattern includes a descriptive approach as primary or secondary then the analysis may seem less 'subjective'. Some of the most accomplished modern historians and creative non-fiction writers are adept at subtly interweaving analytic or argumentative themes and narratives in this way (e.g. in modern German history, see Clark (2007) and Watson (2010)). Yet the subjectivity of author choices structuring the text are only somewhat mitigated with even such developed authoring skills.

How have social scientists sought to respond to worries about subjective readings or uses of documentation? One approach has been to develop deeper or more 'grounded' case analysis accounts, such as 'narrative analysis' or 'thematic analysis' (see the next section). In deciding what to cover here, most researchers make close connections between a single well-articulated theoretical position (or occasionally two positions compared) and a detailed methodology for selecting and interpreting materials from documentary data. However, in almost any causal context, there are now multiple alternative

well-grounded theoretical interpretations and literatures, so that subjective choices of which theory/methods approach to adopt inevitably shape what the findings draw attention to and what explanations they frame. Research teams must also consider whether to proceed with only one perspective (which may accentuate critics' misgivings around author bias in what is covered), or to examine the competing plausibility of two or at most three alternative theories. This is normally the limit of what is feasible in primary research even for research books, and always for articles or chapters. Of course, various types of review articles can obviously consider more approaches, and systematic reviews may also give the analysis more secure foundations (Haddaway et al. 2021) (see Section 10.4 above).

For other researchers and readers seeking to assess the strength or reproducibility of the original researchers' case, mastering the specifics of even one developed theoretical approach is no easy matter. It may seem to demand an upfront time investment that *de facto* makes full reproducibility only plausible for other researchers already inducted into the authors' chosen theories or viewpoints. Compounding these problems, defenders of single-theory analyses (using grounded theory or other perspectives) rarely justify their stance in terms of falsifying their chosen themes. Instead they tend to argue (rather vaguely) that work in their chosen theory can be evaluated only in terms of its accuracy, development of expertise, internal consistency, and its insights or 'fit' in helping to understand empirical evidence.

Both for simpler data checking and more extended assessments of how far original research is reproducible, extra problems arise with information obtained from global or large platform companies, which form increasingly important sources for social scientists. New problems may arise in establishing the precise provenance of research texts and data. For instance, researchers' methods cannot be fully transparent if they rely in part or for key operations on:

- commercial algorithms whose mode of operating remains a commercial secret – such as Google's or other search algorithms;
- the APIs of social media companies like X (formerly Twitter) or Facebook (now part of Meta), also held secret; or
- newer data science/artificial intelligence (DSAI) programmes, such as large language models (LLMs) like ChatGPT, Grok, etc.

Given their salience in modern social processes, many research teams might pragmatically see few problems in embodying non-transparent but widely used software into research designs, even if the validity or lack of systemic biases involved in such systems has to be taken on trust. Is this just an issue for academic purists, still demanding a form of complete control over the research process that is simply impracticable in the modern era? After all, only a relatively small proportion of social science work embodies commercial software as an inherent part of assembling documentary databases. And

all of academia already routinely uses search engines and LLMs in many other ways. Standard advice on similar issues from previous eras – like handling the radically differing citation counts generated by Google Scholar, Scopus, Web of Knowledge, Cross Ref and other search engines – has been to look at the correlation between alternative services' results.

Wherever it is feasible, it is best to avoid a research design dependent upon a single commercially confidential search algorithm. If unavoidable, researchers need to characterize any known or newly found biases or limitations as accurately as possible using alternative software. But the rise of DSAI systems means that the scope of these 'black box' elements in research methods is almost certain to grow quickly over the next two decades. For instance, ChatGPT has already been used to analyse linkages between central banks' policymaking (Leek, Bischl and Freier 2024).

All these issues mean that even 'gold standard' documentary analysis can be questioned on multiple open science (OS) grounds if the authors do not disclose in open fashion the texts analysed, what methods were used and details of how to reproduce their work. How comprehensive and inclusive were the documents or text sources covered? How were they selected for inclusion/exclusion or for more intensive analysis? Were the documents covered the most relevant and authoritative ones for this organization's (or actor's or group's) involvement in key events in the relevant area or time period, plus other situational factors? How far were the theory used and methods of document review chosen ones that were likely to purposefully or inadvertently shape the documents covered and their interpretations? In particular, how can we be sure that evidence has not been selected by the analyst to support a particular (desired) interpretation? Did choices made on document selection or access, and on methods, tend to lead analysts to draw particular findings/conclusions? Overall, is the analyst's interpretation legitimate or credible, given the text sources and collateral information? Is the analysis critically undertaken and assessed? What alternative interpretations of the texts or documentary evidence exist, and have the research team explained why they are seen as not feasible and credible for the situations covered? How does the documentary evidence triangulate with other sources – especially interviews, case studies, reactive survey data, or patterns of objective behaviour? Finally, how far does this work fit with or diverge from other existing work in the relevant sub-field, and what explains salient differences?

11.3 Enhancing openness for 'immersive reading' approaches

The first stage of qualitative research for any documentation source is always for the analyst to read all the source evidence falling within the study's scope. This task is best done just once in a systematic way that is outlined in a pre-registered document – especially where the research authors then create their

own text-only, wholly qualitative account using interpretative or 'constructivist' methods. Tracking a complex story requires careful scoping, asking critical questions and achieving a rich understanding – and depositing some form of curated research archive in ways that can underpin the analysis for later users.

Scoping and setup

Most solely qualitative researchers tackling a new source seek to 'immerse' themselves in the full volume of relevant texts, despite the high time costs involved. With any large text sources the boundaries of what is to be read need to be explicitly and tightly defined from the outset. Sometimes immersive reading will require learning about peripheral or tangential matters – for example, to establish that a firm, agency or NGO always behaved in the same way across other issues as well as the one in focus. On the initial boundary criteria there are bound to be near-miss or tricky boundary cases (e.g. where two readers disagree about whether they are in or out of scope). These cases should be included with the personal archives that researchers construct for working analysis, along with any boundary changes that later reading shows are needed, so that their impact can be explicitly assessed if need be. However, in the first instance, researchers need to keep their attention tightly just on the in-scope text, albeit noting interesting peripheral matters for potential revisiting if time allows.

Do not just plough into careful reading at the beginning of research, by beginning documentation analysis 'on spec' or with the expected 'best bits'. Instead, looking ahead and considering the whole task systematically is sensible, perhaps also making some exploratory 'quick dives' across the whole array of texts – accessing small text samples right across the full range of available materials, to check what is there. Does the format, inclusiveness and information quality of texts stay the same across a whole period of years in a longitudinal study, or across organizations or settings being compared? What is included in documents may often vary, so that materials available in one period or setting are not necessarily there all the way through. Quick looks spaced across the whole texts to be covered may also generate some advance insights, very conditional impressions that may help analysts to surface additional core questions. It may also alert researchers to any of the limitations covered in Figure 11.1.

To select well in and synthesize information from a large text source, a robust framework for recording events, noting interconnections, pulling out full text passages for possible reuse and making notes needs to be set up from the outset. The practices adopted here ideally need to be applied in the same form across the full range of documents to be covered, which is simpler if they are homogenous in format. But it is a more complex task where the texts included vary in format or character as reading progresses, even in small ways. A way of

indexing documents quickly and in standardized fashion is needed (e.g. precise short names or acronyms for sources or authors, dates, and item/document numbers or paragraph/page numbers in folio documents, or chapter/subsection headings and numbering in digital documents). Replicate the conventions used by document authors here. Charts, table and diagram numbers are always needed too. In open access digital documents that are otherwise not broken up, for each salient passage to be retained record a quote of six or seven distinctive words in the text, a unique conjunction that you can re-find using control+F search.

Next, before beginning full reading, it is very helpful to pre-register core research questions (and any prior expectations about them from the literature review, interviews, media or other sources), wherever possible in a date-stamped document (either held privately or made public). This can also be kept up to date as the research develops. Expectations provide an explicit base against which to compare or contrast stories that emerge as more internal or public texts are read and a preliminary analysis begins to develop. The same document can also define a metadata structure for salient records to be used in shaping the analysis, especially recording high-value (quotable) passages that help shape the final analysis in easily reaccessible ways.

Where complex Boolean search codes are used to surface relevant material to read in very wide and large archives, and to generate a focused text dataset (as in much digital humanities research), then full search strings need to be recorded along with the inclusion/exclusion criteria for what is looked at or not. Later on, if it proves better to not stick to these initial boundaries, documenting why a change was made at the time and explicitly defining new criteria or categorizations is important. Later researchers' ability to track code changes made from the initial plan is a key boost to transparency, allowing easier identification of changes that need explanation. By contrast, unacknowledged code changes reduce readers' assurance in the original researcher, and may easily be misinterpreted as 'gotcha' moments, especially where non-academic readers get involved in historical or intellectual controversies.

Achieving 'depth' understanding

The first stage of immersive reading of documents covering a time period is normally to construct a narrative or chronology framework, filling in bits and expanding detail where they can. Such accounts need to be holistic – without unexplained puzzles or bits that do not fit – but different approaches can be taken – including pure date-order ('this happened, then that happened') chronologies, life stories (Lieblich et al. 1998), periodized narratives (with analytic characterization of causation and interactions in each period), and so on. To track an evolving view of what happened and what explains it, most researchers find it helpful to keep some form of journal or use e-notebooks (see Chapter 5). 'Storyboarding' a specific article or chapter or book can be a useful way of

defining an argument core (Dunleavy 2014). Instead of trying to define a detailed article structure in a top-down or sequential ('begin writing at the beginning') way, storyboarding recommends sketching out a set of sub-heads and micro-stages within them, and progressively adding extra materials as they become available (like key exhibits, tables and figures; or quotable quotes; or just key analysis insights and potted arguments). This expanding record can then be actualized in carefully written prose when the complete 'story' has stabilized, with a wealth of materials behind it.

Alternatively, researchers might write a revisable/disposable summary narrative – an instantly disposable ('tomorrow's fish paper') document designed to go out of date and be revised as more is known and understood. Consider posting as a public document either a whole or partial storyboard, or another interim account, or summarizing the progress of analysis periodically in blog posts. Both are great ways to register the project's development and to attract early (formative) comments and feedback from other experts in the area – which can be very helpful in steering ongoing research.

The basic proto-documentation stage needed for analysis to work involves researchers reading from the start in an over-time dataset, or overviewing spatial or organizational variations, making brief 'pointer' notes for a later write-up phase or a storyboard as they go along, and recording ample meta-data needed to easily re-find information. Using paper files or having one-time access to closed or internal sources that might not be reaccessible also means that researchers confront difficult 'take away' decisions. How much full text from source documents should they download and self-archive, or photograph or scan into their phones (assuming this is allowed in a given library or archive), or try to get photocopies of for text content in the most backward libraries? Collecting too much take away material (e.g. downloading everything) can consume valuable time and clutter up study records with distractor or 'confuser' material that later turns out to be unusable or not important (see Chapter 5 on the burden of 'omnibus' record keeping). Recording too little take away materials risks author regrets in writing up results, where perhaps relevant points or referencing support are vaguely recalled but cannot now be accessed or cited in a specific enough way. For instance, referencing that includes precise document titles and quote locations is always needed for PhDs and articles and books at research level.

From the outset researchers need to ask not just, 'What might it be desirable to record?', or, 'What do we absolutely need to record?' (both open-ended questions), but also 'What can we use in our analysis/planned publications and own writing?' (always a smaller, finite answer). Identifying and recording possible high-value quotes verbatim on first reading is often critical, before then moving them (or the best of them) into the storyboard document. Similarly, when the bulk of data selection and compression can only be accomplished via taking notes, it is important to try to photograph/scan and retain some substantial materials that show the precise language and style in which the source documents are written, especially where the analysis throws

up interpretations that may be controversial. For key points it is also helpful to collect 'show and tell' materials (like copies, photographs or facsimiles of documents) that clearly show the original character and import of major text passages or quotes, or the overall style of the documentation.

Sometimes library constraints may mean recording must be done with pen-and-paper notes, or by retyping text into a PC, or scanning/photographs using a phone. Researchers are naturally keen to minimize what they collect for potentially quotable passages or numbers. But around these it is wise to take 'too much' text before and after the key passage for context, as well as carefully recording metadata or indexing details at the same time. Never leave possible quotes on first reading to re-find later, even in fully open digital texts. Using PC Notes software to make and keep notes systematically with clear records and searchable full text is key to developing a useful research archive (see Chapter 5). It can seem time-consuming to build an archive systematically (since reading is quick, but notes/recording slower). But, in itself, a constructed, researcher-selected archive of material is always a useful evidence source in a PhD or research project, and it may be separately depositable in ways that are potentially useful for other researchers.

As analysts read deeper in a dataset or archive there can be a temptation to 'park' the job of beginning analysis. Some researchers avoid formulating or noting down any interim thoughts or evaluations until 'all the materials' have been absorbed and a complete set of evidence has been selected and recorded. Postponing analysis in this way is always a bad idea and often reflects a misconception – that somehow the answers to theoretically interesting questions or solutions to complex causal analysis questions just lie 'out there' in the documentation, waiting to be discovered. This is never the case. Meaningful analyses are always constructed by analysts in a long iterative process of going backwards and forwards between theories, hypotheses or other propositions, and evidence. They are not just picked up lying around in a hiding place. So it is valuable to begin the analysis process (in a formative-only way) even when only a large fraction of documents have been accessed.

Thinking things through early on (even in tentative ways) is also important to speed up and improve how reading progresses, by registering new developments more explicitly and revising the analyst's reading frames. It is also important to regularly monitor the volumes of material being collected and the time being taken, revising quickly if needed. In many qualitative social science fields it is easy to get locked into an infeasible framework for collecting/recording. 'Collectomania' can also be a real problem, and not just with first-time authors or PhD/post doc researchers. Conscientious researchers using big source documentary datasets can routinely gather far more materials than they need to analyse their research question, or than they can ever actually use in their own research texts – a temptation most acute with textual analysis following a descriptive structure, where authors have the fewest structuring cues about what is or is not important.

Curating and depositing a documentation archive

At later stages of research projects qualitative researchers face choices about making self-archived materials easily available for others to read. Academics may fear that their approach was too personal, even idiosyncratic, to interest anyone else, or perhaps that their records are too disorganized and unsuited to being reused, or somehow may reflect badly on them. In fact, however, sets of notes, exhibits and selections of documents derived from large datasets and focused specifically on one topic are *curated* resources that researchers have picked out, organized, labelled and commented on. (The modern concept of digital 'curation' goes far beyond its original museum or aesthetic context meanings (Williams 2009).) So a well-focused and collated collection is almost invariably worth depositing as a research archive with an independent value, separate from and additional to that of the articles or books produced using them, or a grant-funding report sent to a funder. As Pershon (2020) notes:

> Curation leads to tangible and publicly sharable products, beyond the boundaries of academic dissemination. Curation affords researchers a methodology to identify, refine, interpret, and present social constructions surrounding bodies of artifacts while inviting viewers [or readers] to ask their own questions and seek their own interpretations of the artifacts [or materials] presented.

Even if a research archive is solely text and notes, it adds depth to the research project and provides resources that facilitate reproducibility. Including other types of artefacts (images, photographs, videos, etc.) adds extra value. A well-curated and deposited project archive can encourage other researchers to use the same documentation sets or to try and replicate the approach in other temporal or spatial contexts.

Many researchers using open and digitally available document sources (like agencies' or firms' annual reports and minutes, or media outputs) unreflectively believe that there is little point in archiving their research troves – because all the basic materials involved are already easily available. Yet any project's research archive selects for relevance to a specific theme, and gives many insights on how to read and use the documents included – especially if accompanied by a comprehensive chronology or narrative of events, plus additional background evidence for key themes already summarized in note form. Ideally, it would also include a research journal (like an anthropologist's field book) recording the original researchers' journey to understanding the documentation (see Chapter 5). So, almost paradoxically, the value of curated documentation resources is greatest where all the source materials grouped together by the original researchers are themselves open access. Then the hugely time-consuming document-read-and-selection stage has already been completed for follow-on readers. Curation here means picking out a specific theme from the mass of other topics muddled together in the same source

documents. By contrast, the research archive of a qualitative text analysis is a thematically unified set of full texts, assembled in a reputable, reusable and handy way for any reader wanting to pursue the same theme. Researchers with different interests but using similar kinds of documentation may also be able to learn a lot about how to get the most out of that source before starting work.

Curating and depositing a research archive is especially valuable in a different way where the source documents are closed access, inaccessible to ordinary readers, or only available in a single library or archive location. Here the original researchers' careful notes of what they have read take on extra value in systematically summarizing what was available. And researchers can also flesh such (potentially contestable) records by including as many supplementary materials as they can, especially:

- 'Show and tell' evidence, like photographs, photocopies, or self-scans of key documents. The amount that can be copied may be constrained by the policies of the host library or archive, or by copying costs if only institutional copies are allowed, and it is always limited by the researchers' available time. Self-copying in this way may often have to focus only on some documents, but these are the most critical or valuable ones.
- Photographs or copies of images, data tables, statistics, diagrams, formulae, plans or other similar elements are especially useful for follow-on readers.
- Where especially important documentation cannot be copied because of restrictive source archive policies or facilities, then fully retyping verbatim parts (or the whole, if feasible) of salient documents can often be doubly useful. It boosts the original researchers' credibility and greatly assists later researchers or readers without similar access.

Researchers who plan to deposit archives at the project's end, and who want to foster their sub-field by encouraging others to study similar sources, can usefully anticipate some of these needs. If time allows, this might mean collecting and archiving somewhat more materials than are strictly needed for their own purposes. Of course, researchers will need to prioritize what they specifically need over other materials that are only potentially useful in an archive way.

Because they save time, and create more readable resources, using PC or lab notebooks can help research teams from the outset to document-as-they-go in more systematic, depositable and reuseable ways (see Chapter 5). Electronic notebooks and scrapbooks will often be publishable online just as they are, with little extra work involved at this stage, or only a minimum of extra guidance needed in a help document. By contrast, handwritten notes or journals may often be scrappy, enigmatic, or filled with shortcut notations that made sense at the time but even a year later may seem oracular to their authors

('What did I mean by that?'). Some amplification, clarification and additional explanation before depositing notes materials is useful, even for experts in the sub-field. Going the extra mile in depositing helpful guidance and induction advice to other researchers wishing to access the same online source is good practice.

A final factor that has previously caused researchers to doubt the value of archiving their curated research materials has been the feeling that very few people (if anyone) would ever read it. In the pre-digital age this was perhaps an apt reason for not archiving materials, since options for permanently backing up source materials were scarce. And official research archives often seemed to look down their noses at curated archives from unfunded, small or very specifically focused projects. However, just as the advent of digital search massively broadened the range of articles and books being cited in the research literature (Acharya et al. 2014), making good quality work far more easily accessible, so pervasive digitalization and enormous advances in search capabilities mean that deposited research archives are now easily findable, so long as their metadata are done correctly (see Chapter 5).

When linked to from published journal articles or research books, even small project archives can be a potent way of stimulating further research in the same area, encouraging add-on work in the same field, or fostering copycat efforts even in other countries or different periods or social spheres. There is considerable (but still anecdotal) evidence that archives for popular articles and books are themselves accessed and used frequently and generate additional citations in their own right. Obviously in small disciplinary sub-fields, and less common languages, usage will be less. But over a long run of years it can still mount up in surprising ways, helping other authors, adding usefully to the 'dynamic knowledge inventory' (Bastow et al. 2013, Chapter 9). Especially where a project archive (or a proto-archive, in an almost-there condition) already exists for other reasons – for example, for other project team members to be able to access, for inclusion in a funding report, or for PhD supervisors or a committee to see – depositing is a decision with no downsides.

11.4 Quantitative metrics to support documentation analyses

Documentary sources lend themselves to many different forms of numbers-based quantitative analysis that can help analysts to strengthen and pluralize the evidential basis underlying their qualitative interpretations. We consider coding within texts and audits across online web documents. Whatever method is used, counting things within or across documents can create new quantitative data that provides extra context and constitutes a resource that is valuable in its own right.

Counting within full texts

'Integrating digital data and methods can extend the scope of qualitative social research. However, to date, this has hardly ever happened' (Franken 2022).

As with interviews, the essential step involved here is extracting counts from wordage. At its simplest, anyone can use the control+F key in PDFs (or equivalents) to check how often multiple connected words, phrases or concepts (or their opposites) are used across a document (or whole document set). This gives an instant and objective (if only initial) measure of how important particular vocabularies are in the organization's culture or in an actor's mindset. (It also works well in understanding authors' priorities in books and articles or in compiling literature reviews.) For instance, a sustained yet intuitively accessible approach was used by Moretti and Pestre (2015) to analyse 'Bankspeak', quantitatively charting changes in the content of World Bank documents over two decades, allied with an in-depth, periodized analysis of the bank's changing policy stances.

Counting inside a text relies on first coding words or phrases within the text, which is much easier to do if the texts are already in digital database form, avoiding the often time-consuming first stage of entering the text digitally. Manual coding of high-interest items during immersive reading is only feasible with relatively small datasets. Proceeding by stages, at least a pair of researchers in larger projects (to check for intercoder reliability) must familiarize themselves with the text, intuitively surfacing possible codes separately (Jawad et al. 2021). Sole researchers need to handle coding largely alone but should get reviews and comments from others. In a separate phase a possible thematic framework can be sketched to guide detailed coding and then fleshed out as researcher's understanding grows.

> Data coding is more systematic, with reorganization based on similarities which lead to categories and then themes, using methods such as meming and coding that go beyond those that simply echo the questions posed to participants [*in survey work*] or preestablished categories [*for documentation analysis*] (Ankeny and Leonelli 2024, p. 258). [*Our italics*.]

However, researchers must still judge subjectively whether concepts or vocabulary are being used in the same ways across occurrences. Defenders of manual coding argue that a complex qualitative decision must often (always?) be made on whether a word recurrence has the same meaning as other occurrences. A lot may hang on the context of use and on subtle nuances of phrasing – what might be called the 'tone of voice' with oral discourses (Mackieson, Shlonsky and Connolly 2018). Once an overall set of codes has been achieved then grouping and checking them for any obvious gaps, anomalies or oddities (and getting outside comments on this aspect) can identify any researcher blind spots (there are always some) and add extra depth and coverage (Jawad et al. 2021).

With any larger non-digital document scanning using optical character recognition (OCR) or text image analysis using artificial intelligence (AI) developments such as large language models (LLMs), mean that there are now multiple ways of achieving straightforward and reliable digital text that can be recognized and word- or phrase-counted instantly once in that form. Most university researchers should be able to either get computational library support for coding tasks, or perhaps be able to master simple automatic systems unaided (Lamba and Madhusudan 2022). It is key to be able to clearly define the boundaries for either the single long text or set of multiple documents involved (such as a series of reports, minutes, verbatim records of discussions or debates, or same-subject items collated from media or APIs). Deciding what is in and what is out of scope, without ambiguity, becomes far more important the more researchers move away from solely qualitative analysis.

Yet however clear-cut the search universe, the categories involved in word or intra-text counts are always subjectively chosen, either:

- *in pre-coded ways*, based on the researchers' (prior or pilot-stage) intuitions or understanding, concepts taken from the previous literature, or after exploratory work. Pre-coding may be seen as more 'open' and reproducible (akin to pre-registration, see Chapter 4). But it can also seem too rigid and perhaps primitive to many researchers, compared to their mature understandings; or
- *in post-coded ways*, drawing on researchers' fuller understandings of meanings, based on careful qualitative analysis or immersive reading of the full document set. Post-coding may be seen as more flexible, more useful and more accurate (Franken 2022). Yet both the original researchers and reusers may also need to guard against unconscious biases (creating counting categories that best sustain a preferred interpretation). Just as for interview transcripts moved online (see pp. 216–218), emergent coding frames can be surfaced from within a well-defined document corpus to count the frequencies with which vocabulary, concepts, memes and phrasings occur across a single full text, or across multiple different texts; or
- *using externally given codes*, either taken from other studies (e.g. to facilitate comparisons or to follow a clear norm in a given sub-field), or suggested by automatic systems, including now sophisticated LLMs.

Progress in the digital humanities and media studies has also broadened attention beyond texts, with visual or audio referencing systems supplementing natural language processing, and 'found' patterns being important in fields like digital history (as well as those discovered by analysis) (Luhmann and Burghardt 2021). Different strands in digital humanities have also developed, some close to hypothetico-deductive approaches, and others shaped in a critical spirit (Viola 2025).

Using modern text analysis software packages (e.g. NVivo, or Atlas.ti) or most modern AI software set up with appropriate prompt cues can 'industrialize' coding and enable speedy analysis of large texts or documentation datasets. From prior immersive reading researchers should be able to define terms of interest in sophisticated ways, for example coping with different word endings or word formations, the phrasing contexts where they occur, conjunctions and debarred conjunctions, 'partner' words, unusual vocabularies or synonyms, and so on. Expert users in natural language processing may find it useful to use a computational analysis package like Python to first 'tokenize' text at the word or sub-word level (associating word combinations and variations with numerical codes), which can control better for context.

As analysis proceeds, the main software packages can also suggest ways for the coding pro forma to be refined and updated in transparent ways. They can also suggest terms to code based on autocounts and built-in repertoires for recognizing more complex underlying concepts or interactions of ideas. With data science/artificial intelligence (DSAI) elements increasingly added to software, computer analyses of findings and production of fuller reports can also be automated. Critics argue that automated counts and AI analyses should only be undertaken after close reading of the whole text set (Omer et al. 2021; Pope et al. 2000), and that even AI techniques are cruder than manual coding and less able to recognize nuances of meaning. AI coding and counts are perhaps liable to be biased by patterns or usage in wider populations (e.g. the 'training' data for LLMs), which may not be appropriate for many specialized groups or organizations. However, in many professional, organizational and media contexts vocabularies tend to be used in pretty standard ways, and here such problems should be minimized.

As with interviews, an alternative strategy is to define rich, complex meanings using qualitative comparative analysis (QCA) or fuzzy set approaches (see Chapter 10). Here analysts use initial qualitative work (immersive reading of the text and narrative surfacing) to pose specific questions about sets with dichotomized yes/no responses (or modified five or seven category response questions, see Section 10.2). Whether particular bits of documentation are included or excluded within these analyst-defined sets creates a coding. With several (or many) such codes, different counts of set inclusion can be made along with permutations of sets across whole collections of documents. For large *N* document studies, QCA software will be needed for counts charting complex discussions of issues or issue combinations, especially changes over time periods or spatial regions.

Counting across documents and web audits

In multitext analyses being able to recognize the occurrence (or not) of salient terms across documents can be used to trace variations across sources or over time. So instead of the occurrence of terms within a given text corpus, the count here is of relevant items including any use of a target word, phrase

or concept. For instance, in minutes or sequences of press or social media stories, counting the occurrence of documents including concept X by years or months can show breakpoints in the take-off of issues over time, or systemic differences across different types of media outlets (Altheide 1996, 2000). In generating macro-level, audit-style counts of the incidence of themes or components across documents researchers can create new quantitative dataset themselves.

An especially potent and easily accessible source are online websites of organizations (from $N = 20+$ up to maybe 300 cases). Web audits can provide a powerful way to codify and support otherwise subjective-seeming judgements about what a group of connected organizations are doing. While interpreting what organizations say in qualitative terms relies on subjective interpretation, and can be contested, a more objective approach is to code the presence or absence of multiple micro-indicators that together capture key dimensions of differences between organizations like firms, agencies or NGOs. They often make similar claims to be operating in a commendable or progressive manner; for example, improving sustainability or addressing climate change, pressing ahead with digital change, monitoring the social impacts of their activities, or treating different groups in the population equally and fairly. Yet there are clearly variations between them. Where researchers can only offer a qualitatively documented judgement – for example, that organization X is performing well on sustainability but organization Y is failing – the analysis often seems to lie solely in the eye of the beholder. By contrast, far more objective results can be achieved using multiple micro-codings systematically across comparable organizations documents (e.g. firms' annual reports, regulatory reports, or environmental and social governance (ESG) statements). The documents involved are now almost invariably accessible on their public websites

In the past, academic researchers have often failed to use online data for fear of falling for 'hype' or just accepting biased or public relations (PR) statements as facts, given the numerous 'unreliable text' problems listed in Figure 11.1. Organizations may claim to be doing things that in fact they are not, and they may say they are not doing things that in fact they are. This *used* to be a problem when firms' or agencies' online and web content was chiefly static PR text, slides and pictures (in the late 1990s and early 2000s). But it is far less of an issue now that the online content and websites of firms, agencies and NGOs are integral parts of their organizational processes.

However, the problem shown in Figure 11.2 must be tackled. Researchers must critically infer from what an organization claims to be doing to what it actually is doing. The researcher's expertise will always be crucial here, but much of the time analysts will be able to judge accurately that web claims are accurate (the yellow cell in Figure 11.2) or are evidently false (the pink cell) since online content will be aligned with activity. There is still some risk of analysts falling for solely façade or PR activity (bottom right, white cell in Figure 11.2), but it is far less common now than in earlier decades. It is much harder for organizations to engineer misleading content now, because they

Figure 11.2: The extent to which an organization's claims to be undertaking well-regarded activities match with the profile visible in its public documentation

		Does organization claim to be active in an area?	
		Yes	No
Is relevant activity visible online in this area?	Yes	CLEARLY ACTIVE Individual features exist and are visible – cumulative evidence shows that the organization is strong in this area (as it intends to be)	PROBLEMATIC In normal service areas doing good by stealth? inadvertent activity?
	No	FAÇADE ACTIVITY ONLY The organization does PR hype, 'goes through the motions', or has effectively de-prioritized the issue	INACTIVE Individual features not found – cumulative evidence suggests that the organization is either ignoring this area or failing there

must maintain so much content online and it must be consistent. Modern social media and DSAI technologies can both quickly surface and help publicize unfounded claims, creating magnified reputational risks for false claims.

A second problematic possibility is that organizations mask from public view their own ethically acceptable behaviours (grey cell, top right, Figure 11.2). This risk is high in some very specialist contexts (such as the secret activities of national intelligence agencies), and in other organizations at particular time periods (e.g. when firms are developing commercially sensitive new products or agencies are trialling innovations). But in general the risk of firms, agencies and NGOs 'doing good by stealth' is small.

To see how far a web-based audit can be useful here, consider a case where major firms in a given industry, or the top 100 firms in a stock market, are all claiming to be improving environmental sustainability, taking actions against climate change or reducing carbon emissions. In evidence they offer up positive ESG practices and policies in documents that they must issue online, such as their annual reports, regulatory filings, regulator judgements about major firms, and carbon intensity information for market analysts. Researchers here need to identify a detailed framework of indicators that ideally would be presented by a strongly performing firm. An index of good performance could be constructed as a series of individual micro-aspect questions with yes/no answers, each focusing solely on a single, discrete indicator of pro-sustainability behaviour. For instance, possible items might be:

- Is this topic or element covered in the firm's annual report?
- Does the firm publish a developed ESG plan? If so, how long is it? Does it include topic A, topic B, topic C, etc. Does the firm issue statistics on A, B or C, or not?

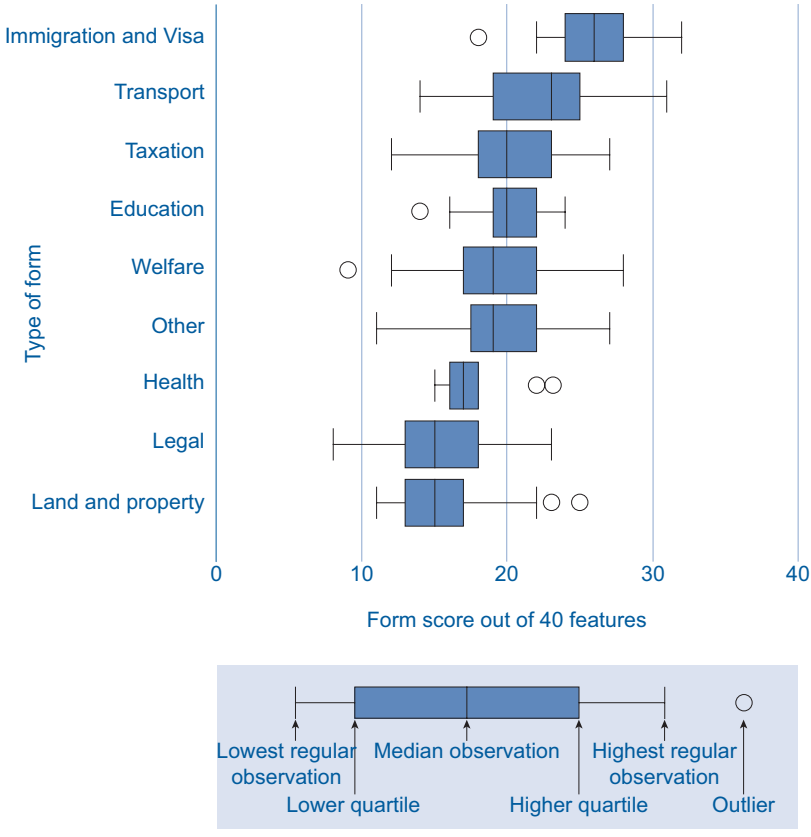
- Has the firm issued a briefing to market analysts on environmental measures?
- Does it identify detailed risks from global warning and climate change?
- Does the firm publish its carbon emission data?
- Does the firm submit to independent audit of its ESG plans and practices?
- Is the firm's data monitored and certified by any independent monitors of climate performance?

Ideally, a large number of such micro-practices should be coded on a *yes, present* versus *no, absent* basis. In addition, quantitatively scored objective items could be added to flesh out the number of indicators. For example, how many words are devoted to sustainability in the annual report? How much statistical information is given on a whole range of different measures? Or how are the firm's policies scored in star ratings by independent monitoring bodies? What rating do regulators give the firm's behaviours? Each micro-item measures a small, even tiny, outward sign of an overall policy stance, ideally in a way that yields yes/no or quantitative number answers.

In themselves, each of the factual micro-indicators included may seem trivial or of contestable relevance for determining the organization's overall performance. The number of feasible micro-indicators will also always be shaped by the specific situation being covered and analysts' judgements of what actions are symptomatic of an aspect of overall behaviour. But if a considerable number of dichotomized or objective numerical indicators can be assembled (at a minimum, say, 20 indicators, and ideally at least 40–50), then organizations' cumulative scores on the whole set of micro-indicators can be insightfully compared against each other. Compiling an overall score for each can be done using simple additive scores of indicators present (ranging from 0 to the maximum number of indicators, with each positive score counting as 1). An organization that scores 5 out of 50 is highly unlikely to be as active in an area as one that scores 40. Using micro-indicators, all of them in themselves providing only small clues about underlying activity levels, also means that analysts rarely need to bother about using a weighted index, which is useful since weighting different items in any objective way is often very hard to do. To give an indication of the outputs from this kind of web-based scoring exercise, Figure 11.3 shows how UK government administrative forms (such as tax returns, passport applications, regulatory permission forms, etc.) scored on a set of 40 micro-indicators of agencies paying attention to users' needs when filling them out, with high scores showing agencies taking care to have accessible and easily understandable forms, and low scores indicating a lack of care. Immigration and visa applications were clearly the most carefully designed for users to fill out, while those for land and property and legal forms were the worst designed. Overall, the central level of ease-of-use was around a 50% presence of desirable features (National Audit Office, 2003).

Figure 11.3: Example – results from a 2004 web census of UK government agencies' administrative forms scored on 40 quality indicators of care and attention for users

10 How the forms score on quality indicators of care and attention for users



Note

The shaded box show the position of the middle half of the data. The lines to left and right show the lowest and highest quaters of the observation, which are less typical. The O shows an outlier, that is highly unusual obsevation which analysis suggests is very different from the rest of the data, and may need separate explanation.

Source: National Audit Office census of forms

Source: Reproduced with permission from National Audit Office (2003).

This example illustrates another desirable feature of using web-based audits of this kind. The government forms covered here included every major UK government digital and paper form (via PDFs), each of which was accessed and scored by a researcher in an hour's work, with monitoring of coder comparability also easily implemented. Where materials are available online, getting full access to relevant documents can be quickly achieved in an unobtrusive

and non-reactive way, without requiring a reactive response from someone in every organization (unlike interviews, for instance). Consequently web audits have no (or hardly any) missing data, an important advantage that simplifies analysis and improves accuracy.

The method has some OSS limitations. Web audits need to be implemented quickly across all the units covered to ensure that changes over time are not an issue. And reproducibility is more problematic if only the analysis coding datasets are saved. The complete source documents are easily accessed by the project team (e.g. all major firms' or agencies' annual reports) but typically cannot be deposited for copyright ownership reasons, unless they have open access certification. However, some selective 'show and tell' examples can be collated to back up numerical analyses.

