Research Data Management in the Canadian Context

A Guide for Practitioners and Learners

Edited by: Kristi Thompson, Elizabeth Hill and Emily Carlisle-Johnston (English)
Danielle Dennie and Émilie Fortin (French)

Research Data Management in the Canadian Context

RESEARCH DATA MANAGEMENT IN THE CANADIAN CONTEXT

A Guide for Practitioners and Learners

EDITED BY KRISTI THOMPSON; ELIZABETH HILL; EMILY CARLISLE-JOHNSTON; DANIELLE DENNIE; AND ÉMILIE FORTIN

JENNIFER ABEL; EUGENE BARSKY; MARTIN CHANDLER; LUCIA COSTANZO; DYLANNE DEARBORN; ALISON FARRELL; ÉMILIE FORTIN; JANE FRY; LOUISE GILLIS; MEGHAN GOODCHILD; KARA HANDREN; ELIZABETH HILL; GRANT HURLEY; SHAHIRA KHAIR; DANI KWAN-LAFOND; OLIVER LAPOINTE; AMBER LEAHEY; CYNTHIA LISÉE; DR. RONG LUO; LACHLAN MACLEOD; STEVE MARKS; DR. JOEL T. MINION; JEFF MOON; KAITLIN NEWSON; MIKAYLA REDDEN; CHANTAL RIPP; ÉDITH ROBERT; DR. ALISA BETH ROD; DANY SAVARD; SANDRA SAWCHUK; FELICITY TAYLER; KRISTI THOMPSON; BERENICA VEJVODA; MINGLU WANG; LEE WILSON; TATIANA ZARAISKAYA; AND DR. BIRU ZHOU

Western University, Western Libraries
London, ON

CONTENTS

Using this Textbook 1

How to Navigate This Textbook 1

Why an Open Textbook? 1

What is an Open Textbook? 2

How to Access and Use this Book? 3

Licensing and Attribution 3

Get in Touch! 4

Reference List 4

About the Editors 6

Acknowledgements viii

Foreword: Reflections on a Career in Data Librarianship x

Jeff Moon

Section I. First Principles in Research Data Management

1. The Basics: An Introduction to Research Data Management 17

An Introduction to Research Data Management

Kristi Thompson

Introduction 17

What Are Research Data? 18

What Is Research Data Management? 20

Reproducibility, Replicability, Traceability 21

Tri-Agency Policy: The Three Requirements 23

Data Management Plans (DMPs) 23

Conclusion 27

Reference List 29

2. The FAIR Principles and Research Data Management 31

Minglu Wang and Dany Savard

Introduction 31

Brief History of FAIR Principles 32

What are FAIR Guiding Principles? 33

How to Make Your Data FAIR: Tools and Guidance 36

Policy Impacts of the FAIR Principles 38

FAIR Principles and Repositories 39

Getting Involved 40

Conclusion 40

Additional Readings and Resources 42

Reference List 43

3. Indigenous Data Sovereignty: Moving Toward Self-Determination and a Future of Good Data 46

Moving Toward Self-Determination and a Future of Good Data

Mikayla Redden and Dani Kwan-Lafond

Introduction 46

The United Nations and Indigenous Self-Determination 47

A History of Indigenous Peoples and Bad Data 48

Indigenous Data: What is it? How Would It Be Different Under Indigenous Self-Determination? 49

Interacting with Indigenous Knowledge 50

First Nations Data Self-Governance in Canada 52

Conclusion 56

Additional Readings and Resources 58

Reference List 58

Section II. A Canadian Context for Research Data Management

4. Canadian Research Data Management: History and Landscape 65

Eugene Barsky; Elizabeth Hill; Tatiana Zaraiskaya; Minglu Wang; and Lucia Costanzo