In some parts of the world or areas of social life archived full text documents may only be available on paper or as uneditable PDFs. However, web scraping can be a useful tool here, so long as the same information is included recurrently in an easily findable way (e.g. using the same language, numbered exhibit or perhaps on the same page) across the series – as with many report series including statistical data tables. Researchers define a simple set of codes that will autoaccess multiple cases of the same document series, then go to the same place (given by the code), and download either the whole table or particular elements included as numbers in a database, without having to scan and transcribe everything (Graham et al. 2020, Section 2.7). Bradshaw's *Web Scraping for Journalists (2017)* provides an easy guide for journalists, which could meet the needs of qualitative researchers with limited coding capabilities.

Conclusions

Moving toward greater openness in documentation analysis has multiple benefits for researchers by improving the *provenance* of findings and conclusions and showing that immersive reading was undertaken with due care. Supporting quantitative analytics enhance the authority of findings. The more curated the document text deposited, and the more transparent the analyses made, the greater the *reproducibility* of the research (a key Turing Way criterion) (Persohn 2020). Providing curated datasets enhances the *robustness* of research if other researchers get convergent findings (another Turing Way criterion). Analysis clearly supported by open data evidence will have a better *reputation* in the discipline and outside.

In the wider sub-field, *replicability* is enhanced if other researchers can reuse well-documented search processes and codes to find and define parallel document datasets (e.g. covering the same theme in a different country or time period) and use the same methods as the original research to see whether similar interpretations and results emerge as plausible there (again a key Turing Way criterion). If convergent findings emerge (e.g. across different contexts or time periods) then *generalizability*, another Turing Way criterion, is boosted. Positive findings here can boost accuracy in a sub-field

(e.g. reaching a consensus on change chronologies and narratives), with *positive learning gains* giving extra assurance to reviewers and readers. Even when results are negative on replication and generalization, a learning gain is made, often prompting further work with the same document sources and texts.

Note

* This chapter was developed from an original online seminar held on 25 January 2023, forming part of the LSE's input into the Civica project. The project was funded by the European Union's research and innovation programme Horizon 2020, under the chapter 'Science in Society'. The PowerPoint slides from the seminar and a video record of the presentation and subsequent discussions are permanently free to download from the Civica Zenodo pages: https://zenodo.org/records/7630488 We thank all the participants at this session for their questions, comments and inputs.

References

Acharya, Anurag; Verstak, Alex; Suzuki, Helder; Henderson, Sean et al. (2014) 'Rise of the rest: the growing impact of non-elite journals', *arXiv*. https://doi.org/10.48550/arXiv.1410.2217

Altheide, David L. (1996) *Qualitative Media Analysis*, London: Sage. https://doi.org/11.4135/9781412985536

Altheide, David L. (2000) 'Tracking discourse and qualitative document analysis', *Poetics*, vol. 27, pp. 287–299. https://perma.cc/S8XQ-CMGS

Ankeny, Rachel A and Leonelli, Sabina (2024) 'Investigating research practices: how qualitative methods enhance philosophical understandings of science', *Qualitative Psychology*, vol. 11, no. 2, pp. 247–262. https://doi.org/10.1037/qup0000289

Bastow, Simon; Dunleavy, Patrick; and Tinkler, Jane (2014) *The Impact of the Social Sciences: How Academics and their Research Make a Difference*, London: Sage. **$** https://doi.org/10.4135/9781473921511

Berger, R. (2015) 'Now I see it, now I don't: researcher's position and reflexivity in qualitative research', *Qualitative Research*, vol. 15, no. 2, pp. 219–234. **$** https://doi.org/10.1177/1468794112468475

Bowen, Glenn A. (2009) 'Document analysis as a qualitative research method', *Qualitative Research Journal*, vol. 9, no. 2, pp. 27–40. https://doi.org/11.3316/QRJ0902027

Bradshaw, Paul (2017) *Scraping for Journalists* (2nd edn), Lean Publishing. **$** https://perma.cc/3LSU-4NG2

Chillcot, Sir John (2016) *The Report of the Iraq Inquiry*, London: HMSO. https://perma.cc/F3JJ-LM49

Clark, Christopher (2007) *Iron Kingdom: The Rise and Downfall of Prussia, 1600–1947*, London: Penguin. **$** https://perma.cc/6HZH-76FC

Dunleavy, Patrick (1986) *Studying for a Degree in the Humanities and Social Sciences*. London. Macmillan. $ https://www.bloomsbury.com/uk/studying-for-a-degree-9780333418420/

Dunleavy, Patrick (2003) *Authoring a PhD: How to Plan, Draft, Write and Finish a Doctoral Thesis or Dissertation*. London: Palgrave Macmillan. $ https://www.bloomsbury.com/uk/authoring-a-phd-9780230802087/

Dunleavy, Patrick (2014) 'Storyboarding research: how to proactively plan projects, reports and articles from the outset', *LSE Impact blog*, 31 October, blog post. https://perma.cc/VC89-EQGM

Franken, Lina (2022) 'Digital data and methods as extension of qualitative research processes: challenges and potentials coming from digital humanities and computational social sciences', *Forum Qualitative Social Research*, vol. 23, no. 2, article 12. https://perma.cc/9TX3-BJD6

Gorsky, Martin and Mold, Alex (2019) 'Documentary analysis', in Pope, Catherine and Mays, Nicholas (eds) *Qualitative Research in Healthcare*, London: Wiley. **$** https://doi.org/11.1002/9781119410867.ch7

Graham, Shawn; Gupta, Neha, Smith, Jolene; Angourakis, Andreas et al. (2020) *The Open Digital Archaeology Textbook*, online resource, 11 July, Section 2.7, 'Scraping data'. https://perma.cc/4PHF-U49K

Haddaway, Neal R; Woodcock, Paul; Macura, Bilana; Collins, Alexandra M. (2015) 'Making literature reviews more reliable through application of lessons from systematic reviews', *Conservation Biology*, vol. 29, no. 6, pp. 1596–1605. http://dx.doi.org/10.1111/cobi.12541

Jawad, Aalaa; Savona, Natalie; Penney, Tarra; and White, Matin (2021) 'Industry responses to the UK government's public consultation on the proposed soft drinks industry levy: qualitative documentary analysis using the framework method and a systems perspective', *Food Policy*, vol. 118, article 102498. https://doi.org/11.1016/j.foodpol.2023.102498

Lamba, Manika and Madhusudhan, Margam (2022) 'Text data and where to find them?', in: *Text Mining for Information Professionals*, Cham: Springer. **$** https://doi.org/10.1007/978-3-030-85085-2_2 OA at: https://perma.cc/W8JB-FGLJ

Leek, Lauren; Bischl, Simeon; and Freier, Maximilian (2024) *Introducing Textual Measures of Central Bank Policy-Linkages Using ChatGPT*, Department of Political and Social Sciences, European University Institute. https://perma.cc/NKB4-DWD6

Lieblich, Amia; Tuval-Mashiach, Rivka; and Zilber, Tamar (1998) 'A new model for classification of approaches to reading, analysis, and interpretations', in Lieblich, Amia; Tuval-Mashiach, Rivka; and Zilber, Tamar (eds) *Narrative Research*: *Reading, Analysis, and Interpretation*, London: Sage Research. https://doi.org/11.4135/9781412985253

Luhmann, Jan Burghardt, Manuel (2021) 'Digital humanities – a discipline in its own right? An analysis of the role and position of digital humanities in the academic landscape', *Journal of the Association for Information Science and Technology*, 25 June, vol. 73, no. 2, pp. 148–171. https://doi.org/10.1002/asi.24533

Mackieson, Penny; Shlonsky, Aron; and Connolly, Marie (2018) 'Increasing rigor and reducing bias in qualitative research: a document analysis of parliamentary debates using applied thematic analysis', *Qualitative Social Work*, 10 July, vol. 18, no. 6, pp. 965–980. $ https://doi.org/11.1177/1473325018786996

Moretti, Franco and Pestre, Dominique (2015) 'Bankspeak – the language of World Bank reports', *New Left Review*, vol. 92, March/April. $ https://perma.cc/J8PT-AFXJ

National Audit Office (2003) *Difficult Forms – How government agencies interact with citizens*, London: National Audit Office, Value for Money report No. HC 1145 Session 2002–3. Research by Patrick Dunleavy, Helen Margetts, Francoise Bouçek and Rosie Campbell. https://www.nao.org.uk/reports/difficult-forms-how-government-agencies-interact-with-citizens/ Also see: https://researchonline.lse.ac.uk/id/eprint/596/1/cover-NAO_2.pdf

Olsen, Wendy (2022) *Systematic Mixed-Methods Research for Social Scientists*. London: Springer. $ https://link.springer.com/book/10.1007/978-3-030-93148-3

Omer, Usmaan; Veysey, Martin; Crampton, Paul; and Finn, Gabrielle (2021) 'What makes a model prescriber? A documentary analysis', *Medical Teacher*, vol. 43, no. 2, pp. 198–207. https://doi.org/11.1080/0142159X.2020.1839031

Persohn, Lindsay (2020) 'Curation as methodology', *Qualitative Research*, vol. 21, no. 1. https://doi.org/11.1177/1468794120922144

Pope, Catherin; Ziebland, Sue; and Mays, Nicholas (2000) 'Analysing qualitative data'. *British Medical Journal*, vol. 320, no. 7227, pp. 114–116. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1117368/

Viola, Lorella. (2023) *The Humanities in the Digital: Beyond Critical Digital Humanities*, London: Palgrave Macmillan. OA at: https://library.oapen.org/handle/20.500.12657/62372

Watson, Peter (2010) *The German Genius: Europe's Third Renaissance, the Second Scientific Revolution, and the Twentieth Century*, London: Simon & Schuster. **$** https://www.simonandschuster.co.uk/books/The-German-Genius/Peter-Watson/9781416526155

Williams, Alex (2009) 'On the tip of creative tongues', *New York Times*, 2 October. https://www.nytimes.com/2009/10/04/fashion/04curate.html

12. Archival data – problems and early solutions

For a number of reasons, mainstream history and social science studies based on archival research have been a lagging area in terms of open science (OS) developments, despite the vigorous growth of digital history. In addition, the wider interdisciplinary field of digital humanities is often concerned with historical themes or artefacts in the arts, literature, languages or philosophy (Berry and Fagerjord 2019). Conventionally studied contemporary history is an overlap area between the humanities and the social sciences, while history for older periods is chiefly a standard humanities discipline (Guldi and Armitage 2014). Archival work forms the central method and has sustained a prestigious and resolutely qualitative approach (Mills and Mills 2019) written up as research 'big book' texts, some of which command a large general audience and media coverage that bridges across into high-sales commercial publishing. Journals have been less important, with chronically low citation rates. And access to academic historian jobs at major universities still hangs critically on early career researchers producing monograph books rather than journal articles. As journals in the UK and Europe increasingly go open access, citations have grown slightly, but books (especially well-cited important books) have remained resolutely paywalled.*

We begin by outlining why archival research is different from systematic documentation studies and has distinctive ways of studying and writing up sources (Section 12.1). Next we examine what scope exists for increasing the openness of archival research using standard open social science (OSS) approaches and others tailored specifically to archival work with non-digital sources (Section 12.2). We conclude by considering the balance of logistical or operational barriers to openness relative to cultural constraints (Section 12.3).

12.1 Archive resources and research

The notion of 'archives' tends to immediately conjure a picture of traditional library shelves packed to the rafters with box files or ancient manuscripts in

How to cite this book chapter:

Dunleavy, Patrick and Monteath, Timothy (2026) 'Archival data – problems and early solutions', in Dunleavy, Patrick and Monteath, Timothy (2026) *Doing Open Social Science: A Guide for Researchers.* London: LSE Press, pp. 285–298. https://doi.org/10.31389/lsepress.dos.l

leather-bound covers, suggesting solely documentary contents. In addition, though, non-documentary or not formally published elements add considerably to the complications of archival work:

> In American usage, the term 'archives' expands to an umbrella concept for all primary documentation, including personal papers (such as correspondence, diaries, and manuscript writings). It is also used for non-commercial photographic evidence, non-commercial videos, and home movies. Also included in the general concept of archives are ephemeral materials such as posters for an event (e.g. an election). The initial intent for these materials is that they will be discarded after the date of their original use passes.
>
> These materials are generated with one specific purpose in mind (for example, sending orders to an army or enrolling a child in a religious congregation). Then later the documents are preserved and used for a secondary purpose: writing history. Archives are documents that are no longer needed for their original purpose, yet have significant informational and evidential value for the purpose of writing history. (Schmidt 2016).

So, 'Archival research entails analysing data that were stored other than for academic research purposes' (Henga et al. 2018) and that additionally have been subject to selection for retention. Increasingly, a prominent historian of industrial innovation noted that 'The mission of historians is to impose narratives on some pretty boring stuff contained in archives' (Howes 2023d).

Even comprehensive-sounding resources, such as 'national archives', normally contain only a selection of diverse papers and resources from earlier periods judged worthy of permanent retention. Digital archives may be much more generous than paper or object archives in terms of what is covered, but retention is still usually essential. Consequently, archival researchers need to be expert not only at reading the records or objects they access, but in understanding how these came to be selected and the relationship they bear to a wider set of information that has not been retained (Clary-Lemon 2014).

Archives are very helpful document sources for contemporary history which bridges into the social sciences, but they are absolutely unavoidable for earlier periods whose study most historians see as defining their discipline. Outside the digital history sub-field, covered at the very end of this section, many mainstream (qualitative methods) historians have been at best sceptical about seeking greater openness in their discipline, or even hostile to it (Blaxill 2023). Openness has been strongly identified with shifting away from nuanced narrative history writing (for which the most prestigious modern figures in the discipline are rightly renowned). Sceptics fear that its replacement might either veer towards a solely digital-history pattern, or lead to a more quantitative history approach (notably developed in modern economic history).

It might follow the direction taken by open archaeology – a humanities/science, technology, engineering, mathematics and medicine (STEMM)/social science overlap area (Graham et al. 2020; see Knöchelmann 2019). There, open methods from STEMM, pre-registration of research and omnireliance on recording every aspect of digs in computerized notebooks (like Jupyter, see Chapter 5) have all taken firm root. Traditionalists also fear that an open history could split into narrow and more intensively sub-specialisms with reduced audiences in order to secure grant funding dependent on meeting open science (OS) requirements. More general historical coverage might then be relegated to textbooks or mass market, pop-history books (a burgeoning market). The end result (they fear) would be a less accessible, less prestigious, more technical and rather off-putting academic discipline, whose public prestige would fall and whose current attractions for students and PhDers would reduce.

Historians also point to a series of major logistical barriers because many archives are still often not digitized. Consequently archival access may be restricted by:

- permissions barriers, such as access rights restricted to professional scholars, students of particular institutions only, or readers who complete some other form of vetting process;
- geographical access – history documents or artefacts are often made physically available only in one home institution;
- linguistic restrictions, such as the prior need to study the historical languages that texts are written in, which in some cases may require years of specialist study to master;
- other specialist archival research skills or contextual knowledge needed to even begin to understand what given texts may mean or why historic artefacts were produced or how they worked; and
- copyright constraints on photographing, copying or reproducing images of documents or artefacts, or including such images in subsequent published academic work.

At the same time, there is a powerful but still minority 'digital history' movement that is committed to the open source assembly of digitized key documents and other artefacts (especially images and cultural objects), so far chiefly on a project-by-project basis dependent on specific funding for initiatives. Underlying the rapid development of this approach from the 2000s was a strong commitment amongst U.S. scholars to 'democratizing history', using open access publication and the creation of dozens of individual specific online archives to enlarge the potential for lay people and general readers (as well as other academics) to access documentary archives in an open source manner (Locke and Wright 2021; Rosenzweig 2006). This effort has been especially linked in the USA to telling the historical stories of minority groups or historically oppressed communities, to redressing the balance of attention in 'mainstream' historiography and to enthusiasm for 'participatory

action research' (PAR) (see Chapter 13). It also explains the importance of open access journals in areas like Latin American history (where indigenous people's histories and cultures were largely submerged for centuries by Spanish and Portuguese invasions) found in a UK study (Finn 2019). At the same time the field had many 'stay away' critics, some highlighting the possible evanescence of digital sources, echoing Richard Rinehart's quip that: 'Digital sources last five years or forever, whatever comes first' (Bennett 2013). That line of attack has subsequently receded as more and more historical sources have become digitally available (albeit with physical backups in most cases) and backup systems have improved. More recent critics argue that the technical know-how involved in digital history and open source archiving methods can seem exclusionary. It appears to shut out not just general readers but also non-technical qualitative scholars from easily finding, understanding or citing this somewhat 'enclaved' body of work. Some counter-critics within digital history in turn accuse the mainstream profession of adopting 'Luddite' positions (Blaxill 2023).

12.2 Why is greater openness needed in historical research?

Critics argue that conventional history has only included relatively sparse descriptions of the sources and methods used, so that the discipline may have a reproducibility crisis (Howes 2023a). Methodological thinking amongst historians remains unusually sparse by modern standards. And, apart from progress in digital history, they have not advanced over unusually long periods of time:

> While there have been extensive debates about the nature and methods of historical research since the nineteenth century, the underlying assumption has generally been that documenting one's sources in a series of footnotes is essential to enable other researchers to test the validity of the research. Even if this approach never amounted to 'reproducibility' in the sense of scientific experimentation, it might still be seen as broadly analogous, since the evidence can be reassembled to see the basis for the explanations that were offered and to test their validity (Burrows 2023).

Of course, the only set of people who could seriously and easily contemplate undertaking any 'reassembly' work for most of the modern period would be other historians working in the same broad sub-field, trained in archival methods and familiar in detail with the specific archives used in the original study.

However, digital era advances have strengthened the problems of history as a 'generalist' discipline – for example, changes such as speedy and standardized communication and data-handling, digital archiving, excellent search, the development of widespread IT literacy, open sourcing of knowledge and

the growth of social media – to name a few. They have strongly increased pressures for the specialization of sub-fields within history. In addition, current historical training methods primarily develop a conventional set of tools and methods, set within a strong disciplinary culture. 'Historians must constantly invent interpretations from the evidence that will be challenged by later generations of historians' (Howes 2023d).

Generalist history accounts can potentially be far removed from the views and experiences of other highly skilled professional people working in the same or an adjacent area. In studying the development of any sub-field – such as science, or medicine, or public policy, or agriculture, or war, or international relations, or poverty – historians are now likely to reach an audience amongst whom are many differently trained academics or practitioners who can understand (and oftentimes better understand) the technical subject matter being covered. Anton Howes gives as an example the still widely disseminated and believed claim that Napoleon Bonaparte stimulated the invention of canning by running a competition for innovations to help his armies wage longer campaigns in the field. In fact there was no such competition (Howes 2023a). Instead, in a completely separate story, a French scientific academy awarded a money prize to an inventor who had invented a canning method 10 years earlier. Howes suggests that if we are to prevent escalating numbers of such dud historical 'facts', historians need to work more closely with professional communities in technical areas and local communities in spatially defined issue areas.

So a key first rationale for more open history, then, is to enhance the quality and reliability of historical 'facts' and to prevent the chronic mass circulation of 'factoids' that are not accurate. This can now best be achieved by making knowledge of archival data more accessible to a far wider range of readers, some of whom may have strong comparative advantages over necessarily somewhat more 'generalist' historians – such as

- professional scientists, engineers, medical, economics or other technical experts, etc. interested in exploring the historical origins and development of their own fields of work;
- academics and researchers in other related fields with expertise to contribute on the technical operations involved; or
- 'lay' people with specific knowledge of a terrain, locality, family history, local skill or other esoteric 'ordinary knowledge'.

For Howes (2023d):

> Historical enquiry can best prosper, and most effectively marshal the evidence, by professional historians being challenged quickly, not just by other professional historians but by the many diverse 'publics' with differently setup expert or intensive 'ordinary knowledge' of research fields.

Expertise in data assembly linked to knowledge of the primary subject matter has an especially potent 'extra' contribution to make to the effective and accurate exploitation of many kinds of the information lying only partly explored in historical archives. For example, amateur climatologists in the eighteenth and nineteenth centuries accurately recorded weather details (Shuttelworth 2015). STEMM researchers trained in marine biology and ecology have found that the detailed logbooks of whalers stretching back to the seventeenth century shed a lot of light both on the whaling ships' practices and methods, but also on estimation methods for the historical population numbers of different types of whales and their recovery times from progressively more intense over-fishing episodes (Ogden 2016). Wider studies have also found that many types of historical fishing records dating back to medieval times (including things like engravings showing large numbers of static shore nets used in some areas) illuminate historic population numbers and trends for many different types of organisms across most marine ecologies (Holm et al. 2010). Even in what might seem a remote field, such as geophysics, collections of folk tales, songs and pictures about past eruptions can provide scientists with key clues about the back history of many different volcanoes (Averett 2022; Holm et al. 2010).

As a second key rationale, the possible canvass for historical research is vast and so choosing areas for productively undertaking new research is a major problem, much as it is in STEMM, but with vastly fewer resources. Allied with steadily improving archive-searching capabilities, open access changes in how historians operate offer a hope of considerably increasing the efficiency of historical research, and reducing the incidence of duplicated or abortive research efforts, which Howe (2023d) estimates are large. From the earliest stages of historical research that is pre-registered and later deposited in open access ways, other academic historians and other expert readers will be able to get a much better picture of what has already been done. The more historians record in detail and openly how they accomplished studies, and post guidance on the strengths, limits and problems of a given source for their own specific research, the more scarce resources can be optimally targeted. If open logs of experience are regularly updated, this feedback can be near-immediate, and may reach other scholars of whose existence the open access authors are completely unaware – for example, researchers planning to research a completely different topic or exploring a similar archive but covering earlier periods.

A third key rationale for open history is that it enhances the subject's reproducibility in the digital era that stretches ahead – which is likely to be vital in maintaining its funding and academic credibility with other discipline groups. New generations of academic historians will make their whole careers in an OS climate where past deference to the limited quality assurance given by conventional 'footnotes and peer review' approaches has weakened a great deal. Partly due to competition from social scientists, in contemporary history minimalist 'legacy' modes of 'standing up' historical arguments are no longer

seen as sufficient evidence tokens for high-quality research. This effect is also likely to move 'downstream' to affect earlier periods, partly due to the development of digital history as a sub-field and the more detailed narratives it generates. Unless conventional historians make changes to mainstream methods, then access to funding and academic status are likely to move away to better-founded disciplines. In addition, being perceived as a laggard area in OS methods terms could become a disincentive for excellent new PhD entrants. By contrast, evolving ways to better demonstrate that historical research is reproducible and giving easier access to more of the evidence base are likely to strengthen conventional history being seen as reliably based, rather than only subjectively argued.

Enhancing the digitization of sources

In the past Howes (2023d) and others argued that mainstream historians (again, excluding digital history here) have largely chosen to just sit back and wait for years (or decades) for someone else (with presumptively infinite funding) to fully digitize the archives they use. Beyond endorsing digital archive work, they have not seen it as part of their own responsibilities to accomplish or even encourage the far greater and more rapid digitization of archival sources. Instead, however, Howes (2023b) suggests that:

- Historians develop the habit of collecting wherever possible greater amounts of 'show and tell' displays of personal archival sources to accompany each article, book or chapter. For instance, if this is feasible in a source archive then taking photographs of all the (key) documents adds an extra dimension to the 'archive of the archive' that every historian builds through their notes.
- In backing up their claims, historians should show examples of what archival records looks like, and advise new readers how to access and use an archive vital for their research. Annotated copies of the actual documents can help educate other readers (especially already skilled archival researchers) about the characteristics and limitations of the specific records or archived objects being drawn on.
- A network of committed-to-open historians in each field of work is needed, who can write-up advice notes specific to their type of archives and rapidly make archive images open to all in a cumulative manner. Extending such open source practices in an incremental way to mainstream history work can have an impact surprisingly quickly, not least in transforming doctoral programmes and skills, and thus rapidly changing early career researchers' practices. And as open practices spread, so 'laggards' in each sub-field can be inspired to join the movement, especially if history journals embrace open depositing of supporting materials wherever possible.

- External funders of history research are also likely to be induced to routinely foster new source digitization efforts, covering its costs and specifying an open access format as an integral part of providing financial support.

Of course, how much progress can feasibly be made depends on a large number of other empirical factors. The devil is always in the details. Historians may be reluctant to see funding diverted to fund depositing data and pre-registered materials if such new demands seem to cut into their scarce funding. Expensive search, travel and subsistence costs are often key factors that determine how much productive time scholars can spend with relevant records, especially in studying things in distant countries or hard to access archives. Yet open access practices can also be very valuable in enhancing *productive* archive time across the discipline – for instance, encouraging the use of computerized notebooks or PC journalling by historians that both enhance their projects' recording capabilities and are potentially more easily made open. The more such practices progress, the more likely it will be that a researcher starting on a new project can explore prior work carried out in the same archive; or can look at other similar work in different archives; or both. Scholars who have accessed other people's reproducible past efforts can learn what worked and what did not. They can better plan, scale and set up their research long before boarding a plane for distant shores or crossing the threshold of an archive that is new to them. Improving the collective efficiency of mainstream history has other benefits, as well – such as stimulating better PhD and student projects, from which a sustained interest in the discipline may in turn spring.

12.3 Cultural barriers around openness

A range of cultural barriers are still likely to impede open history advances, stemming in part from genuine scholarly anxieties and some partly legitimate concerns to keep the discipline in the shape it has previously been. We review four main objections made by historians, and consider how they might be rebutted.

(a) *Managing the production and evaluation of historical evidence in fields with both digitized and non-digitized sources is likely to remain problematic* for the foreseeable future. No full or immediate transition to OSS-type processes for qualitative work is likely to be possible in handling older periods. The supply of collateral documentation is less, there can be no access to interviews outside the contemporary period and the potential bias problems identified in Figure 11.1 are evident in most archives. Historians working with sources that are neither fully digital and reproducible on the one hand, nor fully non-digital or 'closed' resources on the other, will still know what counts as satisfactory evidence provision. But the large mass of scholars are likely to operate in an intermediate zone where conventions for guiding back-story

and provenance-proving information are vague, potentially contradictory, or seem to work only on the basis of case-by-case justifications. Authors, reviewers and editors may all diverge from each other on what counts as adequate historical evidence, let alone 'proof'.

Perhaps researchers in the middle ground must learn to live with diverse possible standards and develop skills in choosing how to give readers the maximum possible information, given the sources and archives being used. Historians clearly cannot wait for a fully digitized utopia to arrive when 'everything' is online, or could easily be made so if a demand arises. This happy state might be more likely now, given advances in data science/artificial intelligence (DSAI). But it is still likely to be realized only in the decades ahead. In the long interim, mainstream scholars cannot abdicate responsibility for making incremental or smaller-scale improvements that prove feasible.

As for the difficulties of handling mixtures of digital and non-digital sources, contemporary history has developed many effective ways for carefully managing their co-existence and choosing appropriate criteria of evidential soundness for each – and the field shows no narrowing of differently accessed sources. In many ways these criteria reflect historians' experiences of handling the co-existence of areas with plentiful and with very sparse data sources, so the problems involved are not really new ones.

(b) *Anxiety that open sources will 'crowd out' closed sources* in ways that bias the discipline's development has been voiced by some scholars. If digitized evidence sources foster reproducibility, and greater reproducibility makes grant funding easier to secure, will historians tilt towards projects with digital sources that confer greater OS comparability? At the level of citations, as journals increasingly expect all sources to be open access ones, historians worry that sources still behind paywalls (like qualitative research books) may be marginalized – assigning undue weight to digitized sources compared with undigitized ones. Could undigitized sources then be neglected or de-emphasized? Until all important sources have been digitized, is there a danger that almost accidental components of how documents are stored may have subtle but strong effects in narrowing or distorting analyses? However, offsetting such worries is the fact that modern historical scholarship is very competitive. Early career researchers, as much as principals with strong funding records, have strong scholarly incentives to seek out and develop new, undigitized sources that offer unique insights, bring them into debates and make them new candidates for digitization.

(c) Some traditionalist scholars have argued against digitization that it could *foster the growth of more inexpert historical analysis*, with amateur analysts generating more disinformation that looks sourced and plausible, which is harder for professional historians to rebut. Reduced barriers to viewing archive documents plus social media may encourage more ordinary readers to 'pile in' on what started out as just historical controversies. People untrained in archival skills may accidentally or innocently produce inaccurate interpretations that can be all too easily magnified for political or ideological reasons.

Yet this seems an outdated 'alchemical' stance, with scholars seeking to deny or control 'mass' access to knowledge development behind barriers kept artificially high. A great deal of historical material does bear on important political, social and ideological controversies. But so too do the materials of the STEMM disciplines – their knowledge is esoteric, potentially dangerous and societally salient. Scientists have worked to evolve successful coping strategies when the need arises to speedily refute misanalyses and to ensure that media dissemination of disinformation is restricted and not magnified. Indeed, part of the drive to OS in STEMM disciplines reflects concerns to combat disinformation and show the evidence to refute it. So too will conventional historians, as the discipline group already has done for misuses of paper archives, and as digital history has done for its area of the discipline group.

Conclusion

Historical enquiry can best prosper, and most effectively utilize, reconstruct and debate historical records, by professional historians being challenged quickly (Howes 2023c), and not just by other professional historians but by the many diverse 'publics' with differently setup expertise or well-developed (often very intensive) 'ordinary knowledge' of research fields (in Lindblom and Cohen (1979)'s terms). By becoming more open, historical archives and archival research can be reconstituted in a way that is more accessible, more transparent and becomes a greater resource for anyone seeking to 'do' history (Locke and Wright 2021). Nothing in this trajectory implies that the existing prestige and widespread general readership of qualitatively written history need be damaged, nor any part of the field's 'character' lost. Instead an incremental or piecemeal transition of conventional scholarship toward open history, proceeding in line with a wave of archive digitizations and open source efforts in digital history, can help broaden the discipline's range of methods and deepen researchers' ability to help each other. Optimists see a potential for sub-fields like the history of science and innovation or contemporary political history to create areas where digital and mainstream approaches might merge more in future. For instance, Blaxill (2023, p. 289) argues that:

> The first wave of historical computing may have faltered [but]…. political history could form a vanguard in a second coming of computational analysis in the wider discipline, not just because of the amenability of its sources and research questions to quantitative analysis, but also because it is entirely in keeping with our rich tradition of methodological innovation.

This has particular resonance in the strong linkages between the development of digital history and the growth of citizen-involving scholarship. 'Only by annihilating the distance between the production and consumption of

knowledge', the authors of *Digital Humanities* (Burdick et al. 2012, p. 82) argue, are digital humanists 'able to revitalize the cultural record in ways that involve citizens in the academic enterprise and bring the academy into the expanded public sphere' (Locke and Wright 2021, p. 1506 – quoting). We turn to this broader issue in Chapters 13 and 14.

Note

* This chapter was developed from an original online seminar held on 21 November 2023 with Dr Anton Howes that formed part of the LSE's input into the Civica project, held on 25 January 2023. The project was funded by the European Union's research and innovation programme Horizon 2020, under the chapter 'Science in Society'. The PowerPoint slides from the seminar and a video record of the presentation by Dr Howes and subsequent discussions are permanently free to download from the Civica Zenodo pages: https://zenodo.org/records/10217610. We are deeply grateful to Anton for leading the debate and kindly allowing us to replay some of his arguments here (albeit in our own fashion). We also thank all the participants at this session for their questions, comments and inputs.

  Anton Howes is a historian of innovation and Head of Innovation Research at The Entrepreneurs Network. His email newsletter, *Age of Invention*, is followed by over 25,000 subscribers – see https://perma.cc/7PRW-RK8S He is the authors of *Arts and Minds: How the Royal Society of Arts Changed a Nation* (2020).

References

Averett, Nancy (2022) 'Studying Volcanoes through Myths, Legends, and Other Unconventional Data', *Eos* blog, 22 April. https://perma.cc/Z79D-9Q9C

Berry, David M. and Fagerjord, Anders (2019) *Digital Humanities: Knowledge and Critique in a Digital Age*, London: Polity. $ https://www.wiley.com/en-gb/Digital+Humanities%3A+Knowledge+and+Critique+in+a+Digital+Age-p-9780745697666

Blaxill, Luke (2023) 'Why do historians ignore digital analysis? Bring on the Luddites', *Political Quarterly*, vol. 94, no. 2, pp. 279–289. https://doi.org/10.1111/1467-923X.13267

Burdick, Anne; Drucker, Johanna; Lunenfeld, Peter; Presner, Todd; and Schnapp, Jeffrey (2012) *Digital Humanities* (Cambridge, MA: MIT Press). https://direct.mit.edu/books/book/5346/Digital-Humanities

Burrows, Toby (2023) 'Reproducibility, verifiability, and computational historical research', *International Journal of Digital Humanities*, 31 August. https://doi.org/10.1007/s42803-023-00068-9

Clary-Lemon, Jennifer (2014) 'Archival research processes: a case for material methods', *Rhetoric Review*, vol. 33, no. 4, pp. 381–402. https://doi.org/10.1080/07350198.2014.946871

Finn, Margot (2019) 'Plan S and the history journal landscape: Royal Historical Society guidance paper', *Royal Historical Society*, Web page, 23 October. https://perma.cc/6DJF-SPBB

Graham, Shawn; Neha Gupta, Jolene Smith, Andreas Angourakis, Andrew Reinhard et al. (2020) *The Open Digital Archaeology Textbook*, Online OA textbook, 11 July. https://o-date.github.io/

Guldi, Jo and Armitage, David (2014) *The History Manifesto*, Cambridge: Cambridge University Press. https://doi.org/10.1017/9781139923880 or http://nrs.harvard.edu/urn-3:HUL.InstRepos:25620510

Henga, Yu Tse; Wagnerb, David T.; Barnesc, Christopher M.; and Guaranad, Cristiano L. (2018) 'Archival research: expanding the methodological toolkit in social psychology', *Journal of Experimental Social Psychology*, vol. 78, pp. 14–22. https://doi.org/10.1016/j.jesp.2018.04.012

Holm, P.; Marboe, A. H.; Poulsen, B.; and MacKenzie, B. (2010) 'Marine animal populations: a new look back in time', in McIntyre, A. (ed.) *Life in the World's Oceans: Diversity, Distribution and Abundance*, Wiley Blackwell, pp. 3–23. https://doi.org/10.1002/9781444325508.ch1

Howes, Anton. (2023a) 'Does history have a replication crisis?', *Age of Invention* blog, 29 August. https://perma.cc/6E9W-SDD8

Howes, Anton (2023b) 'Open history', *Age of Invention* blog, 22 September. https://perma.cc/LXD4-9EPX

Howes, Anton (2023c) 'How to be a public historian', *Age of Invention* blog, 28 November. https://perma.cc/E5Q7-84YK

Howes, Anton (2023d) 'Opening up history', online seminar, 21 November. https://zenodo.org/records/10217610

Knöchelmann, Marcel (2019) 'Open science in the humanities, or: open humanities?', Web page on www.mdpi.com/journal/publications, *Publications*, vol. 7, no. 65. https://doi.org/10.3390/publications7040065

Lindblom, Charles E. and Cohen, David K. (1979) *Useable Knowledge: Social Science and Social Problem Solving*, New Haven, CT: Yale University Press. **$** https://perma.cc/4YWQ-F8UE

Locke, Joseph L and Wright. Ben (2021) 'History can be open source: democratic dreams and the rise of digital history', *American Historical Review*, vol. 126, no. 4, pp. 1485–1511. https://doi.org/10.1093/ahr/rhab534

Mills, Albert J. and Mills, Jean Helms (2019) 'Archival research', in Cassell, Catherin; Cunliffe, Ann L.; and Grandy, Gina (eds) *The SAGE Handbook of Qualitative Business and Management Research Methods: Methods and Challenges* (vol. 2), London: Sage. **$** https://doi.org/10.4135/9781526430236.n3

Ogden, Lesley Evans (2016) 'New data from old treasures: whaling logbooks', *BioScience*, vol. 66, no. 7, p. 620. https://doi.org/10.1093/biosci/biw067

Rosenzweig, Roy (2006) 'Can history be open source? Wikipedia and the future of the past', *Journal of American History*, vol. 93, no. 1, pp. 117–146. **$** https://doi.org/10.2307/4486062

Schmidt, Laura (2016) 'Using archives: a guide to effective research', Society of American Archivists, web page, 26 March. https://www2.archivists.org/usingarchives

Shuttleworth, Sally (2015) 'Old weather: citizen scientists in the 19th and 21st centuries', *Science Museum Group Journal*, vol. 3, pp. 156–176. https://perma.cc/ARJ4-FTYW

Vanden Daelen, Veerle; Edmond, Jennifer; Links, Petra; Priddy, Mike; Reijnhoudt, Linda et al. (2017) 'La publication durable digitale des guides d'archives de l'histoire du 20ème siècle', *HAL* web page, 10 November. HAL Id: hal-01632366. https://inria.hal.science/hal-01632366

13. Doing 'citizen social science'

Denisa Kostovicova, Patrick Dunleavy and Timothy Monteath

Most open science (OS) advocacy has focused chiefly on making research available to and reproducible by other researchers (Hecker, Garbe and Bonn 2018b; Hecker et al. 2018). But in the social sciences numerous different theories and methods have also advocated opening up the research process to include those being studied (its 'subjects'), the groups, communities or individuals participating – even though 'openness' is still seen in different ways across disciplines (Laakso 2022). Although the same is also true of the science, technology, engineering, mathematics and medicine (STEMM) disciplines (Robinson et al. 2018), people involved in (or who 'undergo') social science research projects have a different, special knowledge position vis-à-vis researchers. At its best, new research may represent their situation to the people studied in ways that are new to them – and maybe might even help empower them to make positive changes for themselves and their communities. But equally, research 'subjects' may have their lives, views and behaviours interpreted and presented to others in ways that they might not accept. The knowledge or skills that participants could potentially offer to the research may be overlooked or misinterpreted. At worst, these perspectives stemming from research may be disturbing, demeaning or dangerous for the people studied, as the 'decolonization' movement's critique of Western-framed research has long stressed (Thambinathan and Anne 2021).*

Where individuals, groups or communities actively took part in the research in some way, their lives can especially be changed (for better, or worse) when social research is published. Accordingly, many social scientists acknowledge that a special duty of care arises and that research 'subjects' themselves are the best judges and guardians of their own interests. They are 'stakeholders' in academic work as much as (or more than) the bodies or reviewers that assess projects and give out grants, or the other organizations and interests that finance academic work. Hence, they should be involved in how research

How to cite this book chapter:

Kostovicova, Denisa; Dunleavy, Patrick; and Monteath, Timothy (2026) 'Doing "citizen social science"', in Dunleavy, Patrick and Monteath, Timothy (2026) *Doing Open Social Science: A Guide for Researchers.* London: LSE Press, pp. 299–322. https://doi.org/10.31389/lsepress.dos.m

is organized, focused, implemented and presented publicly – and (potentially) also in how results and findings are analysed. This is the core of citizen social science (CSS).

We define CSS broadly here to cover all efforts to consult and involve more fully the views of the 'subjects' or participants or respondents in how research is undertaken, and to take account of and share fully the interests and voices of all affected non-professional stakeholders. Any scholarly work seeking to 'co-produce' research with the groups, communities or individuals being studied, or to encourage their active participation and sharing in decisions about approaches or interpretations, counts as CSS in this wide definition. We begin by briefly reviewing the multiple approaches that meet these criteria across many different disciplines and theoretical approaches (Section 13.1). Next we consider the key arguments about why it matters to make a change from conventional academic monopolies over research decisions. Citizen involvement and co-creation of research offers important benefits in terms of enhancing the accuracy, timeliness, interpretative depth and social impacts of academic work (Section 13.2). We next review some practical ways in which CSS projects might be made more open. The two approaches, open social science (OSS) and CSS, have substantial potential overlaps and synergies, yet also pull in somewhat different directions that need careful management. However, some tensions and dilemmas (choices where both options have drawbacks) remain in implementing more open CSS (Section 13.3).

13.1 Citizen social science approaches

Increasing the involvement of research 'subjects' has been used across many social science disciplines, including: 'anthropology, communication, education, social and economic geography, political science, psychology, sociology, management, public policy and administration, urban studies and planning, and science and technology studies' (Tauginienė et al. 2020, p. 4). It has also been deployed in some contemporary history work, notably in projects translating masses of historical documents or in oral history projects (Vanden et al. 2017).

Early ideas around CSS:

> converged around notions of mass participation and data collection at scale, with members of the public assisting with research, and recording their beliefs and opinions, generating large volumes of data (Housley et al. 2014).

Any form of crowdsourcing people to undertake social science tasks that is academically directed would count within our definition (Vohland et al. 2021). In any citizen science project (whatever the discipline) how groups, volunteers or other interlocutors are recruited and managed over the study's

life cycle remain important issues (Kullenberg and Kasperowski 2016; Bonney et al. 2016).

Some observers have also argued in the past that CSS projects 'which experiment with the idea of citizen science, are still hard to find' (Heiss and Matthes 2017, p. 24). However, the range of studies with active participation by the people or groups being studied has extended over time, and the diversity and roles of 'interlocutors' have deepened, frequently following paths explored first and earlier on in STEMM disciplines:

> The social sciences can learn from citizen science about new forms of mobilisation, technological platforms, as well as socio-technical skills … [*Yet*] the social sciences have more to offer to citizen science than bridging and mediating [*STEMM 'citizen science' studies*], and citizen social science has many more facets than merely mimicking natural science approaches … Based on the long tradition of participatory approaches in the fields of participatory action research (PAR) and the co-production of knowledge, tools and concepts in the social sciences are available to impart both scientific rigour and inclusivity to knowledge production (Albert et al. 2021, p. 125). [*Our inserts in italics.*]

In addition, the 'citizen humanities' tradition has contributed many innovative ideas to social scientists (Heinisch et al. 2021).

Is 'citizen open science' then just the same as participatory research? According to Stuart and Maynard (2022, pp. 16–17), participatory research includes all the elements shown in green-shaded boxes in Figure 13.1. All forms of 'PAR … draw on a model of community organizing that supports the capacity and expertise of people experiencing issues first-hand'. (Also see

Figure 13.1: Older forms of citizen social science (green cells) and newer forms falling within our broad definition (white cells)

Participatory action research	Community-based research	User-led research
Participatory organizational research	Participatory arts-based research	Participatory rural appraisal
Citizen science	Participatory evaluation	Critical pedagogy
Rights-based research	Co-produced research	Community development
Co-designed research	Co-creation of research	Action research
Post-conflict/ reconciliation research	Grounded experiments	Deliberative democracy

Source: Authors, drawing on Albert et al. (2021).

Friere 2005.) Participatory research more generally seeks to give full expression to the autonomous voices of participants and to uncover and reveal their firsthand (historical and contemporaneous) experiences, meanings, deliberative processes and strategies. Some exponents and critics alike suggest that PAR is solely or primarily about achieving insights, and is not focused on co-ordinated action or collective ways forward.

However, in addition the white shaded cells show other important elements of our deliberately far broader definition of CSS. Co-produced research has long been important in community development, seeking to involve communities holistically (rather than solely as individuals) in framing what is to be researched and what potential policy solutions should be developed or adopted. Co-production developed first in studies conducted and focused 'overseas', such as anthropology and international development and aid research, where researchers were keen to avoid past 'colonialist' cultural biases (Lennette 2022a, 2022b). However, the idea of co-producing research with those affected later spread into medical research, social policy, public administration and 'public sociology' (Housley et al. 2014), especially where finding solutions for problems proved that more nuanced or responsive/co-operative approaches were needed. Making too much of distinctions between PAR and co-produced research seems to us to be a case of the 'narcissism of small differences'. Similarly, 'co-design (also referred to as co-creation) is more of an umbrella term to describe different processes of involving multiple partners in the development and/or provision of interventions' (Hidalgo et al. 2021).

Lastly, there is a set of CSS approaches at the bottom of Figure 13.1 that is clearly more distant from an insight-only PAR focus. Here researchers are not just finding things out collaboratively with research participants, but also taking action or seeking to help others to organize actions on the basis of new knowledge. These projects include evaluating what works best in action research; developing well-grounded 'experiments' with groups and communities themselves; seeking to create foundations for bringing divided societies together after traumatic events (see below); or using deliberative democracy methods to try and reach agreement between opposing interests on community decisions or public policy choices. But note too that in all these cases achieving cognitive gains and better understandings of social meanings and causal processes is the prelude to achieving any worthwhile forms of corrective changes or improvements.

Researchers who specialize in the CSS area acknowledge that involving those being studied can take place at different points along the continuum of the research process: from identifying the research question, collaboration and co-participation in the research process, analysis, to formulation and sharing of the data and findings (Stuart and Maynard 2022, pp. 15–16). CSS also entails being aware of power inequalities between the researcher and the researched, and recalibrating power in the research relationship: from extraction to co-production of knowledge (Kostovicova and Knott 2022).

At the same time some of the most self-conscious CSS practitioners focus primarily on very 'core' or intensive examples of the approach. For instance, Kostovicova's own key CSS-informed work was undertaken to aid the 'healing' process in traumatic post-conflict situations. The research was explicitly designed to help citizens, including victims, human rights activists, teachers, journalists, war veterans and others from different ethnic groups better understand their efforts to establish a common record of war crimes and human rights violations across polarized and traumatized communities in the Balkan states following the highly divisive and conflictual breakup of Yugoslavia in the 1990s. The many war crimes and excesses of violence which occurred then created new sources of grievances and suspicions between communities with long historical memories of previous conflicts. Getting to a common historical record of what happened was seen as a foundational step to bring communities closer to some form of reconciliation (Kostovicova 2023). This is an especially important and challenging form of CSS and illustrates very well the importance of research components going beyond just compiling an isolated academic record or generating new theoretical insights.

Despite both the broadening out of the CSS approach and the 'deepening' of collaborative methods in its core areas, some substantial qualitative research fields still seem to be variants of or very close to CSS, but are rarely included in the lists of relevant areas. An example is 'deliberative democracy' (DD) that we have included in Figure 13.1 above. Perhaps not coincidentally, this has also been a field where an awareness of OS approaches seems surprisingly lacking. For instance, the recent, authoritative Ercan et al. (2023) *Research Methods in Deliberative Democracy* sets out to 'offer thirty-one different methods that scholars use for theorizing, measuring, exploring, or applying deliberative democracy' according to its book blurb. Graham Smith noted in the book's foreword that:

> Deliberative democracy requires openness and respect towards pluralism: to different ways of seeing the world and in the collective search for meaning … [And] the same holds true for the study of deliberation. We require openness and respect towards different ways of seeing our object of analysis (Ercan et al. 2023).

Furthermore, 'the basic qualities of the deliberative mode of interaction' are sometimes listed as 'respect, justification-giving, reciprocity, *openness*, attentive listening, etc'. So we might expect that the DD field above all would be one that takes OS considerations seriously. Yet no awareness of OS imperatives is shown throughout, except briefly in one chapter on survey methods.[1] The editors' preface notes that 'Our aim in publishing this volume open access is to "practice what we preach" ' (p. 19); but across 502 pages an open resource outside the book itself is mentioned only once.[2]

The still evolving breadth and depth of CSS approaches was captured well by Prainsack (2014, Table 7). She suggested seven main groups of factors, covering:

- how work is co-ordinated and who decides how key agendas, stages and processes are defined;
- who participates and on what terms (e.g. as general versus expert respondents, as isolated individuals or as connected groups, or with training/briefing or without);
- whether the work meshes with already-established communities, or their collective leaders or organization;
- how research is evaluated (by the project team or funders only, or by stakeholders or the research 'subjects' themselves);
- how open the project is (and how much it takes OSS concerns seriously);
- how the project is funded, controlled and impelled, and whether outside non-academic interests are involved; and
- where the research physically takes place (e.g. online only; 'fieldwork' only in relevant localities, with analysis located at a distance; or with multiple stages on-site (including analysis and dissemination).

The more that researchers share authority and control of the agenda on each of these issues with the people who are being studied, the more authentic citizens' participation is. For instance, deciding what the results are, what will be done with them, who can access datasets and who owns any intellectual property generated by research have conventionally been determined solely by researchers– but can be opened up for community or group decision also. Similarly, where the people being researched form a community before the research (and after), they might be expected to be better able to share in such decisions, and to guide how research processes evolve more. However, these criteria are chiefly apt ones for older and 'reactive' research approaches that rely on eliciting information directly from respondents (such as surveys, interviews, case studies and field research). They may be less relevant for studies that in part seek to elicit 'cultural' meanings from the analysis of 'objective' text or documentary data or behaviours (such as Leek, Bischl and Freier 2024).

13.2 Why does citizen open science matter?

There are four main reasons why CSS work adds new and distinctive elements to social science research in ways that are more fundamental than for most STEMM areas, namely that research monopolized by academics alone risks wasteful misalignments and mistakes arising; the 'ownership' of research may be important for its perceived success; co-produced research can potentially increase sustainable impacts by leading to more responsive policies better suited to addressing problems; and, lastly, CSS opens up important new areas of methods innovation.

Misalignment of research

Most applied social science work is motivated by a concern to improve societies 'on the ground'. This implies that the research undertaken should mesh closely with both the public policy priorities involved and research participants' own perspectives and priorities. And as we noted above, this is increasingly the language and requirements of grants too. Kostovicova's (2023) analysis sheds a lot of light on how research misalignment can be wasteful, producing work that is seen as inaccurate by those studied and secures little acceptance. For instance, in 2018 the European Commission resolved to create the bases for 'A credible enlargement perspective for and enhanced EU engagement with the Western Balkans.' The European Union (EU) sought to address the problem of elusive reconciliation between opposing ethnic communities in the region, with academic research being a potentially important foundation stone for this effort.

> Together with improved regional cooperation, further efforts towards reconciliation are crucial to firmly anchor peace and ensure lasting stability in the region. The wounds of the 1990s still need time to heal. But 25 years on, it is time to address open issues with new vigour … before accession to the EU. (European Commission 2018).

Yet the communities involved often (correctly) associated academic involvement with past 'extractive' research practices and the obscure ethics of research relationships lacking any transparency for and sharing of knowledge with the groups and communities being researched (Schroeder et al. 2015). This has led to the rejection of reconciliation both as a concept and a policy among communities who seek normalization in the wake of mass atrocity violence (Kostovicova 2023, p. 31).

Kostovicova quotes one of the interlocutors in her different CSS project commenting on the EU Commission's announcement: 'When I saw "reconciliation" in the EU strategy towards Bosnia, I felt sick' [Personal communication, September 2018]. Kostovicova also cites a comment that emerged strongly from a workshop held by the JUSTINT programme (2024), supported by the European Research Council (ERC), which she directs. A workshop in Zagreb, Croatia, gathered civil society activists working on projects addressing legacies of wars in Bosnia and Herzegovina, Croatia, Kosovo and Serbia. The workshop was designed to apply the CSS principle of sharing the research findings with civil society stakeholders, inviting knowledge exchange and co-producing policy recommendations through a joint effort of academics and stakeholders (Sokolić, Vico and Kostovicova 2023). Reflecting on the event, a workshop participant observed: 'Researchers always use us to obtain information. They never invite us to share and discuss their findings.'

Social science researchers have alienated many local communities during their research process, closing off sources of local knowledge. A Serbian civil

society activist in Belgrade, Serbia, argued that the conventional monopolistic academic control over research definition meant that multiple important avenues and sites of investigation go unappreciated:

> In large part, the academic community does not understand what it can learn and has learnt very little. I must say, few exceptions notwithstanding, there are few of those who know how to interpret, really interpret on the basis of what they hear from people…. That is a big problem. That's why people are desperate when they see academics; there's an avalanche of them. (Interviewee, 24 September 2014, cited in Kostovicova and Knott 2022, pp. 66–67).

Like participatory research more widely, a key rationale for using CSS methods is that people and communities 'on the ground' are the best 'knowers' of their own social worlds and issues, while outsiders are often limited in their ability to fully understand the issues of others. The challenge is pronounced in cases involving cultural, social, political and other differences, particularly when there are conflicts between divergent groups or interests. By contrast, 'Co-research produces "better questions, approaches and understandings"' (Stuart and Maynard 2022, pp. 14–15) – it may avoid significant resource waste carried out abortively, which either is based on misconstructions of poorly appreciated, complex situations or lack of awareness about previous applied research and previous efforts to improve matters. Finally, involving research 'subjects' in how research is defined is an effective way of counteracting projects deformed by or overly confined within single-discipline siloes, producing only work that fails to be aware of other academic and ordinary knowledge interpretations.

Ownership of research

In social science the 'researchability' of significant issues may also be severely compromised by extractive modes of working. In many geographical contexts and spheres of social life people's past adverse experiences with solely academic-defined projects can mean that researchers are denied access to interview people and communities, and cannot discuss their lives and situations with them in order to understand what issues are involved. For instance, indigenous peoples have suffered considerably over centuries from western institutes and universities undertaking work in a 'fieldwork visits' mode, with short periods where the maximum amount of local artefacts and other knowledge were 'hoovered' up intensively. All the resources gathered were then taken back to the home base and analysed there in isolation from the groups and communities involved. (For the example of genomic data in STEMM research, see Hudson et al. 2020.) Later the analysis results get published behind scholarly paywalls, fenced off from any response or inspection by those studied.

Until recently, western researchers were (and perhaps still are) often ignorant about indigenous communities' different ways of being and understanding of the world, misconstruing problems and proposing bad solutions (Schroeder et al. 2019). Even apparently neutral approaches, like the FAIR (findable, accessible, interoperable, reusable) principles, can need to be adapted in working with indigenous communities (Carroll et al. 2021). So groups and communities (especially vulnerable groups and indigenous peoples) have increasingly set conditions for co-operation (Liabo, Ingold and Roberts 2017) and some groups ask to vet research study protocols in detail. Project teams that fail to take co-production of knowledge seriously are likely to be locked out of many research sites or opportunities, and to damage access for researchers more widely, even for teams with better practices.

Similarly, researching in post-conflict situations, such as the aftermath of wars, civil wars, massacres, terrorism and other catastrophic events, is likely to mean that researchers must be accepted not just as neutral and well-informed by different 'sides', but also as well-informed and empathetic teams willing to work with their 'subjects'. Participants in such research need to be confident that their traumatic experiences will not be reactivated by insensitive or intrusive activities or conclusions that risk renewing conflicts on a micro-scale. Where past events have been violent, and people have lost families, relatives or friends, research can stir up deeply felt emotions. Kostovicova noted how at a Sarajevo event a research participant reacted to her sharing the findings about the quality of civil society deliberations about war legacies in the Balkans in which a research participant had been involved. Reflecting on the benefits of regional over national deliberations in an aftermath of a regional war that Kostovicova shared, the participant observed: 'That finding of yours is very important to me', adding that it was motivating for him to continue his struggle for truth and recognition for suffering (participant in the REKOM: XI Forum, quoted in 2018 in Kostovicova and Knott 2022, p. 65). In this and similar cases, research is partly redemptive, seeking to get all sides to find a dignified and agreed way to overcome divisions and find an acceptable form of interethnic accommodation. Here and in some participatory research also, emancipation from entrenched attitudes or mores is 'not just a coincidental by-product, but a planned outcome of the research process' (Schubotz 2020).

Research impact

A great deal of social science research is not intended to be immediately 'public-facing'. The theoretical issues motivating a research question can appear esoteric in conception, or a deceptively simple research question can be executed in ways that non-academics find hard to understand. Moreover, general readers are unlikely to ever encounter the primary outputs of research in academic journals and books, even though open access articles are now

increasingly dominant. If they do, they may find obscure 'technical' language and 'jargon' difficult to parse.

However, assuming that audiences outside the academy are uninterested or unable to understand research is harmful and outdated. Indeed, this is far from the case for most applied work, especially in public policy and technological fields, where elite or lay audiences with ordinary knowledge may have a far deeper specialist understanding of some key issues being analysed than even well-educated researchers.

> Digital media and open-source databases, for example, have broadened the scope and number of people who now contribute to science, and accelerated the pace at which they can do this. The threshold for 'doing science' is arguably lower than ever (Prainsack 2014).

For instance, it is a truism that in medicine some patients suffering from conditions can have a great understanding of symptoms, treatments or pain mitigation, perhaps more than even the best trained doctors. Patient groups and non-governmental organizations (NGOs) specializing in one condition have developed as key interlocutors with medical professionals about best practice in their areas of expertise (Pii Schou, Piil and Jarden 2019). So applied researchers who want to have impact must take seriously the job of convincing demanding and potentially sceptical audiences outside academia to accept their research approach, methods and evaluations.

In the open access era, researchers working in highly contested areas (e.g. those that have related closely to historical, political, ideological or cultural controversies) must also expect one set of interests or 'sides' in a given conflict to query, criticize or attack, and potentially misconstrue their major research results, perhaps even making them the focus of misinformation surges or critical social media campaigns. The days are long past of authors worrying only about one or two potentially hostile book reviews by fellow academics (usually published two or three years later, so dulled with an 'old news' patina). Modern applied research of interest to a broad audience in areas of controversy is no sooner presented at a public session (like a seminar or conference) or posted out (in micro-summary format) via social media messages, than it can begin attracting hostile feedback.

CSS helps mitigate and manage potential problems in numerous ways. Researchers who involve all the interests and different types of people affected by their work are much less likely to be surprised by potential 'attack lines' catching them unawares. In addition to getting peer reviewed, a key defence, they will also be able to plan in rebuttals and defences against potential critiques. Involving research 'subjects' in helping to define the orientation, methods and conduct of the study, and also developing the analysis of findings with them, provides a third key line of defence. Intensive contacts with research participants also mean that research teams know how best to phrase findings in ways that command agreement and endorsement from relevant groups

and communities, what language is best understood and where sensitivities need to be protected. Finally, open CSS approaches developed with individuals or groups being studied and relevant communities optimize the chances that applied social science findings may be acted upon by the communities themselves or by government actors seeking practical and (ideally) consensual policy changes.

Methods innovation

CSS approaches have already strongly contributed to opening up applied social science methods and to a revaluation of well-grounded qualitative research. Historically, CSS approaches have also helped to pioneer new modes of inclusion, participation and mobilization of research participants who want to be engaged. In turn, this can help bring about improved and relevant insights and connections for action. In applied work:

> the 'ordinary knowledge' of research participants has often fed directly into debates about the politics of methods, giving rise to questions around what counts as data, who can collect it, and how it can be used. (Albert et al. 2021).

These changes: 'reflect the blurring of boundaries between research objects, subjects, and researchers directly engaged in the everyday realities of [*applied social*] science and society' (see also Figure 9.1). In particular, the consolidation of the public engagement agenda with wider efforts to enhance the impact of *all* academic research has fuelled recent trends towards more systematic qualitative analysis so that researchers can more fully understand the meanings of any social behaviour changes. Included here have been randomized control trials (RCTs), where quantitative work is informed by rigorous qualitative work with the people and groups affected, so that researchers can more fully understand the meanings of any social behaviour changes. On questions like what effect a basic income scheme has on people's welfare and behaviour, or how the public or corporations might be best persuaded to make changes in their behaviours that mitigate global warming and climate change, repeated studies have shown that multiple different social processes are involved that require improvements in learning from participants.

What Bastow, Dunleavy and Tinkler (2014, Chapter 9) term the 'dynamic knowledge inventory' of advanced industrial societies operates in complex ways, and in applied fields especially there are many salient interactions and loop backs between 'lay' and 'expert' knowledge-holders (see Chapter 16). Improving citizens' and organizations' active participation in research about them generates new scientific knowledge and understanding, opens up spaces for more citizen-generated data for the social sciences, improves the chances that research can meaningfully benefit those who participate in it, and accentuates the (positive) societal impacts of the social sciences.

13.3 Achieving greater openness with citizen social science

To help enhance openness, CSS confronts similar problems to other aspects of qualitative research (considered in Chapters 9–12), with a mixed picture of pros and cons. In three areas there are helpful, strong synergies between CSS and OS imperatives:

- pre-planning thoroughly, which is also vital in any CSS project;
- recording fully what happens in inherently unrepeatable research events; and
- maintaining clarity about what is included or excluded from the research.

However, the two approaches pull in rather different directions in three other major areas:

- research boundaries and methods consistency under co-production;
- managing data release versus participants' confidentiality; and
- reconciling academic autonomy with authentic co-production.

Pre-planning

Earlier chapters demonstrated that for open research it is especially important at the beginning of projects to look ahead to the whole set of tasks ahead. The incentives to do pre-planning carefully are also especially high in CSS research. Co-production work is undeniably more complex (and usually time-consuming) to set up than conventional, solely academic research. So, additional attention and time should be given early on to pre-registering initial core expectations, research questions and the methods to be used. In almost all CSS work these three aspects also have to be 'negotiated' early on with research participants themselves. In securing citizen participants' agreement researchers will also have to summarize in accessible ways the earlier academic literature, any already public interviews or materials, and relevant media/public debates, interpretations and controversies.

In the setting up phase it is important to do an exploratory 'quick dive' with your planned set of participants (the 'research subjects') as early as possible, to surface preliminary concerns amongst potential interviewees or focus group members, or amongst group leaders whose agreement will be needed. Key here will be to understand participants' and groups' autonomous concepts and vocabularies, and their likely preferences and expectations about the proposed research. Testing researchers' understandings and recording any corrections made to misconceptions will be especially useful for reproducibility and follow-on researchers devising related or similar CSS projects. At

this exploratory stage of 'updating priors' it will be key not just to 'validate' research assumptions incorporated in the funding bid, but also to surface additional questions or dimensions of analysis that can be included from the outset, rather than being late-stage amendments to the research plan. Doing this using cascade interview approaches or waterfall methods may be harder, since new themes and issues emerge as the work proceeds. But covering all the main angles or different groups involved during exploratory work can somewhat reduce risks here.

Recording unrepeatable events

A great deal of applied social science work, and almost all CSS studies, rely a good deal on research events that can never be exactly redone – such as focus groups, citizens' juries or conversations, interviews, community participation meetings, etc. In CSS studies these processes are also only partially defined by the researchers and instead are co-produced in varying degrees and multiple different formats with the individuals and collectivities being studied. Their views, preferences and needs are built into the study definition and cannot be the same across time or with different sets of participants. Consequently, extra difficulties arise in CSS around how to maximize assurance and trust in the evidence produced, especially if there are restrictions on how far the records of these events can be opened up. Some forms of co-produced knowledge build in permissions from all participants at the outset for the records of discussions to be made public straightaway, albeit normally with some conditions. For instance, deliberative democracy sessions like citizens' juries are routinely filmed and recorded for near-immediate public release. Here, both research participants themselves and later users of the research can easily check on how the event was set up, conducted, analysed and interpreted by the original researchers. But other forms of participatory research (such as evidence gathering in a post-traumatic context) can be very sensitive experiences for the individuals involved. They may object to recording and disclosure, or being recorded may strongly inhibit the frankness and value of what participants say.

Assuming that permissions for full or anonymized disclosure can be obtained (or limited release to other researchers only), CSS participant discussions are typically less researcher-controlled. They may need more sophisticated recording – especially in fast-moving focus groups and group discussions where multiple recording devices are always needed. Interviews are not so problematic with the standard two recording devices per interview recommended in Chapter 9. Remotely filmed participant videos or storytelling are also recommended there, but only one recording is feasible then, with some potential losses in openness for CSS group settings. In CSS work single sources may more often need to be copied and archived reliably. In some studies videoing may be potentially more problematic than audio-only,

because of sensitive subject matter (as with recent post-conflict studies and potential risks to participants). Other recording restrictions may reflect group cultural sensitivities – e.g., Australian First Nations peoples may not agree to the showing of photographs or videos of deceased community members (AIATSIS 2024a, 2024b).

In Chapter 9 we also recommend using remote video via Zoom, Teams or similar software in reaching spatially dispersed interviewees and group-based events that are otherwise very hard to schedule. But in CSS research remote Zoom etc. strategies need special care and planning. They may work best when relatively individual activities are involved and participants control what gets done, for example with participants recording their own stories in their own way rather than following through an interview schedule (e.g. Marzi 2022).

Research boundaries and methods consistency under co-production

We turn next to two main tensions between CSS work and openness in research, with the first concerning boundary-setting and methods stability. In general, open research aims to choose explicit inclusion/exclusion criteria for what materials are ruled into or out of the study frame at the start of research and then stick to these decisions as the process unfolds. We noted in Chapter 3 some important tensions with the whole exploratory research and exploratory data analysis agendas. Where research is co-produced, these potential conflicts are magnified because control over where discussions or explanations go rests more with research participants, and so inclusion/exclusion and direction of analysis decisions are no longer an academic monopoly. Established methods for boundary setting – like the key role of the facilitator in keeping focus groups on topic and preventing any one person being overly dominant, or the skill of interviewers in steering respondents away from digressions they may insist on making – are clearly relevant still. Yet researchers cannot (and normally will not want to) deny a whole group that is formally co-producing a piece of research from going in an unanticipated direction.

Of course, the trick here is not to get blindsided in the first place, by undertaking thorough early preparations, piloting and in-depth exploratory work early on (as discussed above). Yet no advanced preparation can be 100% in anticipating where sets of participants or pre-existing groups will want to steer events and projects. Where changes are made to the planned research, or trajectories shift unexpectedly, the documenting of how this happened needs to show that participants and not the research team took the lead. It may also be helpful to separately identify more routine or incremental types of data excluded – for example, which boundaries were participants' decisions on excluding materials or related sub-topics and which were the researchers' decisions. As with other types of analysis (like systematic review, see above) it is useful to archive near-miss or 'tricky' boundary choices, both those set initially and any decisions about exclusions of topics in later reanalysis.

Managing data release versus participant confidentiality

In CSS work it is of course essential to make sure that only data that respondents have clearly consented to is recorded and stored, and personal data needs to be safely and separately stored from the data to be retained (see Chapter 7; Purdam 2014). Some aspects of the post-research phases may need to be judged carefully. The intellectual property rights for sensitive CSS research may well be assigned to researchers' employing institutions, perhaps as part of securing ethical approval and to ensure that the university or laboratory involved can fully control how it may be used, which could cause complications in getting project data deposited in reproducible ways. The original researchers (individually or collectively) may wish to 'double tap' their completed data in later research, for instance in order to make comparisons, undertake extension work, or remodel analysis in the light of later gains in understanding, so it can be important to safeguard this option, before and after finally depositing data. Finally, it is wise to remember that stewardship of the deposited data rests with the repository involved, who will have responsibility for any future changes or disputes about it. In all these ways, researchers need to pay extra-special attention to how deposit processes are undertaken and how data reuse conditions are set.

In volatile contexts, sharing all data and opting for full transparency may have adverse consequences in terms of research participants potentially becoming visible and suffering harm from being recognized or having their opinions broadcast. The growth of open access, open data and social media has also broadened the range of people and organizations who may gain access to full data from CSS studies, from just academic researchers in the past to a wider range of interested parties. Even where participants have consented to identifiable or anonymized publishing of data, a separate responsibility rests on researchers to ensure that no harm can arise from openness. Bear in mind that changes in social and political situations occur constantly, so that information that seems uncontentious now might well become controversial in future if circumstances change. In general, ensuring the rights and welfare of respondents via appropriate safeguard conditions should always trump the benefits of being open to later researchers exploring how far the study's original results are reproducible.

Reconciling academic authenticity and the authenticity of co-production

We noted above that a great deal of CSS is driven by a concern to empower people, groups or communities hitherto excluded or under-represented in academic research. CSS here is about 'redressing the balance' and committed researchers may only seek to give a voice to the previously voiceless. Other CSS work seeks consensual understandings of issues – but that might be achieved in different ways. Some approaches emphasize 'agonism' (a philosophical

approach where understanding is seen as emerging primarily through a clash of opposites). Others emphasize deliberation where rationalistic decisions are taken on the basis of evidence and logical and even scientific argument. All these approaches may have issues over handling 'silent' voices, potentially producing interpretations skewed in favour of the kinds of people willing to become research participants as opposed to non-participants. Set against that, however, is the value of the researchers' care to achieve inclusiveness, plus the potential improving or even 'therapeutic' benefits of introducing 'subjects' to new theoretical perspectives and other ways of seeing their world – e.g. human security) (Kostovicova et al. 2012)

A particular difficulty can arise in handling potentially sensitive data in CSS efforts, especially those that seek reconciliation or agreement between divided sections of a community, or where the study seeks to surface 'constructive' or consensual contributions, or identify areas of potential agreement or empathy between people on opposing 'sides' of a community or of an issue. At one pole, some research participants accepting 'moderate' or empathetic views may be seen as unacceptable or 'a betrayal' by more polarized people in a given community, attracting adverse consequences for them unless they are anonymized (Kostovicova 2023). At another pole, researchers may face a dilemma when strongly voiced divisive or regressive views are expressed (or re-expressed) in study forums. Do they respect and include 'destructive' or unreconcilable participant contributions in ways that may wreck the project, or instead exclude them at the expense of 'sanitizing' issues in a misleading or artefactual way? 'Citizen open social science' research may entail maintaining boundaries that impede both a degree of openness and more purist understandings of 'autonomous' participation. Are such researcher-imposed constraints that serious, however, in the digital age? Arguably, 'netizens' now have multiple opportunities to develop autonomous views and share them with others, so that later researchers are not confined for information to looking at earlier CSS work. Even on the web and internet, liberal democratic states impose controls on hate speech and certain types of content, making it potentially less 'authentic' but also less harmful to others. The presentation of views in short-form and comment-like ways (akin to many social media inputs) can still be helpful for enhancing reproducibility, even if respondents are anonymized.

Conclusions

Many (perhaps most) advocates of CSS and citizen humanities (Heinisch et al. 2021) have seen it as part of a wider effort at empowerment of hitherto disadvantaged groups to help them better secure social/political/economic change and social justice; or as a way of enhancing societal agreement between initially opposing groups and achieving reconciliation; or as an avenue for advancing policy agreement and ways forward via increasing societal deliberation rather than contestation. Yet the core case for CSS has been that it improves

the quality of research itself, an effort that additional openness takes further. Nor is the case for open CSS confined solely to normatively impelled research. Rather, open CSS approaches have all shared a key insight, that achieving progress in applied social science (and even in many fields of applied STEMM) will entail re-evaluating the role and usefulness of 'local' known information, and the ordinary knowledge of non-expert actors. Broadening the involvement of research participants in the production, analysis and evaluation of all forms of professional social science inquiry can confer demonstrable gains for the quality, usefulness and impact of all academic work, especially in applied fields. Eliciting the greater co-operation and involvement of research participants in how analyses are surfaced, recognized, developed, shared and acted on, once formulated, can entail some challenges in simultaneously enhancing the useability of research for later scholars and its reproducibility. These issues can be handled in ever more sophisticated ways with evolving ethical norms as open social science develops and awareness of both potential pitfalls and the research gains from greater transparency and accountability grow.

Notes

* This chapter was developed from an online seminar held on 7 November 2023 led by Professor Denisa Kostovicova, forming part of the LSE's input into the Civica project. The project was funded by the European Union's research and innovation programme Horizon 2020, under the chapter 'Science in Society'. The PowerPoint slides from the seminar and a video record of the presentation by Denisa Kostovicova and subsequent discussions are permanently free to download from the Civica Zenodo pages: https://zenodo.org/records/10082681 We thank all the participants at this session for their questions, comments and inputs. Denisa Kostovicova gratefully acknowledges the funding by the European Research Council (ERC) Consolidator Grant, Justice Interactions and Peacebuilding: From Static to Dynamic Discourses across National, Ethnic, Gender and Age Groups (JUSTINT), no. 772354, which supported the research on which she draws in this chapter.

[1] The phrases 'open science', reproducibility, repository, 'open social science' or 'citizen social science', 'open access' publishing and 'pre-registration' appear nowhere in the *Research Methods in Deliberative Democracy* book. There is only one developed but brief mention of replication, in the experiments chapter (Grönlund and Kaisa Herne, 2022, Chapter 11), which notes that: "Replication of research is not easily available through naturally occurring data" (p. 165). There are four passing referrals elsewhere. Clearly, while the work is being undertaken, deliberative democracy (DD) practices are designed to playback information and arguments to research participants themselves, so as

to reach shared insights or nuanced agreements. Sometimes also DD research 'subjects' can shape how proceeding operate in limited ways, and more rarely they may be involved in publications and impact work. However, they are paid participants and normally never determine studies' fundamental designs. In other ways too, the concept of openness in the DD handbook appears not to go beyond traditional 'closed' research attitudes about reproducibility.

[2] In a footnote on p. 400. This is also the sole mention of researchers depositing data for later reuse.