Introduction 65

A Brief History of Research Data Management in Canada 66

National Collaboration: From Portage Network to the Alliance 70

Regional Efforts 77

Conclusion 82

Acknowledgement 84

Additional Readings and Resources 84

Reference List 84

5. Research Data Sharing and Reuse in Canada: Practice and Policy 89
Meghan Goodchild; Shahira Khair; Amber Leahey; Kaitlin Newson; and Lee Wilson
Introduction 89
Policies and Practices in Canada 90
Infrastructure, Tools, and Services 92
Support Services 97
Considerations for Data Sharing 101
Future of Data Sharing in Canada 106
Conclusion 108
Additional Readings and Resources 110
Reference List 111
6. The RDM Maturity Assessment Model in Canada (MAMIC) 115
Jane Fry; Jennifer Abel; Dylanne Dearborn; Alison Farrell; and Chantal Ripp
Introduction 115
The Need: How to Assess an Institution's RDM Services 116
What Is a Maturity Assessment Model? And Why Does Canada Need One? 117
How the MAMIC Was Created 118
Using the MAMIC 119
Benefits of the MAMIC 121
Conclusion 121
Additional Readings and Resources 123
Reference List 124

Section III. Working with Data

7. Data Cleaning During the Research Data Management Process 129
   Lucia Costanzo
   What Is Data Cleaning? 129
   Six Core Data Cleaning and Preparation Activities 130
   Data Cleaning Software 145
   Conclusion 146
   Reference List 147
8. Further Adventures in Data Cleaning: Working with Data in Excel and R 148
   Dr. Rong Luo and Berenica Vejvoda
   Introduction 148
   General Procedures to Prepare for Data Cleaning 149
   Data Cleaning Tools 150
   Conclusion 171
9. A Glimpse Into the Fascinating World of File Formats and Metadata 173
   Émilie Fortin
   Introduction 173
   File Formats 174
   Metadata 183
   Conclusion 192
   Additional Readings and Resources 193

10. Supporting Reproducible Research with Active Data Curation 198
Sandra Sawchuk; Louise Gillis; and Lachlan MacLeod
Introduction 198
Platforms 199
Guidelines for Data Storage 200
Data Security 201
Active Data Curation 202
Going Further 205
Conclusion 209
Reference List 210
11. Digital Preservation of Research Data 213
Grant Hurley and Steve Marks
Introduction 213
Threats to Objects Over Time 214
Digital Preservation in a Research Data Context 220
Conclusion 226
Additional Readings and Resources 227
Reference List 228

12. Data Management Planning for Open Science Workflows 230
Felicity Tayler; Mélanie Brunet; Kathleen Gregory; Lina Harper; and Stefanie Haustein
Introduction 231
What Is Open Science? 232
What Are Open Data? 233
Case Study: The Meaningful Data Counts Project 234
What Makes Open Data? Restrictions on Sharing Data 238
Can I Share Data? Determining Data Ownership 239
Conclusion 243
Reference List 245

Section IV. Considering Types of Data

13. Sensitive Data: Practical and Theoretical Considerations 251
Dr. Alisa Beth Rod and Kristi Thompson
Introduction 252
Human Participant Data 252
Other Categories of Sensitive Data 265
Preserving and Sharing Sensitive Data 266
Conclusion 268
Additional Readings and Resources 270
Reference List 271

14. Managing Qualitative Research Data 274
Dr. Joel T. Minion
Introduction 274
The Nature of Qualitative Data 275
Understanding Qualitative Research 278
Data Generation 281
Qualitative Research Data Meets RDM 285
Conclusion 289
Additional Readings and Resources 291
15. Managing Quantitative Social Science Data 294
Dr. Alisa Beth Rod and Dr. Biru Zhou
Introduction 294
Overview of Quantitative Social Science Research 295
Managing Quantitative Social Science Research Data: Files, Formats, and Documentation 296
RDM Issues Regarding Digital Tools and Software for Quantitative Social Science Data Collection 301
Conclusion 304
Additional Readings and Resources 305
Reference List 306
16. Geospatial Research Data in Canada: An Overview of Regional Projects 308
Martin Chandler; Kara Handren; Stéfano Biondo; Amber Leahey; Sarah Rutley; and Rhys Stevens
Introduction 308
Geospatial Data and GIS 309
Regional Geospatial Projects 314
Future Directions 324
Additional Readings and Resources 326
Reference List 326