References

AIATSIS (2024a) 'Cultural sensitivity', Web page. https://perma.cc/99JP-EZD5

AIATSIS (2024b) 'Ethical research', Web page. https://perma.cc/PA6Z-K8H2

Albert, Alexandra; Balázs, Bálint; Butkevičienė, Eglė; Mayer, Katja; and Perelló, Josep (2021) *Citizen Social Science: New and Established Approaches to Participation in Social Research*, pp. 119–128, https://doi.org/10.1007/978-3-030-58278-4_7, in Vohland, Katrin; Land-Zandstra, Anne; Ceccaroni, Luigi; Lemmens, Rob et al. (eds) *The Science of Citizen Science*, Berlin: Springer. https://doi.org/10.1007/978-3-030-58278-4

Bastow, Simon; Dunleavy, Patrick; and Tinkler, Jane (2013) *The Impact of the Social Sciences: How Academics and their Research Make a Difference*, London: Sage. $ https://doi.org/10.4135/9781473921511

Bonney, Rick; Phillips, Tina B; Ballard Heidi; Enck, Jodie W. (2016) 'Can citizen science enhance public understanding of science?' *Public Understanding of Science*, vol. 25, no. 1, pp. 2–16. $ https://doi.org/10.1177/0963662515607406 OA at: https://perma.cc/QE62-A8S9

Carroll, Stephanie R; Herczog, Edit; Hudson, Maui; Russell, Keith; and Stall, Shelley (2021) 'Operationalizing the CARE and FAIR Principles for Indigenous data futures', *Scientific Data*, vol. 8, article 108. https://doi.org/10.1038/s41597-021-00892-0

Ercan, Selen A.; Asenbaum, Hans; Curato, Nicole; and Mendonça, Ricardo F. (eds) (2023) *Research Methods in Deliberative Democracy*, Oxford: Oxford University Press. $ https://doi.org/10.1093/oso/9780192848925.001.0001 OA at: https://library.oapen.org/handle/20.500.12657/60764

European Commission (2018) *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A Credible Enlargement*

Perspective for and Enhanced EU Engagement with the Western Balkans, Strasbourg, European Commission. https://www.eeas.europa.eu/node/39711_en

Friere, Paulo (2005) *Pedagogy of the Oppressed*, London: Continuum. **$** https://www.bloomsbury.com/uk/pedagogy-of-the-oppressed-9781501314162/

Grönlund, Kimmo and Herne, Kaisa (2022) 'Experimental methods', in Ercan, Selen A; Asenbaum, Hans; Curato, Nicole; and Mendonça, Ricardo F. (eds) *Research Methods in Deliberative Democracy*, Oxford: Oxford University Press. **$** https://global.oup.com/academic/product/research-methods-in-deliberative-democracy-9780192873361 OA at: https://library.oapen.org/handle/20.500.12657/60764

Hecker, Suzanne; Garbe, L.; and Bonn, A. (2018b) 'The European citizen science landscape – a snapshot', pp. 190–200, https://doi.org/10.14324/111.9781787352339, in Hecker, Suzanne; Haklay, Muki; Bowser, Anne; Makuch, Zen; Vogel, Johannes; and Bonn, Aletta (eds) *Citizen Science. Innovation in open science, society and policy*, London: UCL Press. https://perma.cc/CF8G-865M

Hecker, Suzanne; Haklay, Muki; Bowser, Anne; Makuch, Zen; Vogel, Johannes; and Bonn, Aletta (eds) (2018) *Citizen Science. Innovation in Open Science, Society and Policy*, London: UCL Press. https://perma.cc/CF8G-865M

Heinisch, Barbara; Oswald, Kristin; Weißpflug, Maike; Shuttleworth, Sally; and Belknap, Geoffrey (2021) 'Citizen humanities', pp. 97–118, https://doi.org/10.1007/978-3-030-58278-4_6, in Vohland, Katrin; Land-Zandstra, Anne; Ceccaroni, Luigi; Lemmens, Rob et al. (eds) *The Science of Citizen Science*, Berlin: Springer. https://doi.org/10.1007/978-3-030-58278-4

Heiss, Raffael and Matthes, Jörg (2017) 'Citizen science in the social sciences: a call for more evidence', *GAIA – Ecological Perspectives for Science and Society*, vol. 26, no. 1, pp. 22–26. OA: https://perma.cc/44DK-3L32

Hidalgo, Enric S; Perelló, Josep; Becker, Frank; Bonhoure, Isabelle et al. (2021) 'Participation and co-creation in citizen science', pp. 199–218, https://doi.org/10.1007/978-3-030-58278-4_11, in Vohland, Katrin; Land-Zandstra, Anne; Ceccaroni, Luigi; Lemmens, Rob et al. (eds) *The Science of Citizen Science*, Berlin: Springer. https://doi.org/10.1007/978-3-030-58278-4

Housley, William; Procter, Rob; Edwards, Adam; Burnap, Peter et al. (2014) 'Big and broad social data and the sociological imagination: a collaborative response', *Big Data & Society*, July, vol. 1, no. 2. https://doi.org/10.1177/2053951714545135

Hudson, Maui; Garrison, Nanibaa' A.; Sterling, Rogena; Caron, Nadine R. et al. (2020) 'Rights, interests and expectations: Indigenous perspectives on unrestricted access to genomic data', *Nature Review: Genetics*, vol. 21, pp. 377–384. $ https://doi.org/10.1038/s41576-020-0228-x

JUSTINT programme (2024) LSE, website. https://perma.cc/BL6N-VR3R

Kostovicova, Denisa (2023) *Reconciliation by Stealth: How People Talk about War Crimes*, Ithaca: Cornell University Press. $ https://doi.org/10.1177/1468794120975657 https://perma.cc/BZ3Y-CMLZ

Kostovicova, Denisa and Eleanor Knott (2022) 'Harm, change and unpredictability: the ethics of interviews in conflict research', *Qualitative Research*, vol. 22, no. 1, pp. 56–73. $ https://doi.org/10.1177/1468794120975657

Kostovicova, Denisa; Martin, Mary; and Bojicic-Dzelilovic, Vesna (201) 'The missing link in human security research: dialogue and insecurity in Kosovo' *Security Dialogue*, vol. 43, no. 6, pp. 569–585. $ https://doi.org/10.1177/0967010612463489

Kullenberg, Christopher and Kasperowski, Dick (2016) 'What is citizen science? – a scientometric meta-analysis', *PLoS ONE*, vol. 11, no. 1, article e0147152. https://doi.org/10.1371/journal.pone.0147152

Laakso, Mikhael (2022) 'Open science and open access publishing in social sciences', in Engels, Tim C.E. and Kulczycki, Emmanuel (eds), *Handbook on Research Assessment in the Social Sciences*, Chichester: Elgar, pp. 105–118. $ https://doi.org/10.4337/9781800372559 OA at: https://perma.cc/C7WN-A3KP

Leek, Lauren; Bischl, Simeon; Freier, Maximilian (2024) *Introducing Textual Measures of Central Bank Policy-Linkages Using ChatGPT*, Department of Political and Social Sciences, European University Institute. https://perma.cc/NKB4-DWD6

Lennette, Carolin (2022a) 'How do we engage in co-research: coproduction and mess', in Lennette, Carolin (ed.) *Participatory Action Research: Ethics and Decolonization*, New York: Oxford University Press, pp. 60–78. $ https://doi.org/10.1093/oso/9780197512456.001.0001

Lennette, Carolin (2022b) *Participatory Action Research: Ethics and Decolonization*, New York: Oxford University Press. $ https://doi.org/10.1093/oso/9780197512456.001.0001

Liabo, Kristin; Ingold, Anne; and Roberts, Helen (2017) 'Co-production with "vulnerable" groups: balancing protection and participation', *Heath Science Reports*, vol.1, no.3, article e19. https://doi.org/10.1002/hsr2.19

Marzi, Sonja (2022) 'Co-producing impact-in-process with participatory audio-visual research', *Area*, vol. 55, no. 2, pp. 295–302. https://doi.org/10.1111/area.12851

Pii, Katherine H.; Schou Lone H.; Piil, Karin; Jarden Mary (2019) 'Current trends in patient and public involvement in cancer research: a systematic review', *Health Expectations*, vol. 22, no. 1, pp. 3–20. https://doi.org/10.1111/hex.12841

Prainsack, Barbara (2014) 'Understanding participation: the "citizen science" of genetics', in Prainsack, Barbara; Schicktanz, Sile; and Werner-Felmayer, Gabriele (eds) *Genetics as Social Practice: Transdisciplinary Views on Science and Culture*, Farnham: Ashgate. OA at: https://perma.cc/U3AQ-96WW

Purdam, Kingsley (2014) 'Citizen social science and citizen data? Methodological and ethical challenges for social research', *Current Sociology*, vol. 62, no. 3, pp. 374–392. $ https://doi.org/10.1177/0011392114527997

REKOM (2018) XI Forum/za tranzicionu pravdu u post-jugoslovenskim zemljama, Sarajevo, Bosnia and Herzegovina, 28 and 29 January 2018.

Robinson, Lucy D.; Cawthray, Jade L.; West Sarah E.; Bonn Aletta; and Ansine, Janice (2018) 'Ten principles of citizen science', pp. 27–40, https://doi.org/10.14324 /111.9781787352339, in Hecker, Suzanne; Haklay, Muki; Bowser, Anne; Makuch, Zen; Vogel, Johannes; and Bonn, Aletta (eds) *Citizen Science. Innovation in Open Science, Society and Policy*, London: UCL Press, pp. 27–40. https://doi.org/10.14324/111.9781787352339

Schubotz, Dick (2020) 'The history and nature of participatory research methods', $ https://doi.org/10.4135/9781529799682.n2, in Schubotz, Dick (ed.) *Participatory Research: Why and How to Involve People in Research*, London: Sage. $ https://doi.org/10.4135/9781529799682

Schroeder, Doris; Chatfield, Kate; Singh, Michelle; Chennells, Roger; Herissone-Kelly, Peter (eds) (2019) *Equitable Research Partnerships: A Global Code of Conduct to Counter Ethics Dumping*, Berlin: Springer. $ https://link.springer.com/book/10.1007/978-3-030-15745-6

Sokolić, Ivor; Vico, Sanja; and Kostovicova, Denisa (2023) *Inter- and Intra-ethnic Dialogues on War-time Violence and its Legacies: JUSTINT Policy Paper.* Justice Interactions and Peacebuilding, London: UK. http://doi.org/10.21953/lse.1bi720lr3oea

Stuart, Kaz and Lucy Maynard (2022) *The Practitioner Guide to Participatory Research with Groups and Communities*, Bristol: Bristol University Press.

$ https://policy.bristoluniversitypress.co.uk/the-practitioner-guide-to-participatory-research-with-groups-and-communities

Tauginienė, Loreta; Butkevičienė, Eglė; Vohland, Katrin; Heinisch, Barbara et al. (2020) 'Citizen science in the social sciences and humanities: the power of interdisciplinarity', *Palgrave Communications*, vol. 6, no. 89, pp. 1–11. https://doi.org/10.1057/s41599-020-0471-y

Thambinathan, Vivetha and Anne, Elizabeth (2021) 'Decolonizing methodologies in qualitative research: creating spaces for transformative praxis', *International Journal of Qualitative Methods*, vol. 20, pp. 1–9. https://doi.org/10.1177/16094069211014766

Vohland, Katrin; Land-Zandstra, Anne; Ceccaroni, Luigi; Lemmens, Rob et al. (eds) (2021) *The Science of Citizen Science*, Berlin: Springer. https://link.springer.com/book/10.1007/978-3-030-58278-4

PART IV
Shifting research cultures

Up to the 1990s academic publications became an ever more professionalized and insulated system, increasingly closed to all outsiders except big corporates, mainstream media and big state agencies that could afford sky-high journal subscription charges, garnering vast profits for a few global publisher giants with the political power to be protected from normal market competition regulations (especially via lobbying the U.S. Congress). Journals multiplied, and their foci became ever more specialized and esoteric. Review processes intensified a search for 'original' work in minute areas and fuelled the non-reporting of replications and non-confirmatory results. The vocabulary of research papers congealed into impenetrable prose, while charts, figures, tables and diagrams became increasingly underdescribed in the main text (because of pressures on word limits) and underlabelled – since only elite cognoscenti would ever access the paper (Plavén-Sigray et al. 2017). Social scientists in fields prone to 'physics envy' (like mainstream economics) and even humanities scholars were quick to copy STEMM (science, technology, engineering, mathematics and medicine) patterns of behaviour (Economist 2024).

Thankfully the open science (OS) agenda has already reversed many of these adverse trends, with scholars and scientists once again encouraged to communicate clearly, including to non-specialist audiences. And huge advances have been made in 'finding new ways of powering intuitive text accounts and visually effective representations' (Dunleavy and Tinkler 2021, p. 116). Yet the social sciences still lag a long way behind modern best practice in getting research results across to the public. To better adopt open methods, social scientists need to change from a legacy 'paywall' professional culture by:

- communicating research and results better, and reaching wider audiences (Chapter 14);
- developing modern publication strategies, ensuring that new generations of researchers appreciate that openness and digital publishing will be essential for their career paths from now on (Chapter 15); and
- modernizing 'legacy' working practices in every discipline, to join up more within an open social science (OSS) culture fitted for the digital era and for making integrated, interdisciplinary advances in applied knowledge (Chapter 16).

14. Presenting results well to reach a wide audience

The key results of research are mostly reported in 'exhibits' or 'attention points', that is distinct evidence elements designed to stand out as especially strong 'argument tokens' – such as tables, graphs, charts, flow charts, graphics, diagrams, boxes for examples or case studies, pictures, and even headings and second colour printing. Attention points seek to attract readers' early and concentrated focus and attention. They are meant to be 'read' and understood on their own, before the main text is considered in detail. Academics, researchers and graduate students normally include exhibits in their quick preliminary scans of articles or books, seeking to quickly assess what value newly found materials have for their own research. Yet, when it comes to presenting data, many researchers nonetheless repeatedly commit the same simple and easily avoidable mistakes that frustrate easy understanding. Nor are they often dissuaded from doing so by editors and publishers, who tend just to focus on curbing text sizes and keeping typesetting simple.*

The open science (OS) approach on exhibits and data presentation covered so far has stressed making full information available to readers. Yet open access publishing also implies that data should be made as simple as possible for the maximum audience. We begin by exploring if these apparent crosscurrents conflict, showing how ideally different types of data can be separated out and presented to different audiences (Section 14.1). Other detailed steps also mitigate any cross-pressures, even when authors must meet stringent length limits. We next briefly review some good practice suggestions for designing clear and immediately understandable exhibits (Section 14.2). Finally, we consider some broader principles for generating effective data visualizations (Section 14.3).

How to cite this book chapter:

Dunleavy, Patrick and Monteath, Timothy (2026) 'Presenting results well to reach a wide audience', in Dunleavy, Patrick and Monteath, Timothy (2026) *Doing Open Social Science: A Guide for Researchers.* London: LSE Press, pp. 323–342. https://doi.org/10.31389/lsepress.dos.n

14.1 Simple exhibits versus full reporting of data

Open social science (OSS) urges authors to maximize the full and precise availability of data, so as to enhance transparency for all readers, and ensure the 'reproducibility' of research where possible. Yet another OSS component, open access publishing (discussed in Chapter 15), also asks that results should be simplified wherever feasible, so as to make them transparent for the broadest possible audience (Divecha, Tullu and Karande 2023). So, the OSS push here is to give readers full and detailed data, while the open access push is for easy-to-understand data in charts, tables and exhibits in general.

Do these pressures conflict? There certainly have been some potential problems in the past. Yet, on inspection, they stem from the survival into the digital era of 'legacy' behaviours by publishers and academics, largely inherited from the demands of printing on paper in the1950s and '60s. Current academic practice is still littered with bits of unexamined and now completely unneeded ways of doing things (Dunleavy 2014). A great example here is how many serious journals still abbreviate their own names in reference lists, in an idiosyncratic fashion that varies from one journal to another and is always completely impenetrable to outsiders. In the digital era why not just always spell out the full journal name, rather than pointlessly saving a few characters as if we were still in the days of 'hot metal' printing? Journal abbreviations impenetrable to outsiders are carried to ludicrous lengths of 'alchemical' mystery by U.S. law journals (CommonLi, no date). And they have even less point following the transition to using DOI numbers (Bond and Green 2014). (No journal names are abbreviated anywhere in this book – which is the transparent, OSS thing to do.)

Data display problems are chiefly due to restrictive word and page limits for articles and book chapters. Some journals still restrict authors to an arbitrary limit on exhibits per article (e.g. no more than six figures) and count each as 500 words towards a tight word limit. Authors with even mildly complex data often feel forced into cramming too much information into their allotted exhibits and may be too close to their data to spot glitches or issues. Journal and book editors rarely improve exhibits, focusing solely on articles' main text. Some limits on attention points can perhaps be 'gamed', perhaps by using 'panelled' figures that count as one exhibit (see Schwarbish 2014), so long as journals and publishers accept how you designate separate figures at face value.

Wherever possible, the best step to reconciling differing OS and open access objectives is to make data available in three formats:

(i) *Accessibly designed and self-explanatory main text exhibits* in the main text of articles, chapters or books, as discussed in the rest of this chapter.
(ii) *A replication archive for articles and chapters or a research methods annex (RMA) for a book,* as discussed in Chapter 8. Each of these

supplementary formats gives readers the background data and details they need to fully understand each exhibit presented. They normally come at the end of the article, chapter or book; or they are deposited reliably online on a secure and continuing basis – either at the journal or book publisher's website, or in a repository – and integrally linked to from within the main text itself (see Chapter 8).

(iii) *A repository version of the full dataset* used in the project, one that is open access and allows reusers to reliably recreate any exhibit (see Chapter 7).

Readers with different levels of 'need to know' questions about the provenance and detailed construction of the data can thus easily access the information they need and in a format they can handle.

However, it seems unlikely that even a critical mass of social scientists providing full availability on these lines will be realized soon, for various constraining reasons, while universal adherence to this three-formats model will take even longer. Some journals and most legacy book publishers are still very reluctant to typeset article or chapter/book annexes placed at the end of their main texts, or even to host them as separate documents on their websites. Consequently, the middle step (ii) above is often excluded. Author promises of data availability also seem to regularly lapse into broken links or once-open supplementary data being subsequently withdrawn (Krawczyk and Reuben 2012). If journal supplementary materials are the weak link, then it becomes even more important to really strive for one-take accessibility in the main text exhibits (stage (i)), and to provide full data reproducibility in a repository at stage (ii) above, as well as in the final stage (iii) (see Chapter 7).

Other issues are also tractable. For instance, in the rest of this chapter we show how with extra care data visualizations can be defined in ways that address common problems. Data scientists' innovations have expanded ways to cope with displaying big data (Schmidt 2022; Kelleher and Braswell 2021), which can be hard to present well in conventional displays. Other newer data approaches address issues such as preserving the privacy of research subjects (Avraam et al. 2021) or highlighting key issues, such as medical harms in clinical trials (Riaz et al. 2022).

14.2 Designing attention points well

In surveying newly found publications for their potential value, most experienced academic readers skim read for their potential value, stopping on the way to try to understand any exhibits encountered. Attention points should help readers to decide quickly between their four main options:

- to move on somewhere else;
- perhaps to save the title and abstract 'just in case' it might come in useful;

- to download a PDF to their PC library for later reading; or
- to commit immediately to a careful read of the paper (Dunleavy and Tinkler 2021, pp. 117–121)

The following *design tips* can help to 'suck in' as many skimming readers as possible to storing, reading carefully and (hopefully) citing your work.

Make attention points as self-explanatory as possible, so that readers can follow (most of) what they mean without having to scan the main text in detail. Because tables, charts, flow charts, heat maps, diagrams and other exhibits are almost invariably looked at early on, *before* their surrounding text, they should be as self-contained and accessible as possible. Captions and axis and variable labels should give full information about what is shown, using only ordinary language, well-known expert language, or abbreviations that are widely understood in a specific research field. Never use a symbol or formula alone in a chart or diagram where space allows for a descriptive label. This is especially the case for novel, unexplained or author-specific labels without an intuitive meaning (e.g. abbreviated variable names), and all acronyms. However constrained the visual space, an accompanying explanatory note (just below the exhibit) should decode what any new acronyms or potentially unfamiliar formulae mean. Ideally, it may also briefly describe in an intuitively understandable way what the key finding or 'take away' message from the exhibit is.

Show readers what they need to know – no more, no less. Exhibits must be designed for their audience, not for authors hung up on demonstrating their 'academicness'. Many authors of academic books and articles, PhDs and other research publications report overly precise results, going beyond three or four 'effective digits'. This trait is often paradoxically allied with mislabelling or underdescribing charts and tables, because the authors are trying to cram into one table or chart all the 'version of record' information that might be needed by any possible reader. Doing better means thinking through in advance who your audiences at each stage are, and what their needs or interests are likely to be at this particular point in your text. Suppose, for instance, that your exhibit in the main text covers the growth of a given variable across a period. The wrong way to proceed – the no-design way – is to compile a large table with seven or eight effective digits in every cell (such as 12,293,417) just because that is what it says on your computer printout of the raw data. But does *any* reader really need to know all these numbers to eight effective digits? Would changing the unit numbers in the column heading to millions and putting 12.3 or 12.29 (instead of 12,293,417) be just as good from (most) readers' point of view? Also, suppose this number is a count of unemployed people. Are official statistics really statistically accurate to ±1 person (which is what quoting a number like 12,293,417 implies that you believe)? *Only quote numbers of digits to a level of accuracy that is credible. So, converting the data unit to 'thousands of unemployed' people and writing 12,293 only implies official accuracy to ±1,000 people.* If the raw data is in numbers <1 (like 0.0215483 to

power −10) bear in mind that the vast majority of people will have little grasp of what that means – here rebase the numbers to get them back into more tractable formats, even if this means going against common conventions in the relevant scientific sub-field. The two most accessible number ranges run either from 0 to 10, or from 0 to 100. With percentage numbers, always use the number form (65.1%) and never the decimal form (0.651) for charts and tables in the main text.

Alternatively, presenting data in a chart or a graph always radically simplifies the numbers involved at a stroke, helping readers to find the main message in a series. Choose the right units to express things in, ideally so that the numbers in the chart cluster reasonably closely in the same number ranges (especially 0 to 100). If some very low and very high numbers are involved, parts of the data may get 'suppressed' by normal linear scaling – so here can you switch to log scales and get all the message into clearer focus? (If not, data with extreme numbers can be better handled as tables.)

In the recent past publishers' cost-cutting often only *detracted* from the value added by charts and diagrams (e.g. by sending academics' carefully designed figures off to be reset on another continent by non-experts, or converting colour charts into black and white ones). This phase has largely passed for journal charts, and most authors can now send well-designed exhibits to journals or publishers to be used just as sent. However, publishers still often mess around tables through more 'legacy' practices – like 'stretching' small tables across the full-page width, making them less readable and full of white space; or using even column widths for large text tables (which normally benefit from variable widths depending on the text to be handled in each column). In general, book publishers also remain far more conservatively inclined about graphics, and have less experience in handling them well, than the top journals now.

Make the messages in attention points as simple as possible (but not more so). Lots of academics live with jumbled data and complex equations that take a whole wall display in the project office to trace out as flow charts or algorithms. Researchers train themselves to operate continuously with complex numbers, visualizations and messages, and for long periods they may only explain them to colleagues like themselves. Naturally enough, this makes it hard for authors to 'see' tangles, issues or problems in thoroughly familiar data, so that some unconscious, 'automatic' reproduction of avoidable complexity makes it into exhibits. Point out examples to academic authors, and they often look at you oddly. 'Well, of course, if you want to go that far,' their expression implies, 'but I'm not writing a nursery book.' However, simplifying the presentation of results as far as possible does *not* mean oversimplifying them or giving an accessible but misleading account of what is shown. Accuracy is vital, as well as accessibility.

An incredibly basic example of such problems occurs when social scientists (and of course, national statistics and governments bodies) reproduce tables, graphs or bar charts that are completely jumbled up because data are

presented in alphabetical order listings (of named areas or policies), or some other 'customary' sequence (such as moving north to south within a country) (Dunleavy 2003, Chapter 7). Sometimes the producers of alphabetically organized data displays imagine that readers are comparing different data across multiple exhibits, and so a common sequence order will help them, a false conviction beloved by statisticians working for national governments, and some process-focused accountants.[1] In fact, readers just have to identify and try to draw out data trends and patterns for themselves. Alphabetization is also rife in data from international organizations (including the United Nations (UN), World Bank, European Union (EU) and Organisation for Economic Co-operation and Development (OECD)), because countries often complain if they come 'bottom of the chart or table' arranged in a numerical progression. Hence the politically 'safe' option for an agency is to issue only alphabetically jumbled data, where readers will be hard put to easily identify which is the worst performing nation.

Wherever possible, never use alphabetical or other customary orderings for data displays. Instead use the ordering of rows/columns in tables and lines/bars in figures to highlight the evidence – ideally using a *descending numerical progression (largest number at the top, going down across rows in the data, with the smallest numbers at the bottom*). With multiple columns in a table or multiple bars or lines in a chart, maximum simplicity means *either* showing each series in a separate exhibit ('panelling'), *or* choosing the most important numerical progression in theory or explanatory terms to organize all the data shown. If feasible in a table, once a numerical progression is in place down the rows, try to achieve a numerical progression across table columns as well.

Finally, *report data in an 'open' way* understandable for general readers, yet also including specific numerical data points for researchers who might want to reuse or 'mash' your data. Wherever possible, avoid abbreviations and instead use fully understandable labels for row variable values (regions, firms, etc.) – so pick horizontal (and not vertical) bar charts or histograms wherever feasible. Also pay attention to issues around ensuring accessibility for disabled readers, which can be handled best through intelligent choices of colours and modes of display (Schwabish, Popkin and Feng 2022).

Use a single sequence of attention points, not separate sequences for tables, figures, boxes, etc., a practice that has always been undesirable because:

- The dichotomy of figures and tables is artificial and thoroughly out of date now. In the digital era 'figures' often include lots of numbers, and 'tables' increasingly use graphic or diagrammatic elements (like colours, boxes, arrows, back-shading, etc.).
- Creating two different sequences of exhibits makes life more complex for readers to check information, and to no purpose. In the digital era (when page numbers have become meaningless with resizable type-sizes) readers can best access all exhibits using just one, easy-to-find numbering sequence. Give readers one Figure *xx* number, which they

can immediately find using the control+F keys (or command+F on some Apple PCs) in the digital text.

- A single exhibits list, where tables, charts, diagrams, photographs, etc. are all labelled as 'figures' already works very well for most reputable research organizations and also in some academic books and journal articles. If you are a journal editor, or on a journal advisory board, why not implement this change right now?
- A single sequence should also kill off another bizarrely pointless legacy publishing convention, that the headings or captions for tables should come *above* the attention point (which helps readers most, and leaves the bottom of the exhibit free for notes, etc.), but the headings for figures should come *below* a chart or diagram (where the heading inevitably conflicts with and easily gets lost amidst the legend, figure key, notes and other impedimenta often needed). All headings (whether for tables or charts/graphs etc) should come immediately above the main attention point, where they are first encountered by and are most accessible for readers.
- Short example or case study boxes and photographs can also be well indexed by listing them all as 'figures'. This designation might also cover other different evidence elements (like short supporting videos or audio-only files).
- Authors can save a lot of time in producing books and articles if they place captions or headings for a single set of figures within the main text at the appropriate point – and never inside the exhibit itself (which creates pointless extra work if either the number of figures (or tables) changes, or they are reordered or relabelled).

Of course, where authors are submitting to a journal that still adheres to the legacy convention separating out figures and tables they may have no choice but to comply. But in books, research reports and 'grey' literature, where authors' preferences can set the format (or at least count more in doing so), always aim for a single sequence of attention points.

Design tables and charts in the main text separately from either exhibits material backing up each exhibit in annexes or the wider research dataset in online repositories. Exhibits within the main text should have a single purpose, supporting a single part of the core argument, without being burdened with unneeded detail. Charts or tables in the replication annex or the research methods annex (RMA) tables that come at the ends of articles, chapters, or books, should be more detailed. They may meet multiple needs, including creating the 'version of record' that demonstrates the accuracy and replicability of that bit of the work. Full datasets lodged in repositories with good metadata, methods documentation and codebooks can fulfil the same role where annexes or RMAs are not feasible.

When explaining multivariate analyses, especially regressions, social scientists often over-report data in ways that blur together information vital to

sustain the main text argument with that needed for reproducibility. A misguided 'professional' anxiety to report all the data here can produce large, 'dead on arrival' tables of effects and standard errors, with variables ordered in a jumbled sequence alphabetically or labelled using only the arbitrary numbers given by where each variable comes in the codebook. Regression variable names are almost all abbreviated in ways that are either meaningless or ambiguous; sometimes their real meaning may only be properly explained in the full data codebook deposited in a repository. For most readers (even for expert regression users unfamiliar with this particular dataset) understanding regression coefficients in such tables can be extraordinarily difficult without carefully checking what kind of variable each one is. Readers always need to know exactly how independent or explanatory variables are set up – for example, are they dichotomous variables, graduated codes, or fully numerical data (and measured in what units)? Expressing the effects of any given independent variable in terms of what happens with a given unit change in the explanatory variables can be much more intuitively understandable than quoting standard coefficients for differently constituted variables. The numbers published often make sense only to the authors and perhaps the most conscientious paper reviewers, if they devote a big chunk of time to the task. Critics note that:

> The numbers published in physics are steppingstones for further inquiry, because they are few and other researchers often make use of them. The numbers published in social sciences are endpoints, because they are profligate and, once printed, hardly anyone makes use of them (Taagepera 2008, p. 82; and see Taagepera 2017).

Always fully describe what numbers are being shown. This means accurately labelling the *x* and *y* axes of *every* chart, and the column and row headings and row labels in all tables, whether in the main text or annexes. Readers should see captions and labels that are full, readable and use intuitively understandable names for the variables being shown, not acronyms or obscure abbreviations (so again, use horizontal bar charts, not vertical ones, except for over time displays). Axis labels for numerical variables should include the unit of measurement used. Take care when showing percentages that readers know exactly what they are percentages of – usually that entails showing all the % numbers summing to 100% across a row or down a column. In data showing changes over time, make sure that labels say 'percentage point change' where needed, and not just 'per cent'.

Many academics and PhDs do not take their obligations to accurately inform readers seriously enough here; for example, presenting results without having any unit axis labels at all, instead relying on the overall table or chart caption to explain what is shown, and often omitting measurement units. Gnomic labels for charts or diagrams can be completely obscure, especially using only algebraic symbols, whole formulae, or function equations or names

that are actually only explained in the main text. As Dunleavy and Tinkler (2021, p. 340, and Figures 11.4 and 11.5) note, a range of easy-to-use mitigation strategies can enhance openness here:

> Researchers need to put themselves in the place of 'lay' people who find the subject matter of their research alien enough without having algebra incomprehension supplementing their difficulties. Replacing equations with diagrams wherever possible [e.g. a flow chart], using intuitively memorable [and understandable] labels, 'translating' symbols [into words], avoiding obscure symbols [like Greek algebra signs and compound formulae] and explaining any algebra 'moves' made (like rearranging terms or cancelling out) can all help readers to follow better. The number one barrier to clear communication here is often just academics' concern to display complex formulae from a specialist exposition, lest any simplifications for a mass audience be taken by other professional academics as indicating a lack of sophistication on their part.

Authors seems to think that readers will process their entire text in a careful linear way, so that they can legitimately assume that readers will understand all the variables, indices or statistics presented in tables and charts from reading the main text – which will not be the case.

A final potent source of underlabelling and underdescription of data comes from some graphics presentation advice books that recommend removing anything not strictly needed from charts. This is an OK approach if undertaken by experts (like *The Economist*'s graphics designers). However, an *over*-concern with parsimony can end up with non-expert researchers omitting axis labels and measurement units in ways that create actively misleading or obscure figures. Simplifying how data are displayed graphically can help a lot when the exposition needs some selected details to be highlighted, or readers also need to summarily appreciate a complex picture. Presenting results well can often be a matter of lots of small cumulative tweaks rather than a radical overhaul. For instance, in a case study-based analysis de-emphasizing some data in order to focus attention on where case studies fit within a larger picture or context can be very helpful, as Schwabish (2014) does with several kinds of 'clutterplot' exhibits (also see Chapter 10, Figure 10.3 in this book).

Strengthening exhibits is especially helpful in improving an article that is 'just not working'. It can be hard to know how to revise a paper that journal reviewers knock back or are underwhelmed by, perhaps more than once, without undertaking too much time-consuming new or extra research work. Authors may tinker with how their text is phrased if an article gets turned down, but they are often reluctant to ask whether more extensive changes could help. For instance, would it help to redo their literature search for newer materials, or refresh their references to show evidence of a more systematic and inclusive literature search, or generate more supportive citations and quotations? Also

often unexamined is the graphics/data presentation. Could all the exhibits be systematically reappraised in ways that make them more convincing? And might their new findings be better integrated with previous empirical evidence? These changes take a bit more time, but are often the key improvements – because they mean going back to things previously done at the early stage of article or chapter writing, when researchers actually only understood their topic rather imperfectly.

'Going back to the start' on exhibits planning may seem offputting. But more effective exhibits that follow sound design approaches can greatly strengthen reviewers' often rushed initial impressions of the work, as well as their mature confidence in it – making it a lot easier for them to give a positive recommendation. Making sure that data is reported in forms that matter to reviewers – and not in dead on arrival mode – also helps. Reviewers do not want to feel forced into having to closely examine jumbled data to work out what is going on. Even on a revise and resubmit, the impetus to improve exhibits needs to come from authors themselves – reviewers themselves may be reluctant to propose such changes.

Avoid 'legacy' publishing practices that only impair readers' understanding and have no contemporary value. The problematic conventions are usually just now pointless customary behaviours deriving from the print era, but somehow still set in stone in current practice. In addition to the defects already listed above, another prevalent mistake is to use smaller typesizes in tables and charts than those in the main text – thereby making the exhibits in the text *harder* to read than the main text, and subliminally encouraging readers (especially those on smartphones and tablets) to skip them. Even on a PC readers may have to load a separate figure file to understand what is shown, diverting their attention from the main text to do so. Instead, authors should do their best to ensure that all text within exhibits is clear and legible for digital readers on the smallest devices, using a typesize in exhibits that is as similar as possible to that in the main text. But what if tiny fonts are apparently needed to 'cram' an exhibit onto a portrait page? Here, consider using a landscape page instead, or splitting the figure or table into its component parts as several exhibits. Using a somewhat different *kind* of font for exhibits is helpful, however, so long as it fits well visually with the main text font. Finally, *no* typesizes inside exhibits should ever be any larger than that in the main text: this off-putting mistake can sometimes arise in poorly edited diagrams, charts or pictures.

Give a full and careful description of the data used in an article or chapter. Authors' obsession with overparsimony (from a desire to seem ultraprofessional in style), plus journal size limits, now often deny readers access to key information needed at the start of reading. 'Quantitative social research has become increasingly dependent on complex inferential methods', with many benefits in terms of analysing big data especially, but also some key risks (Murphy 2022a, 2021, 2022b). Increasing numbers of articles have been published without any explanation of the vital descriptive statistics (previously

given in 'Table 1'). Murphy argues that this key step is now often omitted from articles altogether (for space reasons) or done half-heartedly (e.g. with one line like 'Descriptive statistics are presented in Table 1' without any discussion). By contrast, he argues that in the social sciences:

> Any result that is established based on a complex data analysis that cannot be shown to be at least plausible based on the types of simple statistics shown in [*a descriptive*] Table 1 (e.g., means, standard deviations, correlations) should be treated as suspect and interpreted with the utmost caution (Murphy 2021, p. 467).

Descriptive statistics are always needed if readers are to engage effectively with the content – without them both researchers and their readers may lose touch with 'what these analyses can and cannot tell them … As statistical analysis becomes more complex, the likelihood of serious errors in analysis and interpretation increases' (Murphy 2022a).

Present data to maximize public engagement, which generally means doing science communication ('sci comm') in multiple ways – via blog posts, infographics, social media, podcasts and videos; presentations to non-academic but expert audiences in civil society, business and government; briefings for non-specialist academic audiences (Dunleavy and Tinkler 2021, Chapter 7; Schwabish 2014; Basole et al. 2020); plus 'playing back' main research results to study participants (see Chapter 13). It is important to think through these different audiences' social media needs and capabilities realistically.

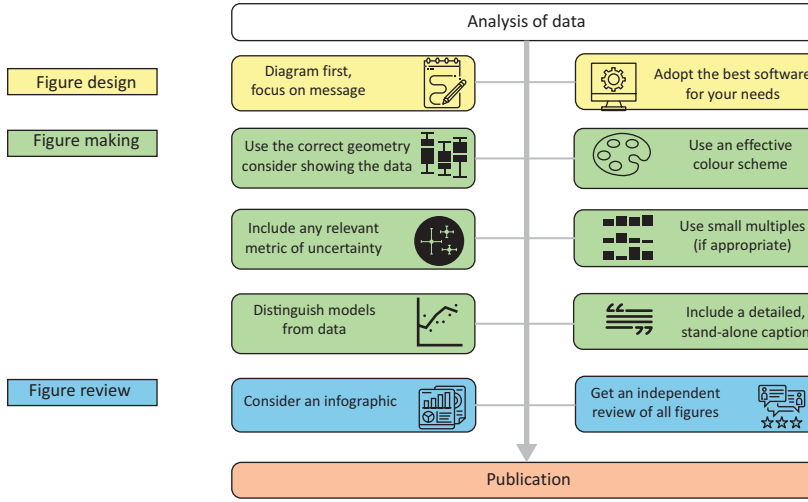
14.3 A broader data visualization strategy

Developing effective data visualizations for an expert or interested audience involves presenting complex and multifaceted data in the most accessible and research-appropriate way:

> Simply put, visualization amplifies human cognition by presenting complex data in ways that facilitate bolder, more appropriate decisions. It is often said that visualization is like a telescope or microscope, enabling users to view and understand processes that are otherwise invisible (Shneiderman 2022, p. 192).

In the past some social scientists resisted accessibility in exhibits from a misguided concern to make their work to look 'professional'. Findings and operations done were explained in a diffident, 'parsimonious' (i.e. over-compressed and inadequate) way to signal the authors' superior technical know-how (Plavén-Sigray et al. 2017). Since then, however, the academic study of data visualization as a field has greatly improved (Engebretson and Kennedy 2020; Divecha, Tullu and Karande 2023). The best modern guides afford strongly

Figure 14.1: Summary of Midway's *Principles of Effective Data Visualization*



Note: This figure is slightly redesigned from Midway's helpful original. We have changed the flow to be from top to bottom; the central grey arrow is added to show that flow; the boxes for 'Analysis' at the top and 'Publication' at the bottom are added; and the original key stage boxes are moved to the left side and separated.
Source: Adapted with permission from Midway (2020).

useful advice and more comprehensive information on numerous types of graphics, and how and where they can be most appropriately used (notably, Schwabish 2021; see also Schwabish 2023; Wilke 2019; Kirk 2024; Miller 2019).

We focus here solely on a single aspect of the topic. Open access publishing implies that presenting complex data well should always tell a clear 'story', supported by exhibits that are easily digestible. Any given data and analysis can be presented in multiple modes, and there is no one 'correct way' to do so. Even in a field like teaching cartography, Tim notes that he would set students the same mapping task with the same data – and get back 10 very different maps. So, in open access work it makes sense to pay most attention to the *core principles* for designing and implementing effective data visualizations. These have been covered effectively by Stephen Midway (2020), whose useful summary is shown in slightly altered form in Figure 14.1. It encompasses 10 core stages from data analysis to final publication (see also Unwin 2020; Basole et al. 2020).

(i) Perhaps the most problematic phase in creating a new visual is the one shown in the top left yellow box for figure design – moving from analysis to thinking of an optimum visualization. Diagramming only here is important in ensuring that diverse early concepts are

considered, instead of a premature commitment to one solution fixing 'tramlines' that later cannot be easily jumped. To avoid any initial 'straitjacket', sketch several possible kinds of visualization *freehand*, focusing solely on overall 'look and feel' issues. Looking in other authors' work for interesting visualization examples and asking team members and colleagues can both help surface multiple contending approaches.

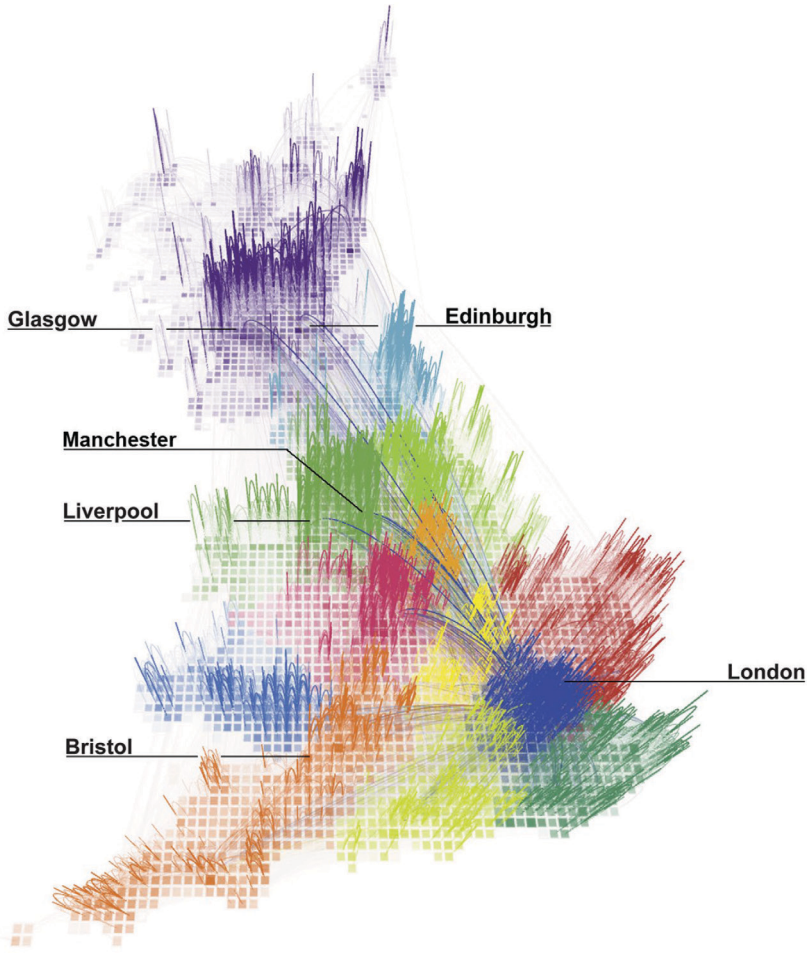
(ii) The second, righthand yellow box above is deciding *what software* to use – either the built-in graphics of analysis packages, or a specialist graphics package that is already known, or new software that can be easily learnt. Most new packages come with transaction costs, but they can probably also be defrayed over a future stream of work.

(iii) Choosing the right 'data geometry' to use goes beyond the conventional options (bar charts, histograms, box plots, scatter plots, regression models, etc.) to take account of multiple other imperatives. For instance, 'heat maps' and many other solutions can cover two-dimensional variations in incidence (Schwarbish 2021); and always try to show both the data and the model (see point (vii) below).

(iv) *Presenting data in colour* should be the norm for all exhibits in the digital era. They convey information immediately, so long as distinctly different colours are used that nonetheless work together harmoniously in a chart or diagram. Many STEMM (science, technology, engineering, mathematics and medicine) articles use multiple strident colours and unmatching colour combinations (as in Midwinter's own article). Social scientists, official documents and many books on achieving 'beautiful' graphics can all tend to the other extreme by seeking an overly 'restrained, professional' image with overly similar colour combinations – such finely graduated shades of the same colour that readers using small digital devices cannot distinguish them. At other times, colours within exhibits may even become hard to disentangle from background shading. Any colour variations in the same chart must always be crystal clear for all readers – but they nonetheless need to go well together, an effect more easily achievable with specialist software. Figure 14.2 shows a good example of effective colour use in a graphic.

(v) Wherever possible, charts and graphics should *include visual indicators of uncertainty* in the analysis, using devices such as error bars, shaded error regions around regression lines and explanatory notes. Showing uncertainty levels in tables also makes sense where it can be done economically. Standard deviations and standard errors are useful in small or medium datasets, but standard errors fall away automatically in 'big data'. So too does the usefulness of all 'significance' tests in discriminating between potentially plausible effects and non-effects (Varian 2014; Dunleavy 2016).

Figure 14.2: The geography of talk by telephone in Great Britain, in 2010

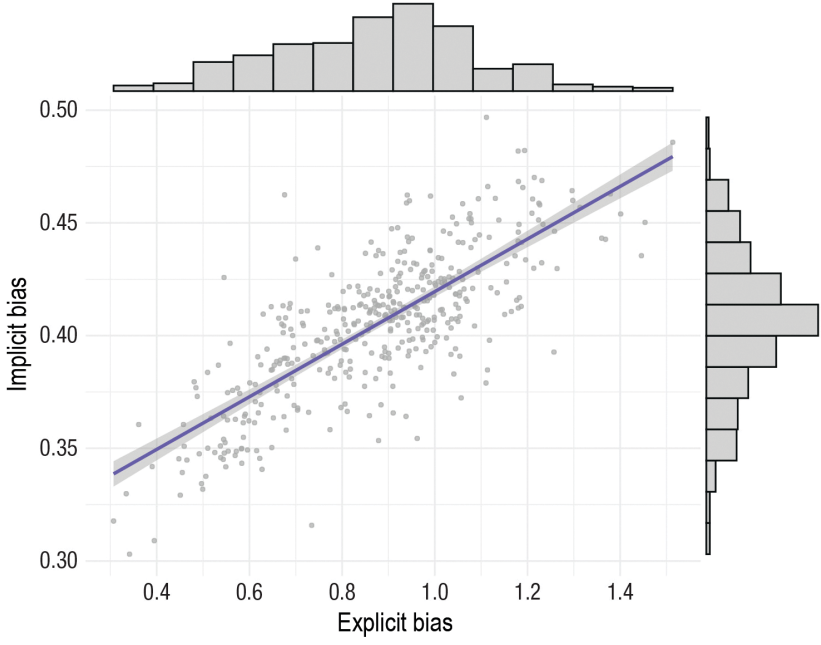


Note: This figure shows the strongest 80% of links, as measured by total telephone talk time, between areas within Britain. The opacity of each link is proportional to the total call time between two areas and the different colours represent regions identified using network modularity optimization analysis.
Source: Ratti et al. (2010). Used under a CC-BY 4.0 licence.

(vi) Use small multiples or 'panels' of charts with the same setup of axes, scaling and geometries to bring out clear differences across variables or to show changes in them – so long as the component charts inside the panels remain large enough to show comparative effects clearly. (Both Midway (2020) and Schwarbish (2014) give examples.)

(vii) Wherever feasible *show both the data and the model* in charts. Do not just rely on presenting the analysis findings of model effects without giving readers a good sense of the underlying data configurations. For instance:
- add a trendline and formula to scatterplots;
- 'jitter' all data points in box plots with small or medium *N* (jittering is spacing out the data points horizontally that occur at the same point on a vertical scale);
- in a dataset that is too large to jitter all data points, it can still be helpful to just jitter the outliers; and
- use 'counting in bins' in one dimension (a histogram) or two dimensions (a heat map or scatter plot) to show the spread of data points where large datasets are modelled (see Figure 14.3, taken from Hehman and Xie (2021)).

Figure 14.3: Hehman and Xie's examples of an 'Improved scatterplot visualizing the relationship between implicit and explicit anti-Black bias, including a 95% confidence band of the slope, with histograms of the variable on each axis in the opposite margins'



Source: Reproduced with permission from Hehman and Xie (2021). Copyright © 2021 by The Authors. Reprinted by permission of Sage Publications.

Similarly, wherever geographical data is available or useful, make it easily locatable, so that skilled readers also easily access in online repositories other relevant geo-located open access data. Extra spatial insights can help reusers in particular to understand the data in depth, and perhaps to think how to remodel it.

(viii) Headings and captions (in which Midwinter includes explanatory notes below charts or tables) should be precise and explain figures in ways that make them self-contained and ideally self-explanatory (see above): 'Captions should do all they can to explain the visualization and representations used (Midway 2020).'

(ix) At the final publication stage, over and above the formal publication or presentation to an academic audience, authors could consider if an infographic (a visual narrative diagram) can help explain a complex chaining of variables and effects. Commissioning an infographic, and especially guiding the artist/summarizer on how to phrase things accessibly, may often aid researchers' own understanding of how to express their ideas accessibly and intuitively, and strengthen the communication of key descriptive statistics. Infographics may well not make it into a finished academic long-form text (like an article or book). But they can still be vital aids in presentations, dissemination work and publicizing project findings on social media.

(x) In an article or book chapter every chart and table (especially those with novel or mixed formats) should be appraised and commented on by one or several independent readers. Close colleagues may not want to criticize your work, so try to have feedback from people who are similar to the intended audience, or from other researchers able to appraise exhibits dispassionately.

Conclusions

In general, there should be no dilemma between focusing in an open access way on being understandable and delivering scientifically accurate and legitimate data in an OSS way. In the digital era, a staged presentation of data in different formats can prevent any clash of these imperatives. As Dunleavy and Tinkler (2021, p. 340) emphasize:

> Working hard to make research data understandable by a wide range of readers does not conflict with improving professional academic communication or with undertaking perceptive and penetrating data analysis. Rather, as the statistics guru John Tukey (1977) constantly emphasized, 'data reduction' – such as clearing away the clutter of over-precise numbers and non-intuitive ways of scaling or basing numbers – can be a vitally important step in

deepening analysis and understanding of the findings for researchers themselves.

For instance, making a big effort to explain formulae in intuitive terms, radically simplifying data in responsible ways and using infographics, visual algorithms and other newly available graphics devices can all help researchers to better understand their data, pick the best form of analysis and explain it better, the opposite of 'dumbing down'. As Midway (2020) also points out, improved data presentation and visualization can improve researchers' scrutiny of their assumptions, inferences and model prescriptions, thereby opening the way for more nuanced theory and model development.

Notes

* This chapter draws on research and seminars forming part of the Civica project, which was funded in 2022–2024 by the European Union's research and innovation programme Horizon 2020, under the chapter 'Science in Society'.

[1] A potential rationale for retaining alphabetical or standardized list orders in tables occurs where many data users must employ 'web scraping' methods to update data without having to retype it. This was an important aspect of data journalism in the early 2000s, when most 'open government' online materials were PDFs in image forms (Bradshaw (no date)), and it may still be needed with 'backward' agencies. However, in western countries that (brief) era of poor official data practices has largely passed. Most governments and international organizations now provide direct data downloads in open CSV formats.

References

Avraam, Demetris; Wilson, Rebecca; Butters, Oliver; Burton, Thomas et al. (2021) 'Privacy-preserving data visualizations', *EPJ Data Science*, vol. 10, no. 2. https://doi.org/10.1140/epjds/s13688-020-00257-4

Basole, Rahul; Bendoly, Elliot; Chandrasekaran, Aravind; Linderman, Kevin (2020) 'Visualization in operations management research', *Informs Journal On Data Science*, vol. 1, no. 2, pp. 172–187. https://pubsonline.informs.org/doi/pdf/10.1287/ijds.2021.0005

Bond, Alexander L.; Green, Jessica R. (2014) 'Journal title abbreviations should be eliminated in the digital age', *PeerJ PrePrints,* vol. 2, article e445v1 https://doi.org/10.7287/peerj.preprints.445v1

Bradshaw, Paul (2017) *Scraping for Journalists* (2nd edn) Lean Publishing. **$** https://perma.cc/3LSU-4NG2

Divecha, C.A.; Tullu, M.S.; Karande, S. (2023) 'Utilizing tables, figures, charts and graphs to enhance the readability of a research paper', *Journal of Postgraduate Medicine*, vol. 69, no. 3, pp. 125–131. **$** https://doi.org/10.4103/jpgm.jpgm_387_23 OA at: https://pmc.ncbi.nlm.nih.gov/articles/PMC10394528/

Dunleavy, Patrick (2003) 'Handling attention points; data, charts and graphics', in Dunleavy, Patrick *Authoring a PhD: How to Plan, Draft, Write and Finish a Doctoral Dissertation or Thesis*, London: Palgrave (now Bloomsbury), pp. 157–196. **$** https://www.bloomsbury.com/uk/authoring-a-phd-9781403905840/

Dunleavy, Patrick (2014) 'Seven upgrade strategies for a problematic article or chapter', *LSE Writing for Research* blog, 31 July. https://blogstest.lse.ac.uk/writingforresearch/2014/07/31/seven-upgrade-strategies-for-a-problematic-article-or-chapter/ Also on Medium: https://perma.cc/3KK5-V92J

Dunleavy, Patrick (2016) 'Big data' and policy learning', in Stoker, Gerry and Evans, Mark (eds) *Methods that Matter: Social Science and Evidence-Based Policymaking*, Bristol: Policy Press. **$** https://doi.org/10.1332/policypress/9781447329367.003.0009 OA at: https://perma.cc/TAP5-L8R9

Dunleavy, Patrick and Tinkler, Jane (2021) *Maximizing the Impacts of Academic Research: How to Grow the Recognition, Influence, Practical Application and Public Understanding of Science and Scholarship*, London: Palgrave (now Bloomsbury). **$** https://www.bloomsbury.com/uk/maximizing-the-impacts-of-academic-research-9780230377608/

Economist (2024) 'Academic writing is getting harder to read – the humanities most of all', *Economist*, 18 December. **$** https://www.economist.com/science-and-technology/2024/12/18/academic-writing-is-getting-harder-to-read-the-humanities-most-of-all

Engebretsen, Martin and Kennedy, Helen (eds) (2020) *Data Visualization in Society*, Amsterdam: Amsterdam University Press. https://perma.cc/6XNZ-AJEX

Hehman, Eric and Xie, Sally Y. (2021) 'Doing better data visualization', *Advances in Methods and Practices in Psychological Science*, vol. 4, no. 4. https://doi.org/10.1177/25152459211045334

Kelleher, Christa and Braswell, Anna (2021) 'Introductory overview: recommendations for approaching scientific visualization with large environmental datasets', *Environmental Modelling & Software*, vol. 143, article 105113. https://doi.org/10.1016/j.envsoft.2021.105113

Kirk, Andy (2024) *Data Visualisation: A Handbook for Data Driven Design* (3rd edn), London: Sage. **$** https://uk.sagepub.com/en-gb/eur/data-visualisation/book285882

Krawczyk, Michal and Reuben, Erensto. (2012) '(Un)available upon request: field experiment on researchers' willingness to share supplementary materials', *Accountability in Research*, vol. 9, no. 3, pp. 175–186. **$** https://doi.org/10.1080/08989621.2012.678688

Midway, Stephen (2020) 'Principles of effective data visualization', *Cell Press*, PATTER 1, 11 December. https://doi.org/10.1016/j.patter.2020.100141 Also at: https://perma.cc/P7TT-XTVS

Miller, Kelsey (2019) '17 data visualization techniques all professionals should know', *Harvard Business School Online*, 17 September. https://perma.cc/8CFM-B6VQ

Murphy, Kevin R. (2021). 'In praise of Table 1: the importance of making better use of descriptive statistics', *Industrial and Organizational Psychology: Perspectives on Science and Practice*, vol. 14, no. 4, pp. 461–477. https://doi.org/10.1017/iop.2021.90

Murphy, Kevin R. (2022a) 'Descriptive statistics are essential to making complex analyses useful', *LSE Impact* blog, 25 April. https://perma.cc/EU58-V55J

Murphy, Kevin (2022b) 'Surviving the statistical arms race', in Murphy, Kevin R. (ed.) *Data, Methods and Theory in the Organizational Sciences: A New Synthesis*, New York: Routledge, Chapter 7.

Plavén-Sigray, Pontus; Matheson, Granville J.; Schiffler, Björn C.; and Hedley Thompson, William (2017) 'Research: the readability of scientific texts is decreasing over time', e-Life, vol. 6, article e27725. https://doi.org/10.7554/eLife.27725.001

Ratti, Carlo; Sobolevsky, Stanislav; Calabrese, Francesco; Andris, Clio; Reades, Jonathan et al. (2010) 'Redrawing the map of Great Britain from a network of human interactions', *PLoS ONE*, vol. 5, no. 12. https://doi.org/10.1371/journal.pone.0014248.g001

Riaz, Qureshi; Xiwei, Chen; Carsten, Goerg; Evan, Mayo-Wilson et al. (2022) 'Comparing the value of data visualization methods for communicating harms in clinical trials', *Epidemiology Reviews*, vol. 44, pp. 55–66. https://doi.org/10.1093/epirev/mxac005

Schmidt, Johanna (2022) 'Visual data science', https://doi.org/0.5772/intechopen.97750, Ch. 6 in Shirowzhan, Sara (ed.) *Data Science, Data Visualization, and Digital Twins*, IntechOpen. February. https://perma.cc/75LP-SV9V

Schwabish, Jonathan A. (2014) 'An economist's guide to visualizing data', *Journal of Economic Perspectives*, vol. 28, no. 1, pp. 209–234. https://doi.org/10.1257/jep.28.1.209

Schwabish, Jonathan A. (2021) *Better Data Visualizations – A Guide for Scholars, Researchers, and Wonks*, New York: Columbia University Press. **$** https://perma.cc/HES4-X8QA

Schwabish, Jonathan A. (2023) *Data Visualization in Excel: A Guide for Beginners, Intermediates, and Wonks*, New York: AK Peters/CRC Press. **$** https://doi.org/10.1201/9781003321552

Schwabish, Jonathan; Popkin, Sue; and Feng, Alice (eds) (2022) *Do No Harm Guide: Centering Accessibility in Data Visualization*, Washington, DC: Urban Institute. https://www.urban.org/projects/do-no-harm-project

Shneiderman, Ben (2022) 'Commentary on "Visualization in Operations Management Research"', *INFORMS Journal on Data Science*, vol. 1, no. 2, pp. 192–193. https://doi.org/10.1287/ijds.2021.0009

Taagepera, Rein (2008) 'Why most numbers published in social sciences are dead on arrival', Ch. 7, pp. 82-92 in Taagepera, Rein. *Making Social Sciences More Scientific: The Need for Predictive Models*, Oxford: Oxford University Press. **$** https://doi.org/10.1093/acprof:oso/9780199534661.003.0007

Taagepera, Rein (2017) 'Science walks on two legs, but social sciences try to hop on one', *International Political Science Review*, vol. 39, no. 1, pp. 145–159. **$** https://doi.org/10.1177/0192512116682185

Tukey, John (1977) *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Unwin, Antony (2020) 'Why is data visualization important? What is important in data visualization?' *MIT Press blog*, 31 January, issue 2.1, Winter 2020. https://perma.cc/93EZ-NNKW

Varian, Hal R. (2014) 'Big data: new tricks for econometrics', *Journal of Economic Perspectives*, vol. 28, no. 2, pp. 3–28. https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.28.2.3

Wilke, Claus O. (2019) *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*, San Francisco, CA: O'Reilly Media. **$** https://perma.cc/VN6M-3K5M

15. Publishing and open access

In the near future (almost) all social science work will be published in digital open access (OA) formats (Laakso 2022). As subscription-model journals die out (even 'hybrid' ones), and all research funders and universities increasingly require OA publication for both journal articles and research books (e.g. LSE 2025), the landscape of career recognition and distinction will change. Professional reputations and scientific attention will both shift away from relying on *where* research has been published (e.g. with old-guard, 'prestige' journals or publishers) to focus instead on *how much research has been used* – whether through other academics citing its findings or reanalysing its data, professionals using it in applied ways, or the research contributing to public debates (Dunleavy and Tinkler 2021).*

So academic mentors, early career researchers and younger faculty urgently need to update past 'legacy' assumptions about where and when to best publish different types of research for career progression. They need to broaden the range of outlets where they publish, diversifying away from a research article monoculture and appreciating how best to use OA outputs to grow their careers (Section 15.1). In fast-changing research fields much more emphasis should also be placed on speedy publication, rather than past 'slow professor' practices of labouring to produce 'tombstone' articles or books published years after their maximum relevance has passed.[1] Moving up a gear to completely OA publication also entails important changes in how funding for journal or book publication fees has been obtained and distributed within social science disciplines and universities.

Reaching any academic audience via digital publishing, let alone a wider public with interests in research, also means that researchers themselves must become far more digitally proactive in disseminating knowledge of their work, using a varied combination of short-form and long-form publications, social media and involvement in wider debates (Section 15.2). Just as fiction authors have to show up in person at bookshop signings and literary festivals to promote their work, if academics' long-form outputs are to be noticed they need to contribute to online debates and hybrid seminars and conferences, making a continuing professional contribution.

How to cite this book chapter:

Dunleavy, Patrick and Monteath, Timothy (2026) 'Publishing and open access', in Dunleavy, Patrick and Monteath, Timothy (2026) *Doing Open Social Science: A Guide for Researchers.* London: LSE Press, pp. 343–362. https://doi.org/10.31389/lsepress.dos.o

15.1 The open access landscape

OA publishing makes academic outputs free to read by anyone with an internet link. In the digital era OA has played a vital role in developing the open science (OS) movement's appeal, especially for open social science (OSS) (see Figure 2.3 in Chapter 2). However, the OA transition was predominantly achieved by commercial publishers shifting their financial base from subscriptions to article processing charges (APCs) as the main form of journal financing (Pinter and Dunleavy 2022b). Some 'hybrid' journals persist, where some authors pay APCs for OA publication, while other authors still pay no charge but their work is paywalled for subscribers only. But this format is generally dwindling away, because fewer university libraries are still willing to pay sky-high journal subscription charges. Alternatively, libraries increasingly require that their faculty can publish OA for free in a publisher's titles, in exchange for maintaining subscriptions to bundles of their journals.

For research books, OA publishing is still an exception in the social sciences, outside the pioneering university OA presses (e.g. ANU Press 2025; UCL Press 2025; LSE Press 2025) and other key OA book initiatives (such as JSTOR 2025; OAPEN 2024; Google Books 2025; DOAB 2024; and the Open Book Collective). Commercial publishers (including most legacy university presses) charge authors a book processing charge (BPC) in return for open access PDF access, and these charges may run into many thousands of dollars. Even a majority of research books available digitally are still charged for, either through individual purchases, or through top university libraries buying digital subscriptions to (collections of) well-used research or teaching books.

Thus, some significant OA issues remain to cover here – the importance of reaching the widest global audience for research; the kind of OA licence to choose; what forms of OA are available for articles; OA formats for research books; and some newer OA arrangements.

Maximizing the audience for research

Full time faculty and students in the global top-rank universities now have relatively wide access to free journal articles and some research books via their library's subscriptions. Yet paywall access still remains an acute problem for the vast majority of faculty and graduate students worldwide (Figure 15.1), and for all other readers who may want to use research in their own professional work. Millions of people in public services, the financial sector, consultancies, and so on have master's or PhD qualifications and can read social science research directly in OA sources. Although these groups are key channels that mediate social science work for practitioner audiences, business corporations and state institutions, they are very unlikely to have access to paywalled books or to many articles (Bastow, Dunleavy and Tinkler 2014, Chapter 9).

Even in the wealthiest countries *not* making research OA can cause myriad problems for academic and external readers. For instance, in the US nearly half of university faculty are not full-time staff (see Figure 15.1) and only a quarter

Figure 15.1: Users of academic work who may not be able to access most published journal articles or research books

Potential user group	Indicators of importance	
Advanced western countries (e.g. EU, US, Anglosphere nations)		
Non-tenure track faculty and academics, especially after leaving a university	In the US, 48% of all faculty are part-time and 7% are on untenured full time contracts (AAUP 2023) and see Figure 15.2)	In the UK, a third of researchers are on fixed term contracts (HESA 2023)
Retired faculty, 'emeritus' professors	Former academic staff often lose their previous library access, yet they often remain active authors and researchers (Crow 2021). In the UK even emeritus professors who retain full library access are excluded from the university–publisher 'bundle' deals on free OA publishing for new articles	
Faculty in non-elite, not top-tier universities, and other college-type institutions	There are 18,000 degree-granting institutions in 180 countries worldwide. Comparing 2024 estimates generated by multiple AI programmes (ChatGPT, Grok, etc.) suggested that there were at least 6 million university faculty staff globally (and perhaps as many as 8 million), with at least 1.5 million professors in the US (*see note below*). Even most university faculty (outside the global top 2,500 universities) have limited or poor access to research journals and books	
Graduate students in non-elite, not top-tier universities, and other college-type institutions	Most graduate students (outside the same top 2,500 universities globally) may have restricted access to research journals and books, depending on how much their university libraries have or can borrow from national facilities. There are at least 2.7 million graduate students in the US alone	
Independent or private practice researchers	May not be able to access university libraries, or not all facilities there. Often they have to pay fees for (some) digital access (although fees are fairly small in the UK and Europe)	

(*continued*)

Figure 15.1: (continued)

Potential user group	Indicators of importance
Corporate researchers	Cannot usually access university libraries except by paying substantial digital access fees. Firms, commercial labs and agency libraries may each have a few journal subscriptions, but only to a few top journals in their area. In OECD countries 1.1 per cent of 25–64 year-olds in the national population also have a PhD; in the US 1.8% of the population has a doctorate (OECD 2017). In the UK, social science graduates in professional roles mediating their discipline's specialism in industry or public services numbered 383,000 in 2014. And in the UK workforce there were 0.54 million people with higher degrees in social science and 1.19 million with first degrees (Bastow, Dunleavy and Tinkler 2014, Figures 9.3 and 9.4)
Government and public services researchers	
Research staff in consultancies	
Research staff in think tanks and institutes	As above, although genuine think tanks inside or closely linked to universities (i.e. not 'junktanks') should have library access
All other countries	
Faculty and students at almost all universities	Outside one or two elite or national universities, plus the national library, academics and students typically only have access to teaching-related materials, a restricted set of journals and those books and journal articles that are fully OA. Comparing 2024 estimates generated by multiple AIs suggests that there were around 9 million graduate students at universities in countries outside the USA (*see note below*)
Corporations and government agencies	Typically have no journals access. Comparing 2024 AI-generated estimates suggests that around 8 million people work as researchers globally, of whom around 40% are STEMM-trained (so may not cite OSS work). Many of the rest work principally in finance, business and government sectors, and for them OSS work is highly relevant
Other civil society organizations	In developing countries only a few local branches of western consultancies, or western-sponsored NGOs/think tanks, may have limited access; otherwise (legal) journals access is very rare

Note: The numbers of faculty, graduate students and researchers given above are approximate, consensus estimates. To derive them we ran multiple, differently phrased questions through four well-known LLM models that undertake AI global searches (ChatGTP, Grok, Google AI main search, Google Gemini). We matched their answers with official data for the US, China, Germany, UK, France and India, and OECD research. In addition, we used data from a full input-output analysis of social science graduates and PhDs in the UK (in Bastow, Dunleavy and Tinkler 2014, Chapter 9).
Source: Authors.

of U.S. academics and researchers are on a tenure track with guaranteed access to paywall sources. In the UK two thirds of academics hold permanent jobs, but a third do not. Graduate students and retired staff leaving even a top ranked university may well lose their library access, falling into the academic 'precariat' shown in Figure 15.1. Outside specific internal labs or research units, firms and agencies almost never maintain libraries of their own, let alone covering regular journal subscriptions or regularly buying paywalled research books. So, for executives and most research staff in business and government agencies, OA to academic knowledge is critical. Outside advanced industrial states, across almost all of the global 'south' and the poorest countries in particular, paywall access articles and books have virtually no chance of being visible to the vast majority of university or external researchers. Most graduate students and even staff in Latin America, Africa and Asia can probably afford to access only 'bootleg' free versions of such articles or books via formally illicit downloading, with adverse implications for publishing as a whole.

By contrast, simple (one click, no conditions, no sign-ups) OA articles and books are vastly more useable and likely to be used worldwide. The trend to 'gold' OA journals has thus been highly positive for multiple groups of readers. OA articles and books also help incentivize using legal versions of outputs, and their digital availability is also key to working from home (WFH) in ways that academics find are superior for working on data, reading literature and producing manuscripts (Aczel, Kovacs, van der Lippe and Szaszi 2021, Figure 2). WFH also cuts transport costs and along with OA lets readers with disabilities participate fully in accessing academic work (Taylor et al. 2022).

Choosing an open access licence

An essential step for OA publishing is to specify a Creative Commons (CC) licence. The basic CC-BY form allows for the free attribution and reproduction of original components so long as author credit and sources details are given; it also allows reused text or materials to be modified. However, there are also more restrictive variants of the CC licences, as shown in Figure 15.2. The SA variant requires that any new materials created using original text or components are made available in forms of CC licence that are the same (or no more restrictive than) the original. It prevents scams where someone reuses OA materials, but then reissues them in a form more restrictive than their original licence. The NC variant allows reuse for non-commercial users only, but bans any *commercial* reuse. For instance, an OA textbook has potential commercial value, so the authors may choose an NC licence to prevent a potential scammer just reissuing the book in paywall form, charging any hapless users for access to their version. The ND variant requires that any materials being reused are kept in exactly the format they were originally issued in, without any modifications or remixing. This stipulation might be appropriate if researchers fear that research findings might be misunderstood by users, or in partisan or polarized contexts where disinformation is rife. Finally, these variants can be combined in different ways, as Figure 15.2 shows.

Figure 15.2: The main types of Creative Commons (CC) licences

Type of licence	What is *not* allowed	What is permitted
CC-BY Attribution	Using text or materials without attributing them to the original source and authors	The materials can be reused with a full source and authorship citation. Materials can also be remixed, adapted, modified or built upon when reused
Attribution-Share Alike CC-BY-SA	As with CC-BY above, but in addition materials used cannot be made available on more restrictive terms than the original licence	As with CC-BY above you can remix, adapt, or build upon a work, but if so any derivative works thus created must be distributed under the same licence as the original
Attribution-No Derivatives CC-BY-ND	As with CC-BY above, but no modifications of or additions to the text or materials are allowed	Text or materials can be reused only if they are *exactly the same* as in the original source, with a full source and authorship citation
Attribution-Non Commercial CC-BY-NC	As with CC-BY above, but no reuse for any commercial purpose is allowed – for example, reselling content is prohibited, along with any use as part of a business site that charges users	Only non-commercial users can reuse and modify materials, with a full source and authorship citation as with CC-BY
Attribution-Non Commercial-Share Alike CC-BY-NC-SA	As with CC-BY-NC above, no commercial use is allowed, but in addition any reusing of materials that creates a more restrictive licence than the original is prohibited	As with CC-BY-NC above non-commercial reuse of materials is allowed, but only so long as any materials created are made available on the same or similar licence as the original elements
Attribution-Non Commercial-No Derivs CC-BY-NC-ND	As with CC-BY-No Derivatives above, but in addition no reuse or modification for any commercial purpose is allowed – for example, reselling content is prohibited, along with any use as part of a business site that charges users	Only non-commercial users can reuse materials, and then only in exactly their original form (i.e. no modifications), and with a full source and authorship citation, as with CC-BY-No Derivatives above

Source: Creative Commons (no date).

Formats for open access articles

Many vague and often inaccurate posts and advice sites cover basic features of 'green' and 'gold' OA arrangements for articles. In fact, different publisher practices now mean that there are five different versions of OA (Tennant et al. 2020), which we compare with the legacy paywall model:

(i) *Paywall articles*. Authors submit papers at no charge, and the journal publisher finances the publication costs by making the accepted articles available only to journal subscribers or to one-off purchasers of the article (normally at high cost either route). The fewest number of potential readers have any access here.

(ii) *Embargoed 'green' OA articles*. Authors submit papers to journals at no charge, and the publisher issues them behind a paywall until an embargo expires. Only then can the authors (legally) deposit one open access digital copy on their university repository and another one on a personal website.

In the STEMM (science, technology, engineering, mathematics and medicine) disciplines the embargo times are lowest (6 months), because most science funder mandates require it. However, embargo times are far longer in the social sciences and humanities. Post-embargo STEMM authors can also normally post as OA on personal or university websites the 'version of record' as finally published in the journal, which benefits from reviewer and editor inputs, better formatting, and late-stage author corrections. But in other disciplines publishers often restrict any OA versions to just the manuscript 'as submitted initially' (ASI) with no reviewer or journal value-added elements.

(iii) *Immediate 'green' OA articles* – as for (ii) on the paywall arrangements, but as soon as a paper is accepted the author can make earlier digital manuscript versions of their paper (i.e. the final copy submitted) OA on their personal website and in a university or other repository. Less often, the OA version allowed may be the version of record exactly as published.

(iv) *'Gold' OA articles* – authors (or their universities or grant funders) pay an APC upfront to create instant OA in the publisher's journal. Authors can post multiple copies of the full article as soon as the article is published, for example on their personal websites, on university and other repositories, and elsewhere. Readers know that any article in an all-OA journal can always be downloaded. In any remaining 'hybrid' journals, however, OA articles funded by APCs may be published alongside closed, paywall-only articles funded by subscriptions. It may not be clear which is which, and some readers here may still be more reluctant to download or cite an article if they think of the hybrid journal as paywalled.

(v) *'Diamond' OA articles* – once an article is accepted the journal publisher makes the final form of the paper digital OA, and the article comes out in an all-OA journal, without the author paying any APC or other payment. Unlike gold OA, this type of OA does not discriminate against research from poorly funded or unfunded authors – publication depends on academic merit alone, not the author's ability to afford an APC.