Section V. Perspectives on Research Data Management

17. Research Data Management and the Open Science Movement: Positions and Challenges 331

Cynthia Lisée and Édith Robert

Introduction 332

Positioning RDM in Open Science 332

The Benefits of RDM in Context 339

Beyond the Optimistic Discourse on Opening Science 340

Conclusion: Being Open About Open Science 344

Additional Readings and Resources 346

Reference List 346

18. A Practical Perspective on the Evolving Field of Research Data Management 349

Dr. Joel T. Minion

Introduction 349

Why the Push for RDM? 350

Whose Responsibility is It? 353

Where is the Leading Edge? 355

The Realities of Managing Research Data 358

Conclusion 360

Additional Readings and Resources 361

Reference List 362

Glossary 363

Appendix 1: Data Management Plan Template 385

Appendix 2: Sample of a Completed Section of the MAMIC 387

Appendix 3: Chapter 10 Exercises 391
Introduction 391
Part 1 (Introductory): Explore the Data and the Code Repository 391
Part 2 (Advanced): Run and Alter the Code 395
Reference List 402
Solutions 403
Chapter 7, Data Cleaning During the Research Data Management Process 403
Chapter 8, Further Adventures in Data Cleaning 405
Chapter 13, Sensitive Data: Practical and Theoretical Considerations 407
Chapter 14, Managing Qualitative Research Data 408
Chapter 17, Research Data Management and the Open Science Movement 409

USING THIS TEXTBOOK

How to Navigate This Textbook

The Table of Contents: Accessing Sections and Chapters

In the top left corner of the screen is a black tab labelled "Contents." Click this to open the Table of Contents dropdown menu. From there, you can navigate to any of the major sections or individual chapters in the book.

By clicking the plus button (+) to the right of a section, you can expand the contents to show each chapter title. These titles are clickable and will take you directly to the chapter.

"Next" and "Previous" Page Buttons

At the bottom left or right of any Pressbooks page (including this one!) are the "next" and "previous" buttons. They are labelled with the title of the previous or next chapter. You can use these buttons to go directly to the previous or next chapter without navigating back to the Table of Contents.

Glossary

At the end of the book is a glossary of terms for your reference. Where applicable, glossary definitions have also been embedded directly within the chapters and appear as underlined in the text. When clicked, the glossary definition will appear as a tooltip window.

Why an Open Textbook?

With the recent release of the Tri-Agency Research Data Management (RDM) Policy, RDM has become crucially important. All researchers who apply for grants to fund data-related research must now meet requirements including writing Data Management Plans and preparing data for archiving. Given the heightened attention to RDM, the need for greater education and the number of courses related to RDM is likely to increase.

In summer 2021, a number of Canadian academics and librarians, including faculty who teach existing RDM courses, formed a group to discuss creating a bilingual, made-in-Canada textbook. The group recognized that at the time, there were no resources suited to the unique Canadian regulatory context and appropriate for use in classrooms. Together, it was decided that an open educational resource (OER) in the form of a textbook would be of the most value to Canadian practitioners and learners, and would capture the spirit of RDM which is meant to encourage openness.

What is an Open Textbook?

An open textbook is a publicly available online resource that is free-of-charge and has an open license that allows others to reuse, retain, remix, redistribute, and revise it. This book has a Creative Commons Attribution-NonCommercial (CC BY-NC) license, which allows for the adaptation and redistribution of this textbook for non-commercial purposes so long as the original creator is attributed (see "Licensing and Attribution" section). Further to the open license, the authors of this open textbook are committed to making this open textbook available immediately, freely, and permanently to anyone who can access the internet.

Benefits to using open textbooks are many. Besides simply providing freely available quality open scholarship resources to students and instructors as a significant cost savings, open resources also ensures that the intention of education is considered. UNESCO's SG4 goal to "ensure inclusive and equitable quality education and promote lifelong learning opportunities for all" by 2030 begins with freely accessible open educational resources (OER). The previous view that education is the business of disseminating knowledge has been challenged by OER advocates who are leading the education reform towards the co-creation and sharing of knowledge (Blomgren & Henderson, 2021; Cronin, 2017; Henderson & Ostashewski, 2018). In addition to the free use of an open textbook, open resources used for instruction are directly applicable to curriculum goals and can remain relevant to the field through the adaptation and revision of the resource (Hendricks et al., 2017).