Open access book formats

There are also five main forms of OA books, that we again compare with conventional paywall publishing:

(vi) *Paywall books* – authors submit a manuscript to a publisher and the firm organizes academic reviews of them. If the firm accepts a manuscript on the basis of reviews, then it is published at no charge to the authors. However, the text can only be accessed or read by purchasing either a digital or a paper copy of the book, both usually high-cost options that only wealthy academic readers or faculty at top global universities can implement. Multiuser ebook versions for university libraries (usually bundled up into composite deals across many books) may broaden students' and others' access a little. But they are sometimes pay per use, sometimes paid via annual subscriptions (often costly) and sometimes restricted to libraries that can afford large, bundled-up 'collections' – which anyway may not include all the publisher's books. Commercial ebooks often come with severe limits on how much readers are permitted to download (e.g. one or two chapters only) and/or on the 'borrowing' (i.e. visibility) time for full-access readers (as little as two days).

(vii) *Embargoed 'green' OA books* – as with the paywall model, authors submit books at no charge to publishers who sell accepted books solely via their paywall site for an 'embargo' period, plus purchases from shops or e-retailers. However, after the embargo time expires the author is allowed to deposit *one* digital OA copy of the book on their university repository and *one* on a personal website. These permitted copies may be the text as published. But usually in the social sciences the publisher only allows authors to deposit the manuscript as sent initially to the publisher (i.e. without any value-added elements or late-on revisions stemming from reviews, editing, formatting, etc.). The publisher carries on promoting and selling their paywall digital and paper formats of the 'versions of record' book (e.g. to libraries and other 'high-end' customers), indefinitely under most book contracts now.

(viii) *Immediate 'green' OA books* – as above in (vii), but authors can make digital versions of the book manuscript instantly accessible in a university or other repository and on their personal websites. Publishers usually only actively promote the digital versions of the final book via the publisher's paywall site, along with sales for physical paper copies. Some publishers may allow them to also be available from OA book aggregator sites like OAPEN.org.

(ix) *'Gold' OA books* – the author (or their university or grant funder) pays a book processing charge (BPC) upfront to make the book OA immediately in digital-only PDF form, somewhere on the publisher's website. Some university presses that promote OA books heavily offer BPCs that cost markedly less than those of commercial publishers (e.g. UCL Press 2024) and guarantee full promotion for the digital OA version of the book. As soon as the book comes out authors can upload copies of their book to their personal websites, to a university repository and to other sites. And the publisher also makes the book OA in some form or other, as well as continuing to sell digital and paper copies via their main website and retailers.

The best practice followed by leading OA publishers (like LSE Press) is to make both chapter and whole-book downloads available on the publisher's site (Watkinson 2024). (Some publishers may also provide the whole book files in multiple formats, such as Kindle, or e-pub, allowing readers to change typesizes and personalize the text – as with this book.) However, some commercial publishers only allow the book to be downloaded in PDF form, or alternatively only in separate chapter form (so that readers who want to get the whole book have to make as many downloads as there are chapters). Free copies of a gold OA book can usually be downloaded OA from somewhere on the publisher's site, but for commercial firms they may be harder to find behind 'confuser' screens (to ensure that search engines always take viewers first to the paywall paper or digital book versions sold on its main website). Some firms place their free ebooks on a separate, unpromoted and often purposedly hard-to-find part of their site that cannot be found from their main website. And most commercial firms offering an OA deal will not allow upload of these books to the key OA aggregator sites (like OAPEN.org or Google Books free books). An alternative commercial strategy (followed by Oxford University Press, for example) is to *only* make the OA version of book available on OAPEN.org, and not at all on their main site. Finally, a book chapter where the author or their university pays an OA fee (akin to an APC) may also be made OA on its own, even if the rest of the book is paywalled.

(x) *'Diamond' OA books* – here all accepted books (or chapters in an edited book) are published in digital OA form, at no charge to

the authors and with no delays or embargoes. The publisher both promotes the OA version extensively on their own all-OA site and also places the book on other key aggregator sites for readers to find via Oapen.org or Google Books free books. The only paywall element remaining is that the publisher normally still also sells paper copies of the book (using print on demand) directly and via e-retailers, sometimes at low-cost prices with OA university presses.

Newer open access formats

As OA develops further for articles and books, four new approaches have grown in prominence:

(xi) *'Flipped' OA books* – after a paywall book has been published for some time, the publisher's revenues normally drop off, sometime sharply. To 'flip' the title from paywall to OA the author (sometimes with support from their university or a funder) can pay the firm a relatively small fee (only a fraction of their normal BPC). The publisher posts the now OA version on their site (but again often in a hard-to-find location) and still sells digital and paper versions of the book at its paywall price to libraries or other high-end customers. If a book's sales have slowed to a trickle, publishers may agree to create an OA version and when first approached by an author will always ask for an unrealistically high price. However, authors should then ask the publisher for a 'net present value' (NPV) justification of this cost – this is an estimate of the money likely to be foregone for future sales through making digital OA copies available. Almost invariably for most academic books, the publishers cannot justify their initial asking price. When annual sales fall to low or even occasional-sale levels, then the publisher's NPV value for future sales will actually be tiny compared with an author offering an OA fee up front. Bear in mind too that even after selling digital OA rights, the publisher can still sell paper copies and their main-branded digital copies at paywall prices (e.g. to libraries). And if a flipped OA work becomes better cited and better known that change may actually generate more library etc. demand for paper or paywall-digital copies than a little known paywall-only version. As a result, if you haggle and demand that publishers back up their fee amounts with evidence from current and past sales, they will normally agree to digital OA in exchange for a much smaller fee than they initially suggest. Authors or their departments/universities can then potentially reboost awareness of the book (and generate more future citations) by proactively relaunching the OA version themselves. Flipping any worthwhile book will always reach a new

audience globally amongst those who could not previously find it in their university libraries or afford personal copies.

If an older book contract goes 'out of print' authors may also be able to freely scan and republish an OA version themselves, so long as the copyright reverts to them. Sadly, though, in the digital era publishers' recent contracts have kept many 'long tail' books theoretically available by listing only high-cost digital versions (that are permanently 'in print') and using print-on-demand technology to generate single paper copies of the book whenever needed.

(xii) *'Freemium' OA* – in this model a scholarly, scientific or serious output is made available free for anonymous users to download, with one-click access. But this OA offer is accompanied by other paywall or 'premium' services, such as additional facilities or extras (like datasets), or access to updates. Often employed by non-governmental organizations (NGOs) or international agencies (like the Organisation for Economic Co-operation and Development (OECD)), charities and business consultancies, this genuine freemium approach is used less by academic researchers.

(xiii) *A corrupted freemium model* gives readers OA, but only after their email address and other 'registration' data has been harvested. Private sector consultancies and firms regularly demand that you 'pay for' access to their reports like this with your personal information – but this should rarely happen for reputable academic organizations. Because many users object to data-harvesting, requiring registration is anyway often self-defeating in terms of broadening dissemination. Most users usually click away at the ID stage, making this a poor option for authors seeking to boost recognition or citations of their work.

(xiv) Other models for books and long reports include

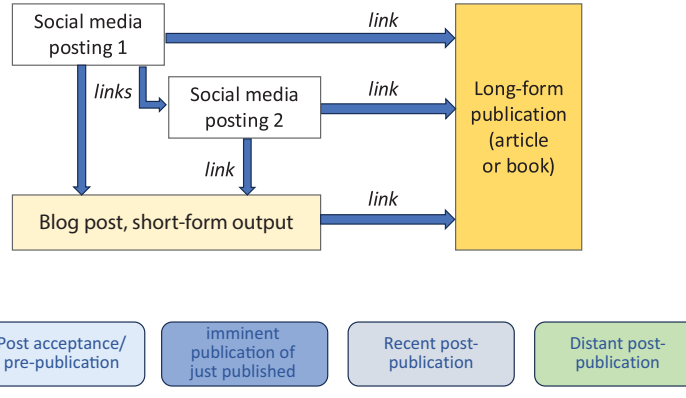
- '*Kick-starter' or crowdfunding*. Authors ask a large online mailing list of possible supporters to each make a small donation (say $10 to $50 each) to cover paying the BPC costs with a lower-cost OA publisher. The book or report only goes ahead if donations reach the critical mass needed.
- *Subscription models* are similar, but authors seek larger per-head donations from a smaller (but still medium-sized) group of potential sponsors, asking them to commit in advance to part-finance a BPC with others.
- *Direct subsidy for the whole BPC* from a philanthropic donor, university or department sources, or grant-funding, may often be available for OA books written by their faculty – either using the university's own press or an external firm. Some university presses developing low-cost OA will also fund the BPC for an otherwise unfunded book themselves.

15.2 Active research dissemination in the digital era

For all research books, whoever publishes them, authors now need to move past a tradition of academic diffidence about publicizing their own articles and especially books (Portwood-Stacer 2021b). Attitudes amongst faculty that deprecate 'self-promotion' are partly dated echoes of eighteenth century aristocrats looking down on 'trade' occupations. However, they also fit with some more legitimate concepts of academic politeness, such as avoiding overclaiming for 'new' work, and respecting scientific professionalism. In practice, none of these considerations has stopped an increasingly improbable rate of claims to having achieved scientific success or innovation (Morris, MacGillivray and Pither 2021). The way to meet legitimate expectations of professional restraint and modesty is not to publish helplessly into a vacuum in a self-defeating manner. Instead, researchers should rediscover responsible agency and promote their new research directly, in factual and proportionate ways that enhance academic debate and seek to inform a critical audience.

In the digital era, all authors (fiction and non-fiction alike) must play a primary role in promoting their books, being proactive online at all stages of the publication process in alerting potentially interested readers and users to their work's key messages and imminent or extant availability. For academic books this can be far less onerous than past practices of authors going to hugely time-consuming conferences, or flying around to give seminars to 'plug' their article or book to typically tiny audiences (Portwood-Stacer 2021a, 2021b). These legacy activities are still OK to do if you have finances or the chance to do so. But in the OA era authors should prioritize more effective things, such as those shown in Figure 15.3.

Figure 15.3: Pro-actively building knowledge of open access research publications



Source: Authors.

These specific steps are helpful:

- Compose and post at different times and in distinct venues several ultrashort *social media messages* – for example, tweets on Blue Sky and X, Facebook content, brief LinkedIn posts, or possibly TikTok micro-videos, etc. Social media messages can normally be copied across from one channel to another at low time cost. Rigorously eschew vacuous, formal or self-congratulatory ('humble brag') postings that give readers no useful immediate information (e.g. just the title of an article or a picture of a book cover or a seminar presentation). Instead, each post or message should give readers or viewers some specific 'added value', substantively explaining key findings or arguments from the research in a snappy way (Dunleavy and Tinkler 2021, Chapter 7).
- All social media messages now should ideally include a JPEG or image of a key figure or diagram from the paper or book, or a short explanatory video. Alternatively, if no graphic is possible, take a picture of the article abstract or the book blurb (in a readable typesize) so that interested readers can see a summary of the whole argument being made, whatever the word limit of the media used.
- For a book, see if you can get advance endorsements from senior people in your field (who are not linked to you) in time for publication. (Incidentally, never seek endorsements from people in your own university or department, or from recent co-authors.)
- If a preprint stage is feasible, post a finished version of an article or book chapter to *arXiv* (no date), or the Open Science Foundation OSF.org site (no date), or a similar preprint server just prior to submitting it for review. Keep the preprint available during the review process itself, which can last a long time – it helps prepare the ground for the later finished article. Meantime, it may help generate reactions from fellow researchers that will improve the final piece and smooth the path with reviewers. When the text can first be downloaded, issue a social media post (maybe copied across forums) to alert this specialist audience.
- Once a publication is pending (having been peer reviewed, revised and accepted), one or several more elaborate social media posts can seek to actively attract readers. Ideally, they should link to an accessibly written blog post or other short-form output. For an OA research book, ideally at least one 'trailer' blog article should come out a week before the book launch, and a different one on publication day (or shortly after), both with full download details for the article.
- Make sure your publisher set up early reviews of books in the main online book review sites for the social sciences and humanities, especially *LSE Review of Books*, *LA Review of Books*, and possibly the *London Review of Books* or *Boston Review* for a major work (links in references). Normally a press person (not the book author) needs to pitch the book to the blog's review editor first, then send them a paper

copy of the book for the reviewer to use and keep. But despatching a digital version before publication day can also greatly speed up getting a really timely review out. Speedy book reviews, endorsements and recommendations online are vital steps for building early downloads of OA books.

- After publication every social media message about a book, article or chapter should include a one-click link to it, via a clear DOI number, a stable URL or a QR code. Potential readers must be able go straight to the full piece, usually by reading the article (or chapter) abstract or a book blurb to decide if they want to download it. For books they may also have to decide if they want the whole book or just a chapter or two.
- Be aware that many readers on social media still find it a daunting prospect to download and decode long-form academic texts, especially if they use maths formulae in the abstract, or esoteric scientific or social science jargon in the title. Will they understand what is said? And is the content as relevant for their interests as it might seem? This is why authors should also always provide a *short-form output* (perhaps even two or three) designed to be accessible for the widest range of readers. Give the editors for possible blogs or outlets several weeks' notice of when you would like blog posts or short-form pieces to come out. Some ask that you pitch a (one-pager) proposal to them, but others want to see a full piece before giving a response.
- The best short outputs tell *a substantive, well-written and front-loaded story* – with the key findings coming right at the start or middle of the piece, and not at the end (Dunleavy 2016). Examples include a 500–1,000-word blog post for a major academic blogsite with substantial viewer numbers, like the LSE's major academic blogs and those for many social science professional associations now. Consider also a long LinkedIn post, an article in a relevant (UK, U.S. or Australian) edition of *The Conversation*, or an OA magazine article (of maybe 1,000–2,000 words). The larger the audience that might be reached, the more effort is justified. It speeds up acceptance by editors if your blog post or other text is in a format that already fits the outlet's style. (For many academic blogs this may mean URL referencing in the main text only, with no footnotes or endnotes.) Wherever possible, try to seek out academic blogs and podcasts with a ready-made audience, shown by their having substantial numbers of LinkedIn, Facebook or BlueSky followers.
- A short (five minutes maximum) podcast or videocast of the author(s) explaining the article or book conclusions, ideally with graphics, might also work well to publicize a major article or a book. Short interviews (under five minutes) on podcasts or webcasts can get intuitively understandable content across accessibly, even for expert audiences.

- Every short-form output (blog posts or podcasts) should include one or two links to the OA full text, ideally using the DOI number in web format (i.e. beginning with https://doi.org/ and then the number).
- By contrast, posting a blog on a small department or university site with few outside readers is not usually an effective step, especially if that blog's outputs are a mix of academic pieces with public relations puff materials for a department or with material written by students. The only usual value of such a post is to create a digital record for you to publicize and link to via social media. However, if a department or lab already has a well-developed and solely academic working paper or research paper series going, then getting a summary of the final research as a blog post or email newsletter item circulated to a large mailing list of subscribers with relevant sub-field interests can be helpful.
- A long conference or seminar presentation posted on YouTube, or another videocast or podcast lasting over five minutes, might be valuable social media accompaniments to a book, or even for a key article. But for most social media users listening to them, or watching them, will be too time-consuming; they are typically seeking to get to the gist of things in the shortest possible time. Consequently, these options are best thought of as *supplementary long-form outputs* (Figure 15.3). They can complement the full text for super-committed readers who are looking for a more accessible understanding than the article or book itself and have the time to devote to doing that. This may include interested people in the general public. Later on, long presentations like this can also get included in reading lists and help students preparing for classes.

All these different types of output are also relevant to consider at the five different phases of publication shown at the bottom of Figure 15.3:

- When submission is near or reviewing is underway, assuming a preprint or finished version of the text is available.
- After an article or book is accepted, in the run-up to publication.
- When the full text output is newly available and fresh (within (say) the first two or three weeks after publication day).
- In later weeks in a more 'consolidating' style for say a month or two after a book has come out (and perhaps an article that is getting a good reaction).
- Well after publication, when the output may form part of the 'long tail' of only occasionally cited outputs. At this stage, if the article or book is digitally available then downloads may be increased by social media messages that stress that it can be accessed easily, making the full text stand out from the whole cloud of digital items suggested in Figure 15.3.

Conclusions

Open access publishing is the easiest 'open' action to start doing, and it is vital in reaching all the readers who can benefit from or improve research. For articles, all-OA journals are now the norm, and so the pace of future change here may slow down, at least for tenured faculty in journal-focused disciplines and located in the top 2,500 universities globally – the institutions best able to fund free APCs with publishers in return for continuing bundled journal subscriptions. And readers across the rest of the world should have near-universal OA to social science articles. In addition, however, less advantaged authors, those outside tenure ranks or top universities, are also likely to see a major broadening of their ability to publish OA articles for free, because of more diamond OA journals publishing by universities, university presses and other bodies, plus the development of other direct funding models (see above).

Publishing of OA books has also greatly increased from universities, university presses, professional bodies and NGOs, while some commercial and legacy publishers have also embraced OA but with high book processing charges. There are still high levels of resistance to OA books from commercial publishers and some professional associations in social science. Even where commercial publishers appear to offer digital OA books, it is important to examine carefully how available and findable these OA texts are compared with their remaining paywalled versions. New university presses (eschewing 'legacy' models), plus progress in operationalizing 'diamond OA' book publishing, are both leading to major changes here. OA books will also be boosted by research funders increasingly demanding that all research that they finance should be immediately open access (UKRI 2025).

The key drivers for OA remain researchers who take seriously their professional responsibility for actively promoting their long years of work to academic and 'impact' audiences alike. Authors who can anticipate and meet both their academic and outside audiences' needs are the best direct exponents of their own work, and its most careful re-expositors. Contemporary digital social media mean that researchers can reach different sets of professional, informed and general public readers without having to fight their way through the fog of 'mainstream' media or 'journalistic' priorities (Dunleavy and Tinkler 2021, Chapter 7). Research authors themselves can best shape and distribute OA messages about their work, in both long-form and short-form ways, not constrained by the medium or reliant on mediators.

Notes

* Section 15.1 was partly developed from an online joint presentation and seminar on OA publishing on 15 September 2020, run by Frances Pinter (of Central European University) and Patrick Dunleavy and forming part of Sciences Po's input into the Civica Programme. (Sciences Po,

Paris is also known as École des Sciences Politique or the Paris Institute of Political Studies.) See Pinter and Dunleavy 2022a, 2022b for the PowerPoint slides and YouTube video, respectively. We thank Frances Pinter for her ideas and inputs and all the participants at this session for their questions, comments and suggestions. CIVICA Research was funded by the European Union's research and innovation programme Horizon 2020, under the chapter 'Science in Society'.

[1] For a few years a backlash movement in the humanities against the faster pace of neoliberal university research actually advocated 'slow professor' practices, allegedly needed to induce more reflective or creative research (Berg and Seeber, 2016; Lefort-Favreau, 2019). Critics argued that some of the arguments involved just promoted the vested privileges of tenured academics amidst university resources shrinking back from past (perhaps unsustainable) 'golden age' levels. In practice, more study and reflection time for tenured faculty may chiefly be gained at the expense of a burgeoning 'academic precariat' who handle major teaching loads on short untenured contracts or get locked into doing long cycles of successive research projects.

References

Aczel, Balazs; Kovacs Marta; van der Lippe, Tanya; and Szaszi, Barnabas (2021) 'Researchers working from home: benefits and challenges', *PLoS ONE*, vol. 16, no. 3, article e0249127. https://doi.org/10.1371/journal.pone.0249127

American Association of University Professors (AAUP) (2023) 'Data snapshot: tenure and contingency in US higher education', Web page, March. https://perma.cc/8Z6K-WWJZ

ANU Press (Australian National University) (2025) Home page. https://perma.cc/DMM9-DH6R

arXiv (no date) Search home page. https://arxiv.org/

Aspesim Claudio and Brand, Amy (2020) 'In pursuit of open science, open access is not enough', *Science Magazine*, 8 May. https://doi.org/10.1126/science.aba3763

Bastow, Simon; Dunleavy, Patrick; and Tinkler, Jane (2014) *The Impact of the Social Sciences: How Academics and their Research Make a Difference*. London: Sage. **$** https://doi.org/10.4135/9781473921511

Berg, Maggie and Seeber, Barbara (2016) *The Slow Professor: Challenging the Culture of Speed in the Academy*, Toronto: University of Toronto Press. https://utppublishing.com/doi/book/10.3138/9781487521851

Boston Review (2024) *Boston Review*, book pages. https://www.bostonreview.net/tag/boston-review-books/

Castille, Christopher M.; Kreamer, Liana M.; Albritton, Betsy H.; Banks, George C.; and Rogelberg, Steven G. (2022) 'The open science challenge: adopt one practice that enacts widely shared values', *Journal of Business and Psychology*, vol. 37, pp. 459–467. https://doi.org/10.1007/s10869-022-09806-2

Creative Commons (no date) 'About CC licenses', Creative Commons Foundation. https://perma.cc/ZEV9-SZ59

Crow, Graham (2021) *Academics Retiring, Scunnered or Otherwise*, Edinburgh: University of Edinburgh, Report. https://perma.cc/2LHG-M9L8

DOAB (2024) Directory of Open Access Books, home page. https://directory.doabooks.org/

Dunleavy, Patrick (2016) 'How to write a blogpost from your journal article', *LSE Impact blog*, 25 January. https://perma.cc/ZLJ3-UR5E

Dunleavy, Patrick and Tinkler, Jane (2021) *Maximizing the Impacts of Academic Research: How to Grow the Recognition, Influence, Practical Application and Public Understanding of Science and Scholarship*, London: Palgrave (now Bloomsbury Press). **$** https://www.bloomsbury.com/uk/maximizing-the-impacts-of-academic-research-9781352010992/

Dr Foster (no date) 'An introduction to pre-print servers', EU project page. https://openscience.eu/preprint-servers-an-introduction

Google Books (2025) Search page for Google Books free titles. https://books.google.com/googlebooks/about/free_books.html

Green, Toby (2014) 'Fremium open access', in *Praxishandbuch Open Access*, Berlin: de Gruyter, pp. 127–136. https://doi.org/10.1515/9783110494068-015

HESA (2023) Higher Education Staff Statistics: UK, 2021/22. https://www.hesa.ac.uk/news/17-01-2023/sb264-higher-education-staff-statistics

JSTOR (2025) 'Path to open – high-quality titles with a high impact', Web page. https://about.jstor.org/path-to-open/titles/

Laakso, Mikhael (2022) 'Open science and open access publishing in social sciences', in Engels, Tim C.E. and Kulczycki, Emmanuel (eds), *Handbook on Research Assessment in the Social Sciences*, Chichester: Elgar, pp. 105–118. **$** https://doi.org/10.4337/9781800372559 OA at: https://perma.cc/C7WN-A3KP

Lantsoght, Eva (2023) 'Book Review: "The Slow Professor"', PhD Talk blog, 17 January. https://perma.cc/X9QY-2EBU

Lefort-Favreau, Julien (2019) 'The 'slow professor' could bring back creativity to our universities', *The Conversation*, 25 August. https://perma.cc/9KRV-P78C

London Review of Books (2024) Home page. https://perma.cc/FS6G-Q9QR

Los Angeles Review of Books (2024) Home page. https://perma.cc/B2X5-RJZT

LSE Press (2025) Home page. https://press.lse.ac.uk/site/

LSE Review of Books (2024) Home page. https://perma.cc/G8YD-AJK2

Morris, Douglas W.; MacGillivray, Erin; and Pither, Eylse N. (2021) 'Self-promotion and the need to be first in science', *FACETS*, vol. 6, pp. 1881–1891. https://doi.org/10.1139/facets-2021-0100

OAPEN (2024) OAPEN 'Online library of open access books', website. https://www.oapen.org/

OECD (2017) 'Education at a Glance 2017 - OECD INDICATORS,' https://www.oecd.org/content/dam/oecd/en/publications/reports/2017/09/education-at-a-glance-2017_g1g7f6b6/eag-2017-en.pdf

OSF Open Science Foundation (no date) Home page. https://osf.io/

Pinter, Frances and Dunleavy, Patrick (2022a) 'Open access publishing for research', CIVICA seminar, Seminar PowerPoints, 15 September. https://zenodo.org/records/7622179

Pinter, Frances and Dunleavy, Patrick (2022b) 'Open access publishing for research', CIVICA seminar, YouTube video, 15 September. https://www.youtube.com/watch?v=4Bc3c7b6XNU&pp=ygUPRHVubGVhdnkgQ0lWSUNB

Portwood-Stacer, Laura (2021a) 'Promoting your academic book', Manuscript Works blog, 21 July. https://perma.cc/59QM-A8RD

Portwood-Stacer, Laura (2021b) 'Do you really have to give book talks?', *Reassurance for Introverts*, 19 August. https://perma.cc/4D3D-JFRY

Taylor, Heather; Florisson, Rebecca; Wilkes, Melanie; and Holland, Paula (2022) *The Changing Workplace: Enabling Disability-Inclusive Hybrid Working*, London: Work Foundation, online report, July. https://perma.cc/4M75-KA7B

Tennant, Jonathan P; Agrawa, Ritwik; Baždarić, Ksenija; Brassard, David et al. (2020) 'A tale of two 'opens': intersections between free and open source software and open scholarship', Preprint at Open Science Foundation. https://perma.cc/8DPH-MPHZ

UKRI (UK Research and Innovation) (2025) 'Progressing open access for books at UKRI', webpage, 10 February. https://perma.cc/DAF4-7N8B

UCL Press (2024) 'Financing your book', webpage. https://perma.cc/A2ZK-FTPU

UCL Press (2025) Home page. https://perma.cc/TWH9-XLSA

Watkinson, Charles (2024) 'University Press redux, Plenary 4; Reclaiming Open Access for the University Press Community' University of Michigan Press presentation to Association of Learned and Professional Society Publishers Conference, University of Edinburgh, 15–16 May. https://umich-my.sharepoint.com/:p:/g/personal/watkinc_umich_edu/ESUdpnl09U1EpNDmquXdHs0BXG9tY4IUIDtYU7-RVL6oJw?rtime=wmgZarJ13Eg

16. Conclusions – achieving open social science

The social sciences face a substantial challenge to match the science, technology, engineering, mathematics and medicine (STEMM) disciplines' established progress in building up open science (OS) as a cornerstone of new academic work (Leonelli 2023). Even this step forward demands a substantial re-evaluation of how research is currently planned, developed and published in many disciplines. But the challenge set in previous chapters is not just to catch up on and adapt earlier versions of OS, but to accomplish a more profound cultural change to openness across the social sciences, one that can renew core foundations for academic progress in more cumulative and sustainable ways. Open social science (OSS) demands some fundamental improvements in how our discipline group undertakes research, how progress is achieved and assessed, and how researchers are rewarded. But the effort is well worthwhile, both in academic and in impact terms. OSS safeguards the future of the discipline group by creating a far more reliable basis for citizens, organizations, firms, governments and civil society organizations alike to use social science knowledge in setting their own policies and planning actions.*

Previous chapters have noted some substantial barriers that remain to accepting OSS approaches. We very briefly recap them here, together with an array of countervailing drivers and mitigations that can help researchers to overcome issues (Section 16.1). These foundations for further progress are now firmly based across the discipline group, and new solutions to past problems are constantly being devised, as earlier chapters have demonstrated. Next we consider the two biggest external factors behind the momentum to greater openness:

- digital decompression changes in knowledge regimes within all modern organizations and civil society; and
- the importance of 'democratizing innovation' and team building in growing the 'dynamic knowledge inventory' of modern civilization (Section 16.2).

How to cite this book chapter:

Dunleavy, Patrick and Monteath, Timothy (2026) 'Conclusions – achieving open social science', in Dunleavy, Patrick and Monteath, Timothy (2026) *Doing Open Social Science: A Guide for Researchers.* London: LSE Press, pp. 363–387. https://doi.org/10.31389/lsepress.dos.p

16.1 Barriers and drivers for open social science

Making changes in any organizational culture can seem threatening for established personnel and their interests, especially a form of change like OSS that is extensive or encompassing. It is perhaps natural enough that a reluctance to embrace openness remains amongst social scientists. Figure 16.1 sets out 15 main issues here, and alongside them also lists the suggestions covered in earlier chapters for mitigating the difficulties involved.

Rather than rerunning our earlier arguments, summarized in Figure 16.1, we turn instead to the nub of much potential resistance to OS, which is a reluctance to rethink the legacy norms and values of career advancement in the social sciences by some already-established faculty and 'high flying' early career researchers (ECRs) at elite universities. The essence of this set of attitudes is that only research publications in top journals count (or should ever count) towards academic appointment, promotion or professional reputation in one's discipline. So PhDers, ECRs and younger faculty (still at the 'cutting edge' of new methods and tech in their discipline) should single-mindedly focus just on crafting articles for prestige, highly cited journals. They should refuse to be distracted by any other considerations than the 'big grind' of gaining top journal slots. This elite model has been minimally updated now, but only to include (i) open access publication in top journals (in practice fundable only by top universities, the key ones for professional advancement anyway); and (ii) depositing data where funders require it.

The progress of OSS on reproducibility has also lagged, and in U.S. and European elite universities data openness remains hit and miss.

> Understanding academia as a reputation economy helps to explain data withholding and selective sharing. Reputation in the current academic system is for the most part tied to high impact (journal) publications … whereas intermediate products, such as data, have little to no exchange value. This explains why researchers strategically withhold datasets until they have published every aspect of them, as sharing too early would risk a competitive disadvantage (Fecher et al. 2017, p. 7).

Studies of author promises about replication archives in social science journals consistently show that either deposits have not occurred or that links to them no longer work (Key 2016; Janz 2015; Krawczyk and Reuben 2012), a pattern that has also endured in STEMM disciplines. Many senior researchers have filing cabinets full of 20 years of unpublished work that they still nominally plan to 'publish when time allows' (Silvia 2007, p. 77). However, more and more top UK and European institutions now go beyond requiring that research datasets be deposited only when they are fully milked for publications, or where funders compel it. A better general

Figure 16.1: Fifteen key barriers for researchers in adopting open social science, and the main countervailing drivers for continuing innovation

Problem	Barriers or problems	Countervailing drivers or incentives for change, or mitigations for problems
A. Researcher insecurity and worries	Fear of scientific vulnerability and external critique	– Share research design anxieties with trusted colleagues – Recognize that there is no 'perfect method' or 'perfect code' – Add new 'open' practices to your approach incrementally, drawing confidence from experiences as you go (Castille, Kreamer, Albritton, Banks, et al. 2022). – Emphasize growth and learning
B. Fears of research ideas being 'scooped'	Characteristic PhDer or ECR worry that work might be pirated or exploited by others. Data is kept private – 'just in case I can get round to analysing or publishing it in future'	– Pre-registering immediately and clearly establishes the precedence for your ideas – Be confident that you know your own data best and so can best exploit it – Be realistic about the likelihood of ever having time or resources to come back to data kept secret 'just in case' (Silvia 2007, p. 77)
C. Time pressures	Time available for 'writing up' or navigating publication processes get compressed. Sharing data at a late stage takes scarce extra time. Starting work on new projects often cuts into the time once pencilled in for wrapping-up old project tasks, like data depositing. Funding runs out fast at the end of projects	– Embed 'open' practices from the outset of projects – Anticipate workflows better or more realistically at the outset – Allow more time for cleaning and depositing data, plus publications – Create a well-founded data management plan (DMP) – see Chapter 7) – Leverage existing data, workflows, models, and especially open source tools (Tennant et al. 2020) to save time – Use existing templates

(*continued*)

Figure 16.1: (continued)

Problem	Barriers or problems	Countervailing drivers or incentives for change, or mitigations for problems
D. Weak incentives	'Open' practices are seen as a community gain but as extra pain/costs for authors, with few individual payoffs for building careers or CVs	– Open access to your data improves your research's provenance and reputation, visibility, citations and potential for impacts – Open datasets on their own create *additional* items attracting citations and build positive views of your concern for data quality in the sub-field – Reproducibility strengthens academic reputations – Pre-registering a research plan can generate early comments that help avoid mistakes and improve the quality of research and analysis – Universities and professions need to better recognize and reward resource-sharing (Shibayama, and Lawson 2021)
F. Complex workflows	Lack of training inhibits reusing previous materials; reusing other people's data requires specific skills that are less taught than how to generate new data from scratch	– Attend an open social science (or open science) training course and implement changes to your research practices after it (Heise et al. 2023) – Write a detailed 'readme' file to lodge with data you deposit to help follow-on users
G. Sharing rights	No experience in setting licences or sharing rights induces keeping data closed	– Use one of the Creative Commons licences for datasets and for publications (Section 15.1) – Follow other authors' approaches
H. Datasets are too large or otherwise unique or hard to handle	This is more of a problem in STEMM subjects and some social science sources, such as market feeds information. Relatively few open social science datasets will breach feasible storage limits for professional software systems. But data limits are restrictive in some software, such as Excel	– Split data into smaller chunks – Share the properties of the data – Lobby for greater access to storage – Seek external storage or upgraded analysis solutions (where feasible, given possible privacy/ confidentiality issues)

Figure 16.1: (continued)

Problem	Barriers or problems	Countervailing drivers or incentives for change, or mitigations for problems
I. Sensitive issues and content in the research or data	Ethical commitments to research subjects, conditions for gaining grants or university research permissions, and often national data laws, all now require close compliance with anonymity or non-attributability conditions	– Anonymize data in intensive ways – Aggregate data above small area level – Report only theme-fragmented data analyses for general readers, with access to full data for researchers only – Generate synthetic datasets – Provide only sample data from big datasets
J. Worries about reusing other people's data	Fears over using existing datasets that are non-academic in origin; or unconventional data; or that are 'messy' and imperfect. Restrictive attitudes amongst PhD supervisors or post doc reviewers on legitimate data sources. Unease over using commercial datasets whose provenance is limited by commercial confidentiality	– Remix new data and iterate with existing data to show its quality – Involve previous dataset-creators – be willing to share work in progress and get feedback – Using genuinely big data resolves or mitigates most commercial dataset provenance worries – Test a dataset for biases before committing to using it – Review models for biases and assess the sensitivity of analysis
K. Risk of critique or inappropriate use	Anxiety that more transparent research will be critiqued by professional rivals or opponents; or may be misquoted or misinterpreted by partisans	– Write detailed metadata to show the data's strengths and limits. – Set repository access conditions that retain some control – for example, the ability to exclude partisan or non-qualified reusers if really needed. – Be willing to help new users to understand the data better

(*continued*)

Figure 16.1: (continued)

Problem	Barriers or problems	Countervailing drivers or incentives for change, or mitigations for problems
L. Diffidence about outreach	Lack of 'impact' training or skills. Weak experience of communicating to non-academic audiences. 'Academic' diffidence and reluctance to be proactive in promoting publications and research outputs	– Take up opportunities for impact and media training – Build ties with research participants and their communities – Use citizen social science techniques where feasible/ relevant – Go beyond open access publishing to encourage or create open educational resources – Exemplify 'open science' virtues – Amplify open social science projects and leaders
M. Who will want to share this data?	Uncertainty over the relevance of research for others may deter open social science practices like data depositing. Researchers often underestimate the extent and nature of other authors' interest in their work, which cannot be easily forecast	– Share research designs as soon as possible – Use new knowledge quickly – Avoid constructing or depositing data in proprietary formats – Use widely available formats for depositing – Provide detailed metadata for deposited information – Disseminate key findings and messages about the datasets used
N. Unclear about what the value added from open social science is	Datasets are seen as very specific or of no value to others	– Value is subjective – Reused data in whole or part can have value for an extraordinary range of scholars, one that is hard to forecast in advance
O. Worries over potentially transient storage or availability	'Keep it in the filing cabinet' locks up data from any potential reuse issues	– Use stable and large data repositories. – Provide and deposit supplementary details for exhibits included in journal articles in replication archives, and for chapters and books in research methods annexes

Source: Authors, but also drawing on (Gomes et al. 2022; Gownaris et al. 2022).

practice now expects researchers to prove the 'replicability' of their work for serious CV-building.

In terms of OSS's push for wide dissemination of research, more social scientists will now diffidently promote their work (usually only once) on social media. And 'impact' factors are included (at some level) in promotion assessments in most universities, but less often in making new appointments. In most social sciences citation counts are still routinely disparaged as indicators of quality by 'old school' academics. Faculty at elite universities with only barely cited articles or books can still achieve tenure, be promoted to full professor, or even be elevated to national academies – especially in the humanities. Many 'hermit' faculty at elite universities *still* have no well-managed Google Scholar Citations (GSC) accounts, even though it has been the most important way in which other researchers and external users can find and understand their work for more than a decade (Dunleavy 2014a; Gerasimov et al. 2024).

Yet the advance of OSS norms and values already implies that the legacy models of career advancement counselling 'closed' research-grind behaviours are increasingly anachronistic. Instead, career advancement now is strongly affected by a great many factors. Figure 16.2 sets out and rates 15 positive influences on contemporary career advancement, both for journal-dominated disciplines and for those where books and chapters remain integral to research alongside journal articles. Eight of the CV-boosting influences are critically dependent on and shaped by adopting OSS practices (shown by the **O**s in the first column), and two more are positively improved by OSS (shown by the diamonds ◊).

A recent paper showed that legacy practices were little changed in U.S. economics in the early 2000s. There, achieving a single publication in one of the big five journals – *The American Economic Review*, *Econometrica*, the *Journal of Political Economy*, the *Quarterly Journal of Economics* and the *Review of Economic Studies* – could sometimes determine whether an early career researcher got tenure at a top university, or was forced to restart a career at a less-rated institution (Heckman and Moktan 2020). Younger economists obsessed about securing a top five (T5) journal slot as a key to securing tenure in a prestigious department, often taking years to achieve that goal. Yet the study authors also concluded that even in their study period (10–25 years ago now) the T5's picks of articles to publish did not include many of the most cited works in the period. Most major economists also published extensively in more specialized field journals. And the legacy system's overall effects were strikingly adverse for creative research in the discipline:

> Emphasis on the T5 in sorting talent creates a culture where vitae length and publication speed in select journals rather than the development of a body of coherent and original ideas is most valued. It incentivizes careerism rather than creative scholarship (Heckman and Moktan 2020, p. 464).

Figure 16.2: Our estimates of the importance of different factors for academic career success, in article-dominated and mixed social science disciplines

Factor	*Importance for career development pathways in*	
	Article 'mono-culture' disciplines	**Mixed article and book disciplines**
Articles in 'top' journals ◊	Very important for tenure in some fields (e.g. economics)	Important for tenure
Articles in leading field/ specialist journals ◊	Important, for appointment and tenure especially	
Articles in other journals	Useful, for all stages	Important, for all stages
Citations **O**	Important for promotion and betterment, useful for tenure	
Research books **O**	Useful for promotion and betterment	Very important for tenure and promotion to full professor ('having a distinctive voice')
Chapter in books **O**	Potentially helpful at all stages (especially if in influential books and available as digital open access)	Useful for all stages, especially if digital open access, and in some particular fields
Depositing research data openly, promptly and in well-documented ways for reusers to access **O**	Very important now for how research outputs are assessed by the profession	Important for how research outputs are assessed by the profession, and resulting in improved citations (Colavizza et al. 2020)
Pre-registering research designs, methods and hypotheses/themes **O**	Important for how research outputs are assessed by the profession	Useful for how research outputs are assessed by the profession
Recommendations or 'letters of support' by senior academics at other universities	Very important for tenure in some uber-hierarchical disciplines and countries (e.g. U.S. economics) Important for promotion and betterment	Important for tenure, promotion and betterment

Figure 16.2: (continued)

Factor	*Importance for career development pathways in*	
	Article 'mono-culture' disciplines	**Mixed article and book disciplines**
External reviewers' reports to home university on an author's submitted works	Important for tenure, promotion and betterment	
Grants and funding success **O**	Important for tenure, promotion, betterment	Useful for tenure, promotion, betterment
Teaching record and student responses	Important, at all stages	
Teaching books/ textbooks	Useful for promotion, especially if widely used or seen as integrating the discipline's consensus view	
Taking part in professional service tasks **O**	Useful for tenure and promotion	
'Impact' profile **O**	Useful for tenure, promotion and betterment	
Public and/or media recognition **O**	Useful for promotion and betterment, so long as academic quality is also kept high. If not, gaining a reputation as solely a 'popularizer' may not always be an asset	

Note: We define career success as having four key stages – securing initial appointment; winning tenure; getting promotion across ranks while in place (e.g. moving up from assistant professor to associate professor and then full professor), and 'betterment moves' by shifting to more prestigious universities or departments. We rate career importance at four settings: Very important; Important; Useful; Potentially helpful.
O indicates a strong improvement with OSS.
◊ indicates a positive effect with OSS
Sources: Authors' personal estimates only – but informed by Heckman and Moktan (2020) and Fortunato et al. (2018).

Similar insights apply in business and management departments, where a grading system is widely used for coding the prestige of journals, with a top category of 4* and the second rank as 4. An article coming out in these prestige journals has greater 'CV boost' value than the same article elsewhere. But often 4* or 4 journals may not be cited as much – for instance, because they occur in generalist, profession-run journals while publishing in more specialist, applied or cross-disciplinary journals can secure more citations for influential papers. A recent study set out to estimate what the 'cost' was in

terms of lost citations that business academics were prepared to pay in order to have their work come out in the conventionally high-prestige outlets:

> We empirically demonstrate that the willingness to forgo citations to publish in 4* journals is strongest among academics who have already published in 4* and/or 4-rated journals. Contrary to our expectations, we find that an individual's existing scholarly impact, as captured by prior citations, has no effect on this preference. We also show that academics working in high-ranked institutions would give up more citations for 4* journal publication than those working at lower-ranked institutions.
>
> It is important to acknowledge that there is still considerable diversity among business and management academics in their willingness to 'pay' in terms of scholarly influence [*or citations foregone*] for publishing in 4* outlets, with many academics appearing immune to the lure of journal status in preference to academic influence (Salandra, Salter and Walker 2022, p. 1254). [*Our italics*.]

Despite the enduring quality of past attitudes, the implications of Figures 16.1 and 16.2 are that the legacy models of closed, paywalled and non-reproducible publishing are fading in the open access and OS era, as a new, more balanced and multivalent set of considerations about building better professional community prevails. In an increasing number of ways, and across more and more social science disciplines, systemic arguments for open access to data and knowledge have begun to edge out solely individual reward incentives. Growing numbers of both senior and early career researchers now embrace OSS ideas as desirable in both research quality and ethical terms. They may be likely to look somewhat less kindly on candidates for positions or advancement whose publishing strategy appears solely self-aggrandizing. The implications for tenure decisions, job openings and promotion decisions are increasingly that researchers who take OSS seriously will come up for tenure or review with core research offerings that are seen as more transparent and effective in academic terms, and hence more impactful. They will also tend to have a full range of ancillary publications, deposited datasets, dissemination and impact outcomes, and professional contributions, which authors wedded to the 'grind out only articles' legacy model will struggle to match.

16.2 Digital decompression and the momentum for OSS

Across the whole history of humanity:

> Collecting, storing, and analysing data were often hard, time consuming, and costly. [So] the difficulties in handling data prompted

> humans to use as little data as possible. The very methods and techniques, the structures, and institutions of discovery were designed so that the most insights could be squeezed out of the least amount of data (Mayer-Schönberger 2016, p. 996).

In every pre-digital organization there was always a pressing need to ruthlessly select a tiny share of data and information to be recorded (making the ditched information unrecoverable), and to code and condense even retained data in radically shortened forms before filing it with some (primitive) form of separate index data, that might later allow for the retained data to be selectively reaccessed.

Digital era advances in numerous analysis, storage and communication technologies have now begun to make such drastic data selection and compression things of the past, in business firms and government organizations (Dunleavy and Margetts 2023) and in science and academic life. The growth of big data plus data science/artificial intelligence (DSAI) approaches has fundamentally altered many organizational operations, creating a distinctive if still emergent post-digital, form of information regime inside advanced organizations – based around full recording and storing of initial data received, using tech that achieves 'lossless compression' (Dunleavy 2022). Digitally recording and storing all data when it is first encountered, or when a dataset is acquired, now obviates much of the need for data compression. It also fosters new forms of post hoc knowledge development, for example via search for new issues hidden in 'verbatim' text or present in images, using machine learning and other forms of algorithmic analysis. The digital revolution hence essentially *decompresses* data. Taken together, digital encoding, transmission and decoding allow the original form of the data to be captured in all its (salient) detail, analysed, repacked, and later unpacked, all in a 'lossless' manner. This information regime change explains why digital changes have had such sweeping consequences across all private and public sector organizations (Dunleavy 2022).

For research, the advent of digital decompression is unambiguously favourable for 'opening up' social science. It ensures the reproducibility of research by radically increasing the amounts of data that can be or are being captured or created, and then stored at extremely low to near-zero cost. Data held in a comprehensively captured form can then be interrogated by later reusers in multiple new ways, some anticipated but others unplanned for. From big data stores or 'lakes', cognitive agents (both human and artificial) can discover information that was not initially foreseen as relevant when the data was retained (Dunleavy 2016). And with faster communication and analysis of more data, new control capabilities are created for tech-based automation in the physical realm, as already with modern inductive search, DSAI analysis and robotics/drones, etc. (Dunleavy and Margetts 2023). For research processes, scientific methods and academic organizations these digital decompression effects have already become super-important – not just in improving

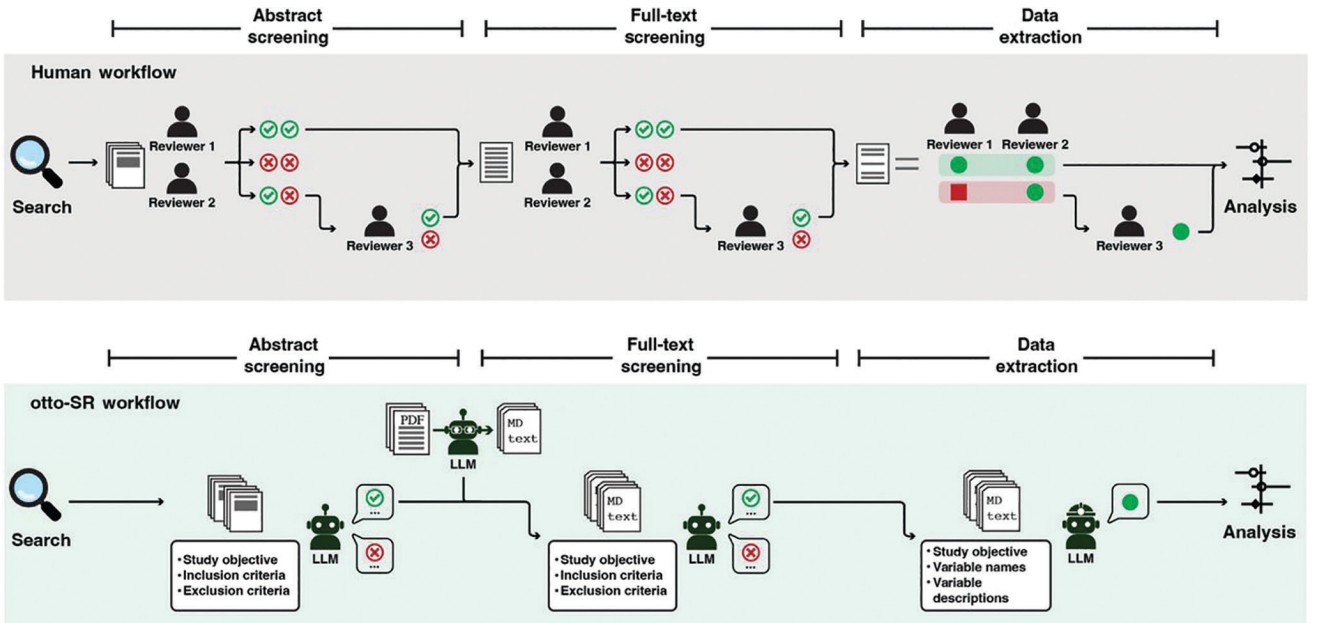
how conventional and AI search processes work in finding and summarizing relevant literatures, or increasing the specialization and mathematization of potential projects. Decompression has also changed foundational research methods, enabling both quantitative methods and qualitative research to use DSAI analyses and new exploratory/inductive approaches. Digital changes also radically cheapen the costs and cut the hassles involved in most data acquisition, cleaning, storage and reaccessing tasks.

Will these trends cease to operate in the next 10–20 years? It seems unlikely, but of course anticipating how digitally changed organizations will develop by projecting futures from current patterns comes with a strong health warning. Digital changes will keep on working through for two decades more at least, but they are likely to take many unanticipated turns in the process, as they have done in the past (Isaacson 2014, 2015). Yet, if anything, IT visionaries' radical expectations of the organizational, political and societal effects of long-anticipated tech changes have been overfulfilled in the last three decades (Brate 2002).

Digital changes are likely to make organizations increasingly flexible and diverse as information-handling mechanisms. The weight of information processing will shift away from still distinctively academic organizational cultures and systems, and increasingly towards partly AI-shaped and probably more generic (and perhaps even partly robotic) tech–social systems, with an ever-deepening use of wider automation and tech-based methods. These directions of digital development will continue to fuel increasing openness in research and more applied knowledge development – as they have in the recent past with open access publishing, expansions in data storage and analysis capabilities, multiple communication and recording tech advances for qualitative information (Jamieson and Lewthwaite 2019), cheaper data depositing, enhanced metadata and intelligent search, and large language models' (LLMs') synthesizing and text-production capabilities. The extent of the changes ahead remains uncertain, but by 2025 one study already claimed significant time and cost savings plus important quality improvements from deploying an AI-driven method (called otto-SR) involving an LLM to undertake two medical Cochrane Reviews (i.e. high-quality systematic reviews). These normally take 18 months to do manually, but the automated approach took only a fraction of that. The infographic in Figure 16.3 shows the key end-to-end differences between the manual and automated processes in outline (Cao et al. 2025, Figure 1).

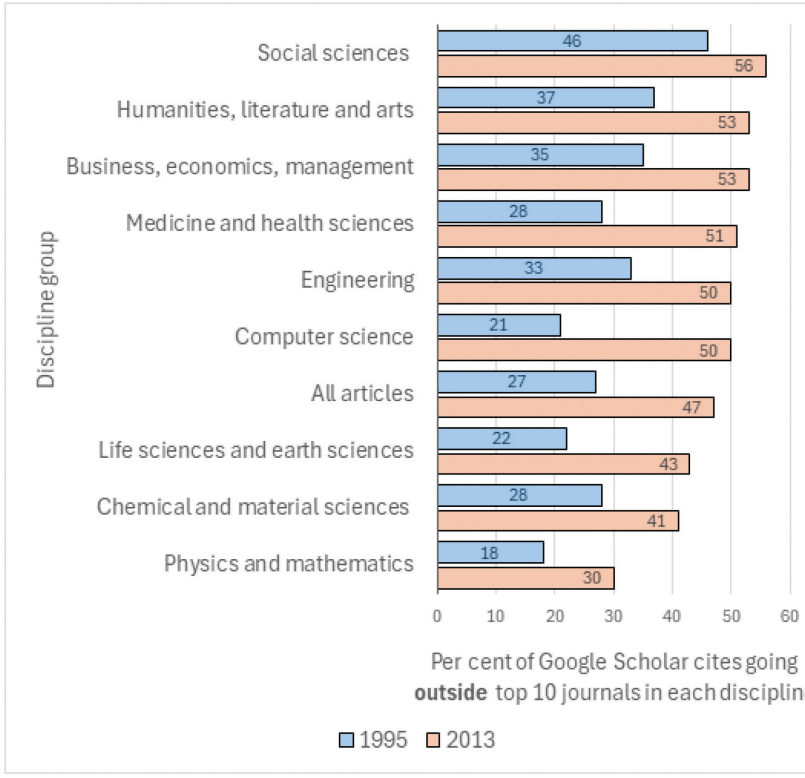
Above all, we know that the digital availability of modern research texts, powerfully boosted by open access at global scale, has already greatly increased the visibility of academic research in non-prestige journals. Figure 16.4 measures the proportion of Google Scholar citations *not* going to the top 10 journals in each discipline during the early digital change period, 1995 to 2013. The social sciences showed a substantial increase in citations to other journals (outside the top 10), up from 45% to 56%. Economics and business increased from a lower 35% to 45% in non-top 10 journals, with a very similar trend

Figure 16.3: Contrasting workflow processes used in manual and AI-driven systematic reviews in a 2025 study



Note: The process for humans is shown with grey background, and the automated otto-SR process using LLMs has a green background. Green ticks show correct inclusion choices and red Xs incorrect decisions.
LLM = large language model.
MD text = markdown text, that is plain text where the mark-up language Markdown has been used to add formatting elements (such as, headings, bulleted lists, URLs) to the text, making it fully readable by LLMs. Using Markdown avoids inserting HTML tags or employing a formal text editor.
Source: Cao et al. (2025).

Figure 16.4: An increasing proportion of citations were to non-elite journal articles from 1995 to 2013, across all disciplines



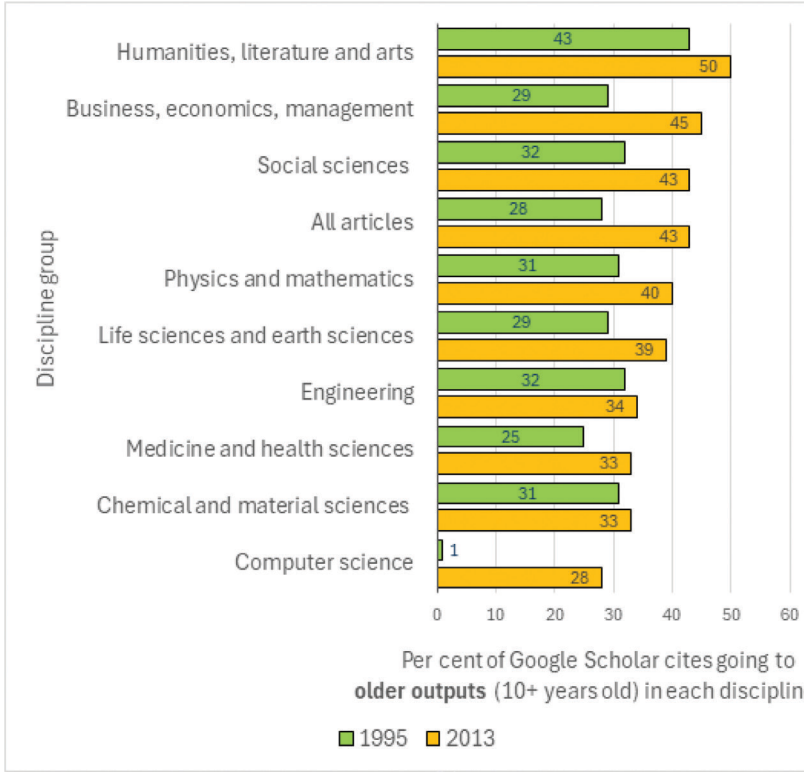
Source: Chart is authors' own, derived from data in Acharya et al. (2014).

in the humanities. Consistent effects over time were also apparent in the STEMM disciplines, marking an across-the-board change that is clearly tech-linked (Acharya et al. 2014).

Perhaps more surprisingly, Figure 16.5 shows that citations to older outputs (those more than five years old) also grew appreciably – from 46% to 56% in the social sciences, and faster in economics and business (from 35% to 54%), with the most rapid change starting in the 2000s (Verstak et al. 2014). Even the humanities changed up from 44% to 56% citations of older work, along with all but two STEMM disciplines (chemistry and engineering). These shifts were additionally made more salient by the changes in how academic research has been evaluated, which now centre more on multiple metrics of citation and reuse in the digital era.

Have these trends continued in the most recent decade? Unfortunately, we do not have more recent data, but digitization has almost certainly kept working in a generally 'democratizing' way – making it progressively easier

Figure 16.5: Older articles (more than five years old) were also cited more from 1995 to 2013 in most disciplines (except in engineering or chemistry)



Notes and key: As in Figure 16.3.
Source: Verstak et al. (2014).

for researchers anywhere in the world to find the most relevant high-quality work for their projects, wherever and whenever they were published – rather than the legacy model of relying solely on outputs in high-prestige journals. Newer DSAI innovations (like LLMs) may possibly result in bigger citation mountains just for top authors in future, perhaps because of AI search effects. But they are equally likely to simply enhance (reasonably expert) users' ability to find and quickly assess relevant materials across a vastly expanded search horizon.

Of course, citations are far from being a 'killer metric' when it comes to assessing academic quality. There is a sound consensus that they should form only one pillar of well-rounded tenure and promotion decisions across universities. Like readership statistics for newspapers or sales numbers for books, we know that counts of 'reads' for articles, books and newspaper or blog items

may or may not mean that readers made any close study of them (or even that users agree with what was said). In a similar way, it is important to maintain a realistic view of just how little an academic citation may mean:

> Although citations are widely used to measure the influence of scientific works, research shows that many citations serve rhetorical functions and reflect little-to-no influence on the citing authors. If highly cited papers disproportionately attract rhetorical citations then their citation counts may reflect rhetorical usefulness more than influence. Alternatively, researchers may perceive highly cited papers to be of higher quality and invest more effort into reading them, leading to disproportionately substantive citations. We test these arguments using data on 17,154 randomly sampled citations collected via surveys from 9,380 corresponding authors in 15 fields.
>
> We find that most citations (54%) had little-to-no influence on the citing authors. However, citations to the most highly cited papers were 2–3 times more likely to denote substantial influence. Experimental and correlational data show a key mechanism: displaying low citation counts lowers perceptions of a paper's quality, and papers with poor perceived quality are read more superficially. The results suggest that higher citation counts lead to more meaningful engagement from readers and, consequently, the most highly cited papers influence the research frontier much more than their raw citation counts imply (Teplitskiy et al. 2022, p. 1).

Compared with other usage indicators, like numbers of reads or altmetrics, though, it is perhaps an advantage that scientists and academics overwhelmingly only cite works they approve of or think are soundly based. Citations of poor or criticized research have turned out in practice to be tiny fractions of the total across the social sciences.

Open science also means that datasets now attract citations in their own right, and there are indications that in STEMM disciplines the numbers of datasets referenced by authors have become substantial. For instance, a study of journal and conference sites in the earth sciences field found '11,000 DOIs associated with [*earth science*] datasets, yielding 17,000 unique journal and conference articles, reports, and book records linked to 3,000 dataset DOIs' (Gerasimov et al. 2024). Additional indicators, such as numbers of accesses, sustained reads and Altmetrics statistics, also show that many more researchers look at datasets than later use or cite them. This 'unsung' lesson-drawing activity can be very important for researchers developing new ideas or approaches, both in the same sub-field and from 'out of field' visitors seeking 'carry across' lessons. Other researchers can learn from full datasets about many disparate issues – for instance, what particular tests or methods were employed and how? What survey questions or behaviour counts worked well? And what areas, regions or countries were covered, for what dates? Open

access databases with good metadata and data stored in easily accessible formats will usually find multiple relevant audiences, especially if the depositors use social media to increase knowledge of their work's availability and the data documentation is also in English or another widely used language.

Applied knowledge and openness

Digital changes also underpin a second key factor supporting the rapid development of OSS, which is the increased importance of maintaining, reconstructing and extending *applied knowledge* of how to manage human-dominated systems (HDS) and human-influenced systems (HIS) in the contemporary world (Dunleavy 2014b; Dunleavy, Bastow and Tinkler 2014; also see Chapter 2). At any given time, Bastow, Dunleavy and Tinkler argue that a 'dynamic knowledge inventory' (DKI) spanning both HDS and HIS areas exists, part of it directly in current use, and part of it potentially accessible but not in use at any given time. Theory-based scientific and academic knowledge is one key source for both used and currently unused DKI knowledge, and applied academic work is the second main source. The far larger third source, though, is the mass of expert or well-founded data, information and know-how generated or used by firms, government and public agencies, civil society organizations, and citizens individually and in groups. This 'ordinary knowledge' has rarely been established via professional social inquiry or certification, but it is vital for modern societies and large areas of public and business policymaking nonetheless (Lindblom and Cohen 1979). There are many different kinds or grades of non-scientific knowledge, including 'contributory' and 'interactional' expertise, plus both disseminated and highly specialized forms of understanding (Collins and Evans 2009). Even more than the tacit knowledge barriers between disciplines (Collins 2010), many specialized and valuable forms of external expertise cannot easily be fully appreciated by academic work alone.

A fundamental aim of OSS (and of OS more widely) is to pervasively improve the quality of non-academic expertise and knowledge, and to constantly combat both information deficits or gaps in it, to correct mistakes from spreading, and to counteract disinformation propagated by vested interests. Advances in STEMM tech areas can be especially improved where innovation is 'democratized' by suppliers or project teams listening very closely to expert consumers using their products in the field, or to service professionals implementing systems and handling problems at first hand (von Hippel 1988, 2005, 2017; von Hippel and von Krogh 2003).

Almost all applied knowledge included in the active parts of the dynamic knowledge inventory is integrative and cross-disciplinary in character. And in the modern world our analysis of HDS and HIS (Chapter 2) stresses that many key processes and phenomena span across both the STEMM and the social sciences fields of study (Dunleavy, Bastow, and Tinkler, 2014; Dunleavy, 2014b). Working to upgrade applied knowledge in OS and OSS ways can help

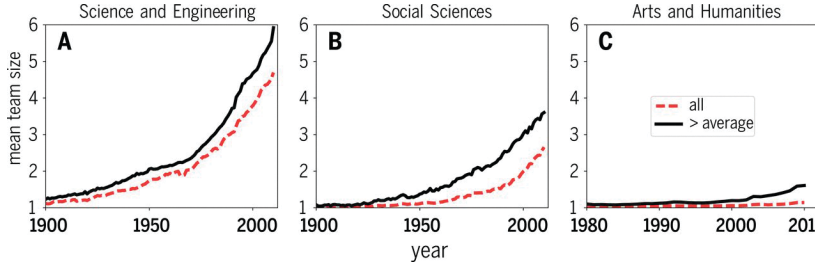
academia to move away from the highly siloed, single-focus, paywalled academic disciplines of the recent past. It can share the modern trend for most major corporate and industrial research to happen partly collaboratively (Ramella and Manzo 2021) and in some open source ways (see Chapter 3). Digital change

> continues to reorder whole disciplines and industries. Entrepreneurs, policy-makers, and researchers have recognized that increased collaboration through these socio-technical systems offers compelling opportunities for business, education, national security, and beyond. It is time for researchers in science to take network collaboration to the next phase and reap the potential intellectual and societal payoffs … Successful scientific collaboratories among genomic researchers, engineering innovations through open-source software, and community-based participation in cultural heritage projects are all early indicators of the transformative nature of collaboration (Shneiderman 2008).

The so-called Science 2.0 or ABC research model is 'applied and basic combined', and it can only be accomplished in teams that are inherently interdisciplinary (Shneiderman 2016, and 2019). Despite some remaining disciplinary hang-ups, academics and scientists in universities have made consistent efforts in the last decade to get better at building project-focused teams. Interdisciplinary teams committed to realizing projects that involve both STEMM and social science knowledge (and even applied humanities via design and future-orienting) have also become increasingly common. In addition, many social scientists now work in 'embedded' ways inside large teams within government agencies, big corporations, or consultancies (Bastow, Dunleavy and Tinkler 2014, Chapter 9). They regularly share know-how with university faculty in their 'home', local or linked departments, and with students (via internships and 'capstone' projects) (Dunleavy and Tinkler, 2021, Chapter 10). Some professionals also carry that knowledge with them if they return to work in universities or become 'professors of practice'.

As in corporations' research, the STEMM disciplines have moved strongly towards using larger teams. To a lesser degree and in a more laggard way, the same trend has characterized social science research. The numbers of authors per research paper grew strongly from 1950 to 2000, especially towards the close of the century (see Figure 16.6). Furthermore, a growing gap opened up between the mean size of teams that produced papers with above average citations in their field (shown by the black lines in Figure 16.6) and the mean size for all papers (shown by the red lines). Large teams clearly produced research that was more cited than average, although there may be a small 'mechanical' effect to offset here, arising from multiple authors each citing their own work. In the humanities only pretty flat minor trends were detectable in this long period.

Figure 16.6: The mean size of STEMM and social science research teams has been growing steadily, and work from larger teams is more cited than average



Notes: The red dashed curves show the mean number of co-authors over all papers in a field. The black curves shows the mean number of authors for just those papers with above average citations per field. 'Black curves are systematically above the dashed red ones, meaning that high-impact work is more likely to be produced by large teams than by small ones'.
These charts are based on Web of Science (WoS) data, and WoS only started to include Arts and Humanities chart in 1980.
Source: Reproduced with permission from Fortunato et al. (2018).

OS and OSS have strongly contributed to the transition to teamwork as the 'new normal' in scientific endeavour, creating a momentum that has eroded some restrictively individualist authorship notions associated with past 'closed science' practices:

> [A]lthough … humanist notions of authorship – including the connotations of reputation, individual creativity, ownership, authority, attribution, responsibility, and originality they came to carry – seem to be an integral part of the scholarly method, they are not 'natural' or 'normal' (Adema 2021).
>
> We believe the hero scientist is a myth and that all science requires teamwork, even as the current incentive structure continues to reward individual achievements almost exclusively. This has remained the case despite the achievements of team science (Gentemann, Erdmann and Kroeger 2022).

Over time we may expect larger teams to facilitate greater 'open exploration' (Vicente-Saeza, Gustafsson and Van den Brande 2020; Cheruvelil, Spence and Soranno 2018) and data sharing in several different ways:

- allowing an effective data management plan to be set up and implemented from the outset, with well-defined team roles (Chapter 7);
- boosting the pre-registration of outputs and methods (Chapter 4);
- improving documenting-as-you-go, vital in larger teams (Chapter 5);

- making data depositing easier and more certain; eroding past patterns of data from completed projects being kept closed; encouraging collective and open access authoring (Chapters 7 and 14);
- helping to reduce incentives for individual CV building with outputs that are little read or used; and
- equalizing out credit-sharing across authors and data contributors.

Conclusions

OSS should be understood as a broad and long-lived cultural change, one where the shared momentum is now in no doubt thanks to an autonomous and strongly rooted change of heart amongst researchers themselves, plus the driving impacts of digital decompression and increased momentum behind more applied knowledge development. However, the transition to OSS will unfold in different stages over the next decade and in different modes across subjects within the discipline group. Over time we can expect many new streams of both academic and digital innovation and adaptations to occur. Some of them may not be easily foreseeable at the present time – as past experience proved (e.g. the rapid rise in the salience of LLM software in universities amongst students and researchers).

In tandem with continued digital changes, we expect open research processes to help foster an increasingly convergent research methods toolkit across the whole social science discipline group. That change should also open up greater linkages and research pooling with the STEMM disciplines, and an increased emphasis upon generating reliable, reproducible and robust research into the human-dominated systems and human-influenced systems that now shape almost all that happens in our world.

Note

* This chapter draws on research and seminars forming part of the CIVICA Research programme, which was funded in 2022–2023 by the European Union's research and innovation programme Horizon 2020, under the chapter 'Science in Society'.

References

Acharya, Anurag; Verstak, Alex; Suzuki, Helder; Henderson, Sean; Iakhiaev, Mikhail et al. (2014) 'Rise of the rest: the growing impact of non-elite journals', *arXiv*, 18 October. https://doi.org/10.48550/arXiv.1410.2217

Adema, Janneka (2021) *From Romantic to Posthumanist Authorship*, Boston: MIT Press, 'Living Books 2', 31 August, online text. https://perma.cc/G6PT-2DAK

Bastow, Simon; Dunleavy, Patrick; and Tinkler, Jane (2014) *The Impact of the Social Sciences: How Academics and their Research Make a Difference*, London: Sage.

Brate, Adam (2002) *Technomanifestos: Visions from the Information Revolutionaries*, New York, NY: Texere.