While there are many commercial publishers that offer similar textbook quality, they have limitations that reduce the impact that they could have. Specifically, they are rarely permanent or freely available which limits the accessibility of these resources to many students, educators, and practitioners. This open textbook, Research Data Management in the Canadian Context: A Guide for Practitioners and Learners, responds to this call for education reform by meeting the gold open access standards of an immediate, free, and permanent open education resource that can be revised, redistributed, retained, remixed, and reused for non-commercial purposes under a Creative Commons Attribution-NonCommercial license.

In the next section, "How to Access and Use this Book," we will explore the book's intended uses.

How to Access and Use this Book?

This book is expected to meet the needs of instructors looking for resources to support their teaching in RDM topics as well as supporting the needs of librarians, students, and researchers who are seeking up-to-date materials for guidance on RDM practices. By publishing Research Data Management in the Canadian Context: A Guide for Practitioners and Learners with a CC BY-NC license, it is our intention that this book be adopted in full as required reading in the classroom, adapted in part as supplemental information, or revised with current or compelling information that the resource may lack. We are excited to offer this open educational resource as a starting point to advance the RDM field with, and for, RDM practitioners and hope to shed light on the need for more resources in this field.

Licensing and Attribution

Examples of attribution language are as follows:

Redistributing the complete book:

Research Data Management in the Canadian Context: A Guide for Practitioners and Learners created by Kristi Thompson; Elizabeth Hill; Emily Carlisle-Johnston; Danielle Dennie; and Émilie Fortin published with Pressbooks. The original is freely available under the terms of the CC BY-NC 4.0 license at https://ecampusontario.pressbooks.pub/canadardm.

Redistributing chapters:

Chapter title, authors, in Research Data Management in the Canadian Context: A Guide for Practitioners and Learners created by Kristi Thompson; Elizabeth Hill; Emily Carlisle-Johnston; Danielle Dennie; and Émilie Fortin published with Pressbooks. The original is freely available under the terms of the CC BY-NC 4.0 license at https://ecampusontario.pressbooks.pub/canadardm.

Revised or adapted versions:

This material has been adapted/revised from Research Data Management in the Canadian Context: A Guide for Practitioners and Learners created by Kristi Thompson; Elizabeth Hill; Emily Carlisle-Johnston; Danielle Dennie; and Émilie Fortin published with Pressbooks. The original is freely available under the terms of the CC BY-NC 4.0 license at https://ecampusontario.pressbooks.pub/canadardm.

For more information, see Creative Commons Attribution FAQ and Creative Commons best practices for attribution.

Get in Touch!

If you like our work and are planning to use it, we would love to know! Please send us a note to let us know how you are using the work by emailing rdmoerteam@gmail.com.

You will find the French edition of the book at this address: https://ecampusontario.pressbooks.pub/gdrcanada/. If you are interested in adapting, translating, or otherwise have suggestions for editing and updating this work, we would also love to hear from you and answer any questions you may have.

Reference List

Blomgren, C., & Henderson, S. (2021). Addressing the K-12 open educational resources awareness niche: A virtual conference response. Alberta Journal of Educational Research, 67(1), 68-82. https://doi.org/10.11575/ajer.v67i1.56965

Cronin, C. (2017). Openness and praxis: Exploring the use of open educational practices in higher education. The International Review of Research in Open and Distributed Learning, 18(5), 1-21. https://doi.org/10.19173/irrodl.v18i5.3096

Henderson, S. & Ostashewski, N. (2018). Barriers, incentives, and benefits of the open educational resources (OER) movement: An exploration into instructor perspectives. First Monday, 23(12). https://doi.org/10.5210/fm.v23i12.9172

Hendricks, C., Reinsberg, S. A., & Rieger, G. W. (2017). The adoption of an open textbook in a large physics course: An analysis of cost, outcomes, use, and perceptions. The International Review of Research in Open and Distributed Learning, 18(4), 78-99. https://doi.org/10.19173/irrodl.v18i4.3006

Henderson, S., McGreal, R., & Vladimirschi, V. (2018). Access copyright and fair dealing guidelines in higher educational institutions in Canada: A survey. Partnership: The Canadian Journal of Library and Information Practice and Research, 13(2), 1-37. https://doi.org/10.21083/partnership.v13i2.4147

ABOUT THE EDITORS

Emily Carlisle-Johnston has been Research and Scholarly Communication Librarian at Western University since 2020. She works with faculty looking to incorporate Open Educational Resources (OER) in their teaching, which includes helping faculty find and evaluate OER, assessing licensing options for re-use and adaptation of OER, and supporting the use of Open publishing software such as Pressbooks. Prior to this role she worked at eCampusOntario, where she led the editorial workflows for creation of OER. Emily completed the SPARC Open Education Leadership Fellow program in 2022. ORCID: 0000-0002-5391-723X

Danielle Dennie has been the Head, Vanier Library at Concordia University in Montréal since 2021. She has also been the Research Data Librarian at Concordia since 2018. She has a Masters in Applied Microbiology from INRS-Institut Armand Frappier as well as a Masters in Library and Information Studies from McGill University. She was the lead on Concordia University's Institutional RDM Strategy. ORCID: 0000-0003-3771-2450

Émilie Fortin has been Research Data Management and Digital Preservation Librarian at Université Laval since 2021. Prior to this, she was the librarian responsible for digital production, preservation and conservation of collections. She completed her Master's degree in Information Science at Université de Montréal, spending a year at the Haute école de gestion in Geneva. She is involved in the Digital Research Alliance's Preservation Expert Group as well as the Partenariat des bibliothèques universitaires du Québec (PBUQ) working group on research data management, and is also a regular participant in iPRES conferences on digital preservation. ORCID: 0000-0002-9717-6840

Elizabeth Hill is the Data Librarian at Western University in London Ontario. She provides access and data literacy instruction to data sources at Western. She has an external advisor role with Statistics Canada. Elizabeth is active in various data communities and working groups in participant and leadership roles. Her research interests include supporting researchers, and she has published on topics related to data delivery systems and data librarianship in Canada. ORCID: 0000-0002-9715-238X

Kristi Thompson is the Research Data Management Librarian at Western University, and previously held positions as data librarian at the University of Windsor and data specialist at Princeton University. Kristi supports research projects, administers data archiving software, works with Western's Research Ethics boards, and is involved at a national level with developing research data infrastructure. She co-edited the book Databrarianship: the Academic Data Librarian in Theory and Practice and has published on topics ranging from data anonymization algorithms to intergenerational psychology. ORCID: 0000-0002-4152-0075

ACKNOWLEDGEMENTS

Research Data Management in the Canadian Context would not have been possible without the collaboration and participation of members of Canada's academic data community, as well as representatives from agencies supporting Research Data Management.

The initial idea to create a resource of this type came from a Canadian RDM-OER listserv, which brought together RDM supporters across Canada. Lachlan MacLeod was instrumental in getting this group formed and talking about developing an open textbook on RDM. The RDM-OER listserv continued to provide input, feedback and support throughout the life of the project.

We are thankful for the project support we had from Serena Henderson during the initial phases of the project, with financial support from Dalhousie University. Yeliz Baloglu Cengay provided assistance with inputting chapters to Pressbooks.

This project could not have happened without the financial support of a number of different groups. Research Data Management in the Canadian Context is supported in part by funding from the Social Sciences and Humanities Research Council. We additionally gratefully acknowledge financial support from Compute Ontario; a Western University Research Mobilization, Creation & Innovation Grant; the Western Libraries, Western University Academic Activity Support Fund; a Concordia Library Research Grant; and a University of British Columbia OER Rapid Innovation grant. Dalhousie University provided support for hiring a Project Coordinator in the early days of the project. The Digital Research Alliance of Canada provided graphics support.

The cover design is by Amy McConchie, CCGoodwin Consulting.

Copyediting services for the original English chapters were provided by Paula Chiarcos and Amanda Feeney from Colborne Communications. Copyediting services for the original French chapters were provided by Suzanne Aubin from Colborne Communications and Jonathan Dorey. Translation from French to English was done by Jonathan Dorey and Amanda Feeney. Translation from English to French was done by Manon St-Jules and Suzanne Aubin. An additional review of the French version of chapter 3 "Indigenous Data Sovereignty" was carried out by Wintranslation.