Cao, Christian; Arora, Rohit; Cento, Paul; Manta, Katherine; Farahani, Elina et al. (2025) 'Automation of systematic reviews with large language models', *medRxiv* preprint. https://doi.org/10.1101/2025.06.13.25329541

Castille, Christopher M.; Kreamer, Liana M.; Albritton, Betsy H.; Banks, George C.; and Rogelberg, Steven G. (2022) 'The open science challenge: adopt one practice that enacts widely shared values', *Journal of Business and Psychology*, vol. 37, pp. 459–467. https://doi.org/10.1007/s10869-022-09806-2

Cheruvelil, Kendra Spence and Soranno, Patricia A. (2018) 'Data-intensive ecological research is catalyzed by open science and team science', *BioScience*, vol. 68, pp. 813–822. https://academic.oup.com/bioscience/article/68/10/813/5088531

Colavizza, Giovanni; Hrynaszkiewicz, Ian; Staden, Isla; Whitaker, Kurstie; and McGillivray, Barbara (2020) 'The citation advantage of linking publications to research data', *PLoS ONE*, vol. 15, no. 4, article e0230416. https://doi.org/10.1371/journal.pone.0230416

Collins, Harry (2010) *Tacit and Explicit Knowledge*, Chicago, IL: Chicago University Press. **$** https://perma.cc/A9DE-9QGR

Collins, Harry and Evans, Robert (2009) *Rethinking Expertise*, Chicago, IL: University of Chicago Press. **$** https://perma.cc/M2TG-V7RP

Dunleavy, Patrick (2014a) 'Are you an academic hermit?', *Writing for Research* blog, 1 March. https://perma.cc/HBZ7-EUSB

Dunleavy, Patrick (2014b) 'The social science of human-dominated and human-influenced systems', Annual Lecture to the Academy of Social Sciences. London: AcSS. https://perma.cc/T38Q-SL2C and https://www.youtube.com/watch?v=nY4ra5QWJeE

Dunleavy, Patrick (2016) ' "Big data" and policy learning', in Stoker, Gerry and Evans, Mark *Methods that Matter: Social Science and Evidence-Based Policymaking*, Oxford: Oxford University Press, pp. 143–168. **$** https://doi.org/10.1332/policypress/9781447329367.003.0009 OA at: https://perma.cc/TAP5-L8R9

Dunleavy, Patrick (2022) 'Information regimes in government bureaucracies and "digital decompression" ', Paper to the UK Political Studies Association Conference, 11 April, University of York. https://perma.cc/JJX3-Y5YK

Dunleavy, Patrick and Margetts, Helen (2023) 'Data science, artificial intelligence and the third wave of digital era governance', *Public Policy and Administration*, 25 September, vol. 40, no. 2, pp. 185–214. https://doi.org/10.1177/09520767231198737

Dunleavy, Patrick and Tinkler, Jane (2021) *Maximizing the Impacts of Academic Research: How to Grow the Recognition, Influence, Practical Application and Public Understanding of Science and Scholarship*, London: Palgrave (now Bloomsbury Press). **$** https://www.bloomsbury.com/uk/maximizing-the-impacts-of-academic-research-9781352010992/

Dunleavy, Patrick; Bastow, Simon; and Tinkler, Jane (2014) 'The contemporary social sciences are now converging strongly with STEM disciplines in the study of "human-dominated systems" and "human-influenced systems"', *LSE Impact blog,* 20 January. https://perma.cc/JJX3-Y5YK

Fecher, Benedikt; Friesike, Sascha; Hebing, Marcel; and Linek, Stephanie (2017) 'A reputation economy: how individual reward considerations trump systemic arguments for open access to data', *Palgrave Communications*, vol. 3, article 17051. https://doi.org/10.1057/palcomms.2017.51

Fortunato, Santo, Bergstrom, Carl T.; Börner, Katy; Evans, James A.; Helbing, Dirk; Milojević, Staša et al. (2018) 'Science of science', *Science*, vol. 359, no. 6379, article eaao0184. https://www.science.org/doi/10.1126/science.aao0185

Gentemann, Chelle; Erdmann, Christopher; and Kroeger, Caitlin (2022) 'Opening up to open science', *Issues in Science and Technology*, vol. XXXVIII, no. 3, pp. 57–59. https://www.jstor.org/stable/27161933 OA at: https://perma.cc/FMC8-B5R8

Gerasimov, Irina; Binita KC; Mehrabian, Armin; Acker, James; and McGuire, Michael P. (2024) 'Comparison of datasets citation coverage in Google Scholar, Web of Science, Scopus, Crossref, and DataCite', *Scientometrics*, vol. 129, pp. 3681–3704. https://doi.org/10.1007/s11192-024-05073-5

Gomes, Dylan G.E.; Pottier, Patrice; Crystal-Ornelas, Robert; Hudgins, Emma J. et al. (2022) 'Why don't we share data and code? Perceived barriers and benefits to public archiving practices', *Proceedings of the Royal Society B*, November, vol. 289, no. 1987, article: 20221113. https://doi.org/10.1098/rspb.2022.1113

Gownaris, Natasha J; Vermeir, Koen; Bittner, Martin-Immanual; Gunawardena, Lasith et al. (2022) 'Barriers to full participation in the open science life cycle among early career researchers', *Data Science Journal*, vol. 21, no. 2, pp. 1–16. https://doi.org/10.5334/dsj-2022-002

Heckman, James J. and Moktan, Sidharth (2020) 'Publishing and promotion in economics: the tyranny of the top five', *Journal of Economic Literature*, vol. 58, no. 2, pp. 419–470. https://doi.org/10.1257/jel.20191574

Heise, Verena; Holman Constance; Lo, Hung; Lyras, Ekaterini M. et al. (2023) 'Ten simple rules for implementing open and reproducible research practices after attending a training course', *PLoS Comput Biol*, vol. 19, no. 1, article e1010750. https://doi.org/10.1371/journal.pcbi.1010750

Jamieson, Lyn and Lewthwaite, Sarah (2019) 'Big Qual – why we should be thinking big about qualitative data for research, teaching and policy', *LSE Impacts* blog. https://perma.cc/GM2P-59EK

Janz, Nicole (2015) 'Is withholding your data simply bad science, or should it fall under scientific misconduct?', *LSE Impact blog*, 3 July. https://perma.cc/Y6VX-UTYC

Isaacson, Walter (2014) *The Innovators: How a Group of Inventors, Hackers, Geniuses and Geeks Created the Digital*, New York: Simon and Schuster. **$** https://www.simonandschuster.com/books/The-Innovators/Walter-Isaacson/9781476708706

Isaacson, Walter (2015) *Steve Jobs*, New York: Simon and Schuster. **$** https://www.simonandschuster.com/books/Steve-Jobs/Walter-Isaacson/9781982176860

Key, Ellen M. (2016) 'How are we doing? data access and replication in political science', *PS : Political Science & Politics*, vol. 49, no. 2, pp. 268–272. https://perma.cc/XJA2-W9RB

Krawczyk, Michal and Reuben, Erensto (2012) '(Un)available upon request: field experiment on researchers' willingness to share supplementary materials', *Accountability in Research*, vol. 9, no. 3, pp. 175–186. **$** https://doi.org/10.1080/08989621.2012.678688

Leonelli, Sabina (2023) *Philosophy of Open Science*, Cambridge: Cambridge University Press. **$** https://doi.org/10.1017/9781009416368

Lindblom, Charles E and Cohen, David K. (1979) *Useable Knowledge: Social Science and Social Problem Solving*, New Haven: Yale University Press. **$** https://yalebooks.yale.edu/book/9780300023367/usable-knowledge/

Mayer-Schönberger, Victor (2016) 'Big Data for cardiology: novel discovery?', *European Heart Journal*, vol. 37, no. 12, 21 March, pp. 996–1001, https://doi.org/10.1093/eurheartj/ehv648

Ramella, Francesco and Manzo, Cecilia (2021) *The Economy of Collaboration: The New Digital Platforms of Production and Consumption*, London: Routledge. OA at: https://perma.cc/V8W-26Q9

Salandra, Rossellal Salter, Ammon; and Walker, James T. (2022) 'Are academics willing to forgo citations to publish in high-status journals? Examining preferences for 4* and 4-rated journal publication among UK business and management academics', *British Journal of Management*, vol. 33, pp. 1254–1270. https://doi.org/10.1111/1467-8551.12510

Shneiderman, Ben (2008) 'Science 2.0', *Science*, vol. 319, pp. 1349–1351. **$** https://www.science.org/doi/10.1126/science.1153539

Shneiderman, Ben (2016) *The New ABCs of Research: Achieving Breakthrough Collaborations*, Oxford: Oxford University Press. **$** https://global.oup.com/academic/product/the-new-abcs-of-research-9780198812173

Shneiderman, Ben (2019) 'Creativity and collaboration: revisiting cybernetic serendipity', *Proceeding of the National Academies of Science*, vol. 116, no. 6, pp. 1837–1843. https://doi.org/10.1073/pnas.1807200116

Shibayama, Sotaro and Lawson, Cornelia (2021) 'The use of rewards in the sharing of research resources', *Research Policy*, vol. 50, no. 7, article 104260. https://doi.org/10.1016/j.respol.2021.104260

Silvia, Paul J. (2007) *How to Write a Lot: A Practical Guide to Productive Academic Writing*, Washington, DC: American Psychological Association. **$** https://psycnet.apa.org/record/2006-23317-000

Tennant, Jonathan P.; Agrawal, Ritwik; Baždarić, Ksenija; Brassard, David et al. (2020) *A Tale of Two 'Opens': Intersections between Free and Open Source Software and Open Scholarship*, Charlottesville, VA: Center for Open Science. https://doi.org//10.31235/osf.io/2kxq8 and https://osf.io/preprints/socarxiv/2kxq8

Teplitskiy, Misha; Duede, Eamon; Menietti, Michael; and Lakhani, Karim R. (2022) 'How status of research papers affects the way they are read and cited', *Research Policy*, vol. 51, no. 4, article 104484. https://doi.org/10.1016/j.respol.2022.104484

Verstak, Alex; Acharya, Anurag; Suzuki, Helder; Henderson, Sean; Iakhiaev, Mikhail; Yu Lin et al. (2014) 'On the shoulders of giants: the growing impact of older articles', Google Scholar, research paper, 4 November. https://doi.org/10.48550/arXiv.1411.0275

Vicente-Saeza, Ruben; Gustafsson, Robin; and Van den Brande, Lieve (2020) 'The dawn of an open exploration era: emergent principles and practices of open science and innovation of university research teams in a digital world', *Technological Forecasting & Social Change*, July, no. 156, article: 120037. https://www.sciencedirect.com/science/article/pii/S0040162518316378

von Hippel, Eric (1988), *The Sources of Innovation,* New York: Oxford University Press. https://library.oapen.org/handle/20.500.12657/46053

von Hippel, Eric. (2005), *Democratizing Innovation*, Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/2333.001.0001

von Hippel, Eric (2017), *Free Innovation*, Cambridge, MA: MIT Press. https://perma.cc/JL8Y-FH3H

von Hippel, E and von Krogh, G. (2003), 'Open source software and the "private-collective" model: issues for organization science', *Organization Science*, vol. 14, no. 2, pp. 209–223. OA at https://dx.doi.org/10.2139/ssrn.1410789

von Hippel, Eric; Ogawa, S; and de Jong, J.P.J. (2011) 'The age of the consumer innovator', *Sloan Management Review*, vol. 53, no. 1, pp. 27–35. OA at https://perma.cc/6KHV-P3FE

Index

Note: Page numbers in italics denote figures.

80/20 rule 140

A

ABC research model 380
academic conservatism 15, 187–188
academic diffidence 354, 367
academic versus co-production authenticity 313
accessible data *see* data/datasets
accidental data losses or damage 161–162
acknowledgement of data reuse 154
action research 302
  see also participatory action research (PAR)
active participation of the participants 134, 300, 301, 309
  see also citizen social science (CSS)
adversity, in interviews 204
AI 60, 245, 274, 374, 377
  in data cleaning 141
  search tools 132
  see also data science/artificial intelligence (DSAI) approaches
All European Academies (ALLEA) 174
Amazon's 'mechanical Turk' system research scandal 207
analytic accounts 261
annex positioning 184–187
  see also research methods annexes (RMAs)
annual reports 256, 257, 269
anonymization 158, 160, 163, 192, 199, 201, 202, 211, 212, 213, 214, 219, 227, 230, 231, 241, 311, 313, 314
Anscombe's Quartet 139
application programming interfaces (APIs)
  data cleaning with 142
  full text databases 255
  of social media companies 263
applied knowledge 374, 379–382
apt illustration cases 192, 225, 231, 232, 233, 242
archives and archival data 285, 294–295
  access restrictions 287
  cultural barriers around openness 292–294
  digital archives 286
  digitalization of 291
  resources and research 285–288
  see also historical research
argumentative structures 261–262
article formats 349–350
article processing charges (APCs) 24, 49, 344, 349, 350, 358
Arxiiv 355
'as submitted initially' (ASI) 349
Association of Social Anthropologists (ASA), data management guidelines 158
Astropy 57
attention points 177, 323, 325–326
  audience-oriented 326
  data description 332
  description of numbers 330
  exhibits, strengthening of 331–332
  legacy publishing practices, avoiding 332
  limits on 324
  public engagement maximization 333
  self-explanatory 326
  simplicity 327
  single sequence of 328
  tables and charts 329
Australia 234
authenticity 201, *259*, 313
autonomous participation 314
auto-recorded interviews 212, 218
axis labels 330

B

backup systems 108, 117, 126, 161–162
bad science 62–63
Bankspeak 272
barriers and drivers, for open social science 364–372
Bayesian models 80–81, 92
big book thesis 178, 182, 184, 186, 188
big data 41, 96, 134, 138, 325, 332, 335, 373
Binder 126

black box elements 143, 264
black hole imaging *58*, 59
blogposts 333
book processing charges (BPCs) 344, 351, 358
bookified version of the PhD text 188
Boolean search codes 266
boosterism 205
Boston Review 355
'Build, Blur, Corrode' criteria 183

C

career advancement/development 47, 364, 369, *370*
case studies 225, *228*, 249–250
  apt illustration cases 231
  business 199
  case selection 227, 231
  casing 226, 229
  de-emphasizing data 331
  deep studies, openness of 229
  diagnostic cases 229–231
  external research 242–243
  fuzzy set (QCA) cases 239–241
  interviews 191
  issues 192
  limiting cases 233
  matching case 232–233
  mixed-methods studies 236–239
  pre-registration 234
  purpose of 226
  with quantitative aspects 234–242
  rationales 226–228
  systematic documentation analyses 230
  systematic review 244
  systematic review and meta-analysis 242–247
  triangulation of results 228
causal influence interviews 197
causal mechanisms 226
CC (Creative Commons) licenses 169, 170, 347, *348*
  CC-BY 169, 347
  CC-BY-NC 169, 347
  CC-BY-ND 347
  CC-BY-SA 347
censorship *258*
Center for Open Science (COS) 24
CERN project 57
ceteris paribus clause 40, 56, 59, 82
ChatGPT 132, 141, 264
Christakis, Nicholas A. 15
Christensen, Garret 18, 23
chronologies 266
citizen humanities 301, 314
citizen social science (CSS) 32, 170, 299, 314
  academic versus co-production authenticity 313–314
  approaches 300–304
  early idea around 300
  insight-only PAR focus 302
  managing data release versus participant confidentiality 313
  methods innovation 309
  older and newer forms *301*
  openness 309
  ownership of research 306–307
  as participatory research 301–302
  Prainsack framework 304
  pre-planning 310–311
  recording unrepeatable events 311–312
  research boundaries and methods consistency 312
  research impact 307–309
  research misalignment 305
  researcher and researched 302
  safeguard conditions 313
  sensitive data in 314
citizen-involving scholarship 294
Claude 141
closed science xvi, 25, 39, 40–43, 61, 182, 372, 381
closed/restricted data 169, 270
co-constructed interview approaches 198
codebooks 138, 162, 165, 330
collaborative methods, deepening 303
collective resources 65
Collins, Randall 16
commendable behaviours, masking 276
commercial algorithms 263
commercial confidentiality 260
commercialization of science and technology 43
common identifiers, in datasets merging 144
common sense 161, 167, 247, 261
Commons Open Repository Exchange (CORE) *172*
complex workflows *366*
comprehensiveness 208, 245, *259*
computational analysis package 274
computational and mathematical sciences, OS in 50
confirmatory causal analysis 94

confirmatory data analysis (CDA) 79, 86–89, 99
confirmatory research 227
contemporaneous notes 199, 205, 214, 248
contemporary (modern period) records 255
  special access 255
contemporary history 285, 286, 290, 293, 300
contextualized analysis 130
contributory expertise 379
conventional wisdoms 45, 261
co-production 32–33, 300, 302
  authenticity 313
  complexity 310
  policy recommendations 305
  research boundaries and methods consistency 312
copyright issues 159
copyright license 169–171
corrupted freemium model 353
costing and resourcing data preparation 158
counting
  across documents and web audits 277
  within full texts 272–274
Creative Commons license *see* CC (Creative Commons) license
critique, risk of *367*
crowdfunding 353
crowdsourcing 300
CSS *see* citizen social science (CSS)
CSV (comma-separated values) format 152
cultural barriers, archival research openness
  closed sources, crowding out 293
  fear of inexpert analysis and disinformation 293–294
  historical evidence management 292–293
cultural biases *259*
curated datasets 279
curating and depositing research archives 269–271
cyber security 64, 160, 161, 166

D

Data Access Committees 151
Data and Code Disclosure policy 180
data and datasets 373
  accessibility 29, 131, 132, 151
  alphabetization 328
  anonymized versions 160
  backups 115, 161–162
  as barrier *366*
  citations 378
  coding 272
  collection of 90
  complexity in merging 144
  confidentiality 143, 214
  copyright license 169–171
  customary orderings 328
  description of 332
  evaluation framework 134–137
  exclusions 109
  exploration 133, 137–140
  findable 150–151
  interoperable 151–153
  labelling 165
  large 96
  large scale data collection 164
  merging 129, 143–146
  new 157
  noise in 87
  pre-existing 156
  recording and sorting 373
  reporting in open way 328
  reproducibility 29
  reverse engineering approach 133
  sharing *368*, 381
  significant 267
data cleaning 129, 140–141
  automation 140
  regular expressions (RegEx) 141–142
  using AI 141
  with Application Programming Interfaces (APIs) 143
data display problems 324
data license agreements 162
data management plan (DMP) 149
  costing and resourcing data preparation 158–159
  data identification and assessment 164–167
  data storage and security practices 160–161
  data sustainability 161
  depositing completed data and documentation 168–171
  documentation 186
  documenting and describing data 165–166
  ethical obligations 158
  FAIR data principles *see* FAIR principles
  new data 157

ownership of dataset 159–160
pre-existing data 156
repository selection and access level 168–169
responsibilities within the research team 166
scope and overview 154–160
secondary data access 162–163
securing personal and sensitive data 163–164
standard 155–156
templates 154
writing within a restrictive word count 167
Data Protection Impact Assessment (DPIA) 164
data repositories 150, 151, 171–173
cross-national 188
free-to-use 168–169
selection and access level 168–169
data reuse 30, 60, 129, 153–154, 162–163, *367*
data sources 133
decision framework 129–131
identifying feasible data 131–133
limitations 136
meanings 135–136
provenance 136–137
relevancy 135
reputation 136–137
data sampling 138
data science/artificial intelligence (DSAI) approaches 60, 64, 263, 264, 274, 276, 293, 373, 377
data sharing agreements 166
data visualization 139, 333–338
charts 336
creating new visual 334–335
data geometry 335
data presentation in colour 335
feedback 338
heading and captions 338
infographic 338
Midway's principles *334*
showing data and model in charts 336
software choice 335
visual indicators of uncertainty 335
DataCite 174
data-dredging *46*, 62
data-protection legislation 149
data-sharing work 155
date-stamped documents 85, 211, 266
defensive writing 183
deliberative democracy (DD) 302, 303, 311
democratization
of history 287
of innovation 363, 379
democratization/transparency
globalizing access to research 68
interdisciplinary understanding 66–67
public belief and confidence 68–69
public engagement with research 67–68
scientific/academic knowledge and ordinary knowledge, building link between 67
descriptive accounts 261
device security 160–161
diagnostic cases 229–231, 232
'diamond' open access articles 350, 358
'diamond' open access books 351, 358
digital archives 286, 291
digital changes 373, 374, 380, 382
impacts of 374
in organizations 374
and OSS development 379
as pushing factor of OS agenda 49
and STEMM science 52
Digital Curation Centre 174
digital decompression 66, 363, 372–382
digital development of government IT systems (case study) 234–236
digital history 193, 273, 285, 286, 287–288, 291, 294
digital humanists 295
digital humanities and media studies 273
digital meeting software 196, 208, 214, 218, 312
digital publishing 49, 343
digital recording 373
digital revolution 49
digital timestamp 90
direct subsidy for the whole BPC 353
disciplinary over-siloing on historic lines 15
disinformation 50, 65, 257, 293–294, 347, 379
DMP *see* data management plan (DMP)
document/documentation analysis 230, 255
advantages and limits of 256–257
contemporary records 255
counting across documents and web audits 277
counting within full texts 272–274

full text API (Application Programming Interfaces) 255
gold standard 264
immersive reading approaches *see* immersive reading approaches
internal documents 256
openness 259–264
postponing 268
quantitative metrics 271–279
documenting data 165–166
documenting research 107–108, 132
backup systems 115
developing approaches for 117–118
digitization system 119
file naming and organizing 112–115
logging 122
open notebooks 122–123
practices 110–112
problems with 108–110
reference management software 123–125
scratch pads and short-run records 118–119
transaction costs 108
documents
biases, controlling strategies 257
indexing 266
see also document/documentation analysis; documenting research
DOIs (digital object identifiers) 150, 153, 168, 324
downscaling, datasets merging 146
DSAI *see* data science/artificial intelligence (DSAI) approaches
duplicating falsified hypotheses 57
duty of care 299
Duverger's Law 82
dynamic knowledge inventory 271, 309, 363, 379

E

ecology and evolutionary biology, OS in 61
EDA *see* exploratory data analysis (EDA)
efficiency
funding and resources usage 65
in-use knowledge versus knowledge not in use 65–66
mistakes debugging speed 64–65
efforts of OSS 1
ehtim library 59
electronic notebooks 122, 266, 270
elite interviews 198, 199, 208, 212
elite universities 3, 364, 369
embargoed 'green' open access (OA) articles 349
embargoed 'green' open access (OA) books 350
embargoed government documents 260
empirical intimacy 226
ESRC Research Data Policy 167
ethics 140, 146, 158, 305
ethnographic data 158
European Commission 305
European Economic Area (EEA) 163
European Research Council (ERC) 305
Event Horizon Telescope team 57, 59
evidence synthesizing methods 244
evolving practice 188
examiners 187–188
Excel spreadsheet
automation 140
coding errors 109–110, 137
size limits 164
exhibits 177, 178, 180, 323, 329
versus full reporting of data 325
strengthening of 331–332
experimental and laboratory-only science, OS in 56–55
experimental methods, researchers using 18
exploratory data analysis (EDA) 93–98
and CDA, relationship between 93–94, 97
central position of 94
inductive approaches to data analysis 97–98
large datasets 96
and pre-analysis plan 97
preliminary work 95
exploratory descriptive interviews 197
exploratory research 226–227
generating evidence but not data 98–99
explorer approach 133
externally given codes 273
extractive research practices 305

F

FAIR principles 17, 29, 149–154, 167, 307
accessible data 151
findable data 150–151
interoperable data 151–153
in mathematics 51
reusable data 153–154

false positives 23, 87, 146
false starts, for open social science 17
  associating with integrity issues and surveillance 20–22
  methodological sectarianism 22–25
falsification *259*, 263
falsificationist optimism 48
fears of research ideas being 'scooped' *365*
field sciences, OS in 59–61
fieldwork research 225, 247–249, 250, 306
  interviews 191
Figshare 173
files
  backup systems 117
  naming codes 114
  organization, archiving as a search engine 114
filtering bias *258*
filtering data 257
findable data 150–151
Finnish Social Science Data Archive (FSD) 168
'flipped' open access books 352–353
focused supplementation 188
footnotes 290
FOSTER project 56
framing approach 3–5
France
  STEMM disciplines in 42
free and open source movement 57
Free and Open Source Software (FOSS) community 52–55
free platforms *54*
free software movement 53, *54*
'freemium' open access 353
Freese, Jeremy 23
full reporting of data versus exhibits 325
full text API (Application Programming Interfaces) 255
funding and resources, efficient usage 65
fuzzy matching 145–146
fuzzy set (QCA) set *see* qualitative comparative analysis (QCA)

G

GDPR (General Data Protection Regulation), EU 149, 160, 163
generalist history accounts 289
generalizability of results 28, 34–32, 279
geocoding 143
Git 126
globalizing access to research 68
go-fair.org 150
'gold' open access articles 349
'gold' open access books 351
Google Correlate 179
Google Scholar 132, 369
government funded research 211
grant-funding 48–49
graphics presentation advice books 331
grey literature 245
Grok 132, 141
grounded analysis 98, 262
grounded theory 227

H

hand coding interviews 216
handwritten notes/journals 270
harking 21, *46*, 62, 89
Harvard Dataverse 168, 172
Herndon, Thomas 109
historical research 288
  costs 292
  digitization of sources 291
  new research areas 290
  primary subject matter 290
  reproducibility 291
  see also archives and archival data
history work 300
Howes, Anton 289
human artefacts 12
human readable data 137
human-dominated systems (HDS) 13–14, *15*, 29, 33, 67, 379
human-influenced systems (HIS) 14, *15*, 29, 33, 379
Hume, David 261
hybrid journals 349
hype *258*
hypotheses 81
  significant level 87
  testing 87
hypothetico-deductive (HD) model 45, 81

I

ideological bias *258*
immediate 'green' open access articles 349
immediate 'green' open access books 351
immersive reading approaches 264–265
  achieving depth understanding 266–268
  documentation archive 269–271
  scoping and set-up 265–266
implicit knowledge *258*

in vivo text analysis software 216
inappropriate use *367*
inclusion/exclusion criteria 312
independent stance about text sources 260
in-depth qualitative interviews 216
indexing documents 266, 268
indigenous research 306
inductive models 81
infographics 333, 338
informed consent 313
insight-only PAR focus 302
integrated OS themes 61
  democratization/transparency 62, 66
  efficiency 62, 64
  integrity theme 61, 62–64
integrated systems 13
  human dominated systems (HDS) 13
  human influenced systems (HIS) 14
integrity theme
  bad science, emergent effect 62–63
  individual research, strengthening integrity in 62
  reproducibility, improving 63
  surveillance 64
intellectual property (IP) right 116, 140, 146, 162, 169, 304
  CSS 313
  policy 157
interactional expertise 379
inter-connected discipline 9, 12
interdisciplinary understanding 66–67
interlocutors 163, 198, 300, 301, 305, 308
intermediate cases 225
intermediate comparative cases 232
internal documents 255, 256
interoperable data 151–153
interviews 195–196, 218–219, 256, 272, 273
  advantages 203–204
  Amazon's 'mechanical Turk' system scandal 207
  analysing and quoting 205
  auto-recorded 212, 218
  biased and fraudulent 206–208, 213
  citizen social science (CSS) 311
  coding methods 215–216
  coding short open text comments 215
  common risks 204–205
  details of 205
  dialogues and conversations 196–199
  in-depth qualitative interviews 216
  interviewee confidentiality 230
  interviewee status 198
  LaCour-Green gay marriage study scandal, 2014 206–207
  limitations 204–206
  non-attributable 199–201, *209*
  off-the-record 199, *200*
  on-the-record 201–202, 208, *209*
  paying for 198
  pre-registration 211–212, 218
  purpose of 197
  quantitative metrics 215–218
  quoting 213
  recording modes *209*
  remote but face-to-face interviews 196
  reproducibility of interviews 213
  structure of 197–198
  talk-based data 202
  types and statuses of 199–202
  uncertainties around 206–208
interviews, openness 208–215
  non-attributable interviews 212–214
  off-the-record interviews 211–212
  on-the-record interviews 214–215
  recorded interviews 208
inter-war social sciences 227
in-use knowledge versus knowledge not in use 65–66
Ionnidis, John 47
iterative improvement 188

J

Japan 235
joint copyright 170
journals 285
  abbreviation of names 324
  high status 21
  replication archives for articles 178–180
Jupyter 126
JUSTINT programme 305

K

kick-starter 353
knowledge exchange 305
knowledge position 299

L

LA Review of Books 355
lab notebooks 111, 122–123, 270
LaCour-Green gay marriage study scandal, 2014 206–207
legacy practices 324, 332, 369, 372

legal data protection 160
library constraints 268
library-like file organization 114
life stories 204, 266
life worlds 204
lifeworld interviews 197
limiting cases 226, *232*, 233
literate programming 122
LLMs (large language models) 132, 141, 245, 264
local knowledge 305
logging, research record 120
London Review of Books 355
LSE Review of Books 355

M

machine learning (ML) 146, 373
mandatory reproducibility requirements 180
Mantra (online course) 174
manual coding 272
mass atrocity violence 305
mass education 45
matching variables, datasets merging 145–146
matching/matched cases 226, 232–233
matrix patterns 262
'me too' research 48
mediated access 151
medium N studies, using QCA methods 241
metadata 151, 153, 171, 266, 268
methodological pragmatism 234
methodological sectarianism 22–25
methodology chapter 182–184
methods adaptation 184–187
methods wars 22–25
metricization feeds over-competition *47*
micro-indicators 277
Miguel, Edward 23
misinformation surges 308
mistakes debugging 64–65
mistakes in record-keeping *259*
mixed-methods studies 192, 225, 234, 236–239
  case selection 237–238
  openness 238–239
  process-tracing 238
  quantitative analyses 236–237
model cases 227, 242
multi-national and multi-team research 29
multiple publication tries *46*
Murphy, Kevin R. 333

N

narrative analysis 262
national archives 259, 286
natural language processing 273, 274
natural/physical sciences and social sciences, false polarities between *14*
nested analysis model 236, *237*
netizens 314
new data 157
NextCloud 126
non-attributable interviews 199–201, 230
  openness 212–214
non-digital document scanning 273
non-digital public sources, access to 260
non-disclosure agreements (NDAs) 166, 199
non-prestige journals, visibility of 374
non-reactive data 134, 136
non-reproducibility of studies 21, *47*
non-scientific knowledge 379
Notes software 118, 268
Null Hypotheses Significance Testing (NHST) 86, 88
null hypothesis 81, 90
numerical scoring systems 241
Numpy library 55, 59

O

off-the-record interviews 199, *200*, 230
  openness 211
omitted variables bias *46*
on-the-record interviews *200*, 201–202, 208
  openness 214–215
open access publishing 25, 28–29, 169, 287, 324, 334, 343, 358
  article formats 349–350
  book formats 350–351
  landscape 344
  license 347
  maximizing the audience for research 344–347
  newer formats 352–353
open archaeology 287
open exploration 381
open history 287, 289, 294
open notebooks 122–123
open science advocacy 299
Open Science Foundation website 85, 355
open science movement 2, 9
open source publishing *54*
  embracing 3
  integrated rationale for 69

open source technology 52–56
OpenAlex 132
openness of academic work, enhancing 1–2
OpenStreetMap 143
optical character recognition (OCR) 273
optional stopping 90
ordinary knowledge 1, 7, 29, 62, 63, 67, 289, 294, 306, 308, 309, 379
originality of research 48–49
otto-SR 374
outreach, diffidence about *367*
over-publication of positive results *46*
over-report data 329
OwnCloud 126
ownership
  of data 159–160, 170
  of research 306–307

P

page limits 324
paper documents 279
paperize 186
papers model PhD 177
participatory action research (PAR) 288, 302
participatory research 202, 301–302, 306, 307, 311
  see also citizen social science (CSS)
paywall access 10, 45–47, 293, 306, 344
  articles 349
  books 350
  proprietary software *54*
PC 93, 116, 118, 123, 270, 332
PDFs 152
peer review 10, 39, 42, 43, 44, 243, 290
  closed 182
  of PAPs 92
  passing *46*, 47, 118
  of *Science* 207
  STEMM sciences 41, 42
periodized narratives 266
personal data, securing 163–164
p-hacking 21, *46*, 62, 88, 89, 91
PhD examiners 182, 183, 187
physical science endeavours
  cumulation 40
  focus 40
  handling puzzles 40–41
  quantitative methods 41
  regularities 40
  scrutiny 41
  tested models 40
pluralistic models of scientific enquiry 45
podcasts 208, 214, 215, 333, 356
political ideological bias *258*
POPIA (the Protection of Personal Information Act), South Africa 149, 160
positive learning gains 280
positive results bias 21, 63
post-codings 216, 218, 273
post-conflict situations, researching in 307
postwar paywall science period 40–43
potentially falsifiable hypotheses 80
pre-analysis plans (PAPs) 85, 86, 89, 90, 91, 92, 95, *96*, 97
pre-coding 273
pre-digital organization 373
pre-existing data 153, 156
  see also data reuse
pre-planning, citizen social science 310–311
pre-print stage 355
pre-registration 60, 79, 99
  case studies 234
  and CDA research design 90–91
  exploratory research 98–99
  historical research 290, 292
  immersive reading 266
  interviews 211–212, 218
  limits of 92–93
  non-attributable interviews 213
  on-the-record interviews 214
  quantitative social research 85
  rationale for 84–86
  role in CDA and EDA research *96*
  and scientific method 80–84
prestige of social sciences 16–17
primary subject matter 290
process-tracing 227, 238
pro-ethics bias *258*
professional prestige 33
proprietary IT systems 180
proprietary social media data 29
proto-documentation stage 267
provenance xvii, 1, 2, 97, 108, 136, 177, 182, 219, 238, 242, 249, *259*
proxy variables 130
psychology, replication crisis 47–48
public belief and confidence 68–69
public engagement with research 67–68, 333
  see also citizen social science
public policy priorities 305
public trust in scientists 44
publication bias 85, 88

publication glut effects *47*
publication phases 357
'publish or perish' system 42, 45
p-values 88
Python 59, 274

Q

qualitative comparative analysis (QCA) 239–241, 274
  case selection 239
  combining and comparing cases 240
  data reuse 242
  five-point scale for outcomes 239
  medium N studies using 241
  openness 241
  Ragin's analysis 240
  set-defining questions 240
  software packages 241
qualitative research 18, 23
  data depositing 171
  interoperable data 152–153
  logging key documentation in 122–123
  research interviews 195
quantitative research 18, 41
  interoperable data 151–152
  logging key documentation in 120–121
  privileging 22–25

R

Ragin's analysis 240
randomized control trials (RCTs) 85, 242, 309
reactive data 133
readers guidance 180
Readme file 180
reductionist approach 56
reference management software 123–125
registered reports 63, 92
regular expressions (RegEx) 141–142
Reinhart and Rogoff case 108–110
remote interviews 196, 212, 218, 312
replicable research 28, 31, 88, 250, 279
replication archives 177, *181*, 324, 364
  as a condition 30
  for journal articles 178–180
replication code 140
replication crisis 21, 47–48, 63, 86, 92, 182
replication study 260
representative cases 227
representativeness *259*
reproducibility 259, 263, 324, 364
  of interviews 195
  maximization of 26–28
  of interviews 199, 206, 213
  qualitative research 30–31
  web audits 279
reproducibility crisis 50, 56, 227
  historical research 288
reputation 279
research collaboration 131
Research Data Management and Sharing (online course) 174
research dissemination 29, 354–357, 369
  blog posting 357
  book reviews 355–356
  chapter/article with one-click link 356
  conference or seminar presentation posting 357
  endorsements from seniors 355
  long-form output 357
  podcast/videocast 356
  pre-print server posting 355
  short-form output 356
  social media messages 355
research impact, CSS 307–309
research journal 266
research memory 107
research methods annexes (RMAs) 178, 182–188, 230, 324, 329
  academic conservatism 187–188
  methodology chapter 182–184
  positioning 184–187
  time needed for 186
  transparency 185
research misalignment, and CSS 305–306
research networks 131
research skills 130
research subjects 299, 308
research team
  responsibilities within 166
  size of 112
research workflow metaphor 108
researchability 306
researcher insecurity and worries *365*
ReShare 168
resource waste 306
re-typing verbatim parts of salient documents 270
reverse engineering approach 132
reviewers 187–188
revisable/ disposable summary narrative 267
rich text format (.rtf) files 152
Rinehart, Richard 288

robustness
of knowledge 250
of result 28
of social science research 31–32

S

safeguarded data 169
SafePod 162
sample size 90
sanitization *258*
Science 2.0 380
scientific method
loss of credibility 56
and pre-registration 80–84
scientistic evidentialism 22
scite.ai 132
scratch pads 118–119
search engines 114, 132, 245, 264
secondary data, access to 162–163
see also data and datasets; data reuse; reproducibility
seed numbers 180
selectiveness *259*
self-archived materials 269–271
self-serving bias *258*
semi-structured interviews 197–198, 206
sensitive data 117, 151, 158, 162, 163–164, 241, 314
sensitive issues *366*
set-defining questions 240
sharing rights *366*
short-form output 356
short-run records 118–119
'show and tell' materials 2, 192, 268, 270, 279, 291
significance levels 87, 88
Simon, Herbert 13
single-discipline siloes 306
single-theory analyses 263
skill development 130
social media 29, 49, 68, 131, 276, 333, 357, 358
campaigns 308
impact on OS 49
messages 96, 308, 355
social sciences
discipline group 10–17
recognition of 9
software choice 335
South Korea 235
special categories data 163
special pleading cases 234
specialist interviews 198
standard deviations 87, 335
standard errors 330, 335
statistical software 91
status 198, *259*
STEMM disciplines 9, 380
citizen science 301
components, transferring to OSS 17–20
controllable systems 56
credibility issues 56
cumulation 40, 41
dataset citations 378
external case studies 244
field research 247–248
focus of 40
handling puzzles 40–41
importance of documenting research in 111
interdisciplinary efforts 14
pre-registration and PAPs 85
privileged status of 43
probabilistic connections and causal connections 56
push for open science 43–50
quantitative methods 41
regularities 40
scandals 43
tested models 40
translation studies, need for 44–45
varieties of OS in 50–61
Stockholm syndrome effects 260
storyboarding 266
string manipulation 141
structured interviews 197
subscription models 353
subsidies 353
supervisors 187
supplemental documentation 177
supplementary long-form outputs 357
supplementary materials 270
supporting documentation 166
synthesis story 246
systematic documentation analyses *see* document/documentation analysis
systematic review, of cases 242–247

T

tables and charts 329
tacit knowledge 79, 92, 111, 186, 379
talk-based data 202
tampering data 160
technical complexity *259*
tech-social systems 374

Test for Excessive Success 91
text analysis software 216, 274
text image analysis 273
thematic analyses, of interviews 216
thematic analysis 262
three-point transparency approach 185
time pressures *365*
tokenization 274
transaction costs of documentation 108
transient storage *368*
translational studies 44–45, 60
transparency 2, 153–154, 324
  affecting factors 263
  RMAs 185
  see also open access publishing
Transparent and Reproducible Social Science Research: How to Do Open Science 23
triangulation 228
Trusted Research Environment (TRE) 162
Tukey, John 93
Turing Way project *27*, 28, 31, 52, 174, 279
t-value 90
two-step research process 95
type I errors 87, 89, 90, 91, 240
type II errors 87

U

UK 144, 163, 285, 288, 347
  COVID-19 pandemic 137
  Data Service 151, 152, 159
  government administrative forms 277
  research funding 49
  statistical geographies 145
  STEMM disciplines in 42, 111
  study of hospitals in 217
underpinning publications 177
un-editable PDFs 279
unity of the discipline group 12
unrepeatable events, recording of 311–312
upscaling, in merging datasets 144
USA 9, 138, 344
  archives 287
  government funding 211
  STEMM disciplines in 42, 50

V

value-added from OSS *368*
Vance, J. D. 50
variable bias 81
variable-focused methodologies 240
verbatim documents 256–257
videocasts 208, 214, 215, 333, 356
'vox pop' extract quotes 197

W

weak incentives *366*
web audits 277, 279
Web of Science 132
web scraping 279
Wikipedia 57

Z

Zenodo xii, 168, 173
Zotero 123, 125

DOING **OPEN SOCIAL SCIENCE**

"This book will undoubtedly be transformative amidst the move towards open research across disciplines. It looks not only at how open social science has been done but also what its many futures may become as they unfurl. It is a landmark text."
Dr. Matthew Hanchard, University of Sheffield, UK

Open science is a set of principles and practices that aims to make research from all fields accessible to everyone for the benefit of researchers and society as a whole. *Doing Open Social Science: A Guide for Researchers* is the first comprehensive book setting out the principles and practices of open research, tailored specifically for those in social science disciplines, at every career stage, offering practical advice on how to make research more transparent, trustworthy and reusable.

Divided into four parts, Part I examines the overall theoretical and methodological aspects of changing from long-established 'closed' approaches to research, setting out an ambitious and inclusive open social science agenda. Part II addresses the practical issues of improving the reproducibility of research through open approaches, including chapters on the principles and tools of documenting research as you go and on open data practices. Part III focuses in on open practices within qualitative social science disciplines. Chapters examine interview-based research, case studies and fieldwork, systematic documentation analysis, archival data and the role of openness in citizen (social) science. Part IV addresses shifting research cultures, including chapters on strategies for presenting research clearly and accessibly to maximise reach and impact, and on open access publishing. The book ends with a discussion of the future of open social science, arguing that openness as a wider cultural change can renew the social sciences and the core foundations for academic progress in more dynamic and sustainable ways than current approaches allow.

This is an essential guide for anyone working in the social sciences who wants to engage more effectively with open research, from doctoral candidates and early career researchers to experienced academics and practitioners.

ISBN 978-1-911712-59-6
90000
9 781911 712596
LSE Press
press.lse.ac.uk