We would especially like to acknowledge the efforts of our peer reviewers, who helped ensure the academic integrity and quality of this manuscript. The following individuals provided assistance:

Jennifer Abel
Fatoumata Bah
Lacey Cain
Alicia Cappello
Erin Clary
Mathieu Clouthier
Alexandra Cooper
Lyne Da Sylva
Sarah Forbes
Jane Fry
Meghan Goodchild
Monique Grenier
Alex Guindon
Melissa Helwig
Laurence Horton
Jasmine Hoover
Fiona Inglis
Erin Johnson
Sandra Keys
Marjorie Mitchell
Nora Mulvaney
Kaitlin Newson
Paul R. Pival
Isaac Pratt
Kharah Ross
Kimberly Silk
Tara Stieglitz
Robyn Stobbs
Carolyn Sullivan
Felicity Tayler
Arielle Vanderschans
Minglu Wang
Susie Wilson
Shiloh Williams
Nadia Zurek

FOREWORD: REFLECTIONS ON A CAREER IN DATA LIBRARIANSHIP

Jeff Moon

Recognition of Research Data Management (RDM) as a key pillar in the research enterprise has increased dramatically in recent years, driven by the efforts of data librarians and specialists, research facilitators, policy makers, funders, journal publishers, administrators in higher education, and a growing number of frontline researchers. But how did we get here? Reflecting on my 36 years working in this space, the answer is clear: community. It is the collegial and collaborative nature of the Canadian data community, working over decades, that has brought us to where we are today through the shared belief that together we can do better. Tracing the history of this progress will help frame the origins and purpose of this new Open Educational Resource (OER) RDM textbook. My recounting of our shared history will be personal and necessarily selective; far more thorough and thoughtful coverage can be found in the excellent works of Gray and Hill (2016) and Humphrey (2020).

I arrived at Queen's University in 1987, armed with a background in biology, a library degree, and a basic knowledge of statistics and mainframe computers — with the latter ultimately getting me hired as Queen's first data librarian. I believe Queen's University was one of only six Canadian institutions with data librarians at that time. Early on, I learned that data librarianship was an aerobic activity: run 9-track data tapes to the computing centre, run back to the library, execute your batch job on the mainframe, run back to the computing centre to collect printed results, run back to the library, find and fix errors, repeat. I was in the best shape of my life.

At around this time, the Federal government of the day imposed cost recovery measures that effectively raised the price tag for Statistics Canada data tenfold, from $25 to $2500 per file, putting these data well out of reach for most researchers and universities. Laine Ruus, a veteran Data Librarian at the University of Toronto, thought together we can do better. Collaborating with the Canadian Association of Research Libraries (CARL), Laine spearheaded negotiations to purchase a single set of all Census data files from Statistics Canada, to be copied and shared under license with participating institutions. The gargantuan, and wholly altruistic task of copying and shipping hundreds of magnetic tapes across the country ensured these data remained affordable and accessible for the 25 institutions who joined in.

With this success, however, came challenges — what were academic libraries supposed to do with these tapes? Librarians, more often than not those responsible for government documents, were assigned 'data librarian'

roles but in most cases had no background or training in this field. As part of the response to this, the Canadian Association of Public Data Users (CAPDU) was established in 1988, with training as one of its primary mandates. Early drivers of this training included Wendy Watkins (Carleton University) and Laine Ruus. Training was first offered informally, often one-on-one, and later more formally in conjunction with various conferences.

Wendy later partnered with Ernie Boyko from Statistics Canada to undertake a watershed project — developing and resourcing what became known as the Data Liberation Initiative (DLI), a national data service model designed to provide access to Statistics Canada data and, importantly, targeted training, for a fixed and affordable annual subscription fee. But this success took much buy-in, time, and effort. In a 1995 regional report to ICPSR, Wendy wrote: "To date, all parties are enthusiastic. What remains to be found are firm commitments to funding." By its launch in 1996, over 50 institutions had joined with each designating a 'DLI representative' and taking advantage of the dual benefits of cost savings and much-needed training. Another less tangible benefit to emerge from DLI was a nascent hub-and-spoke community of practice, with more-experienced data librarians and specialists offering support, guidance, direction, and encouragement to a growing number of new data professionals across Canada. This de facto network of expertise and mentorship helped build relationships, trust, and credibility — and is a community-building model that we are benefiting from to this day.

Fast-forwarding through time, I see the blur of progress from magnetic tapes to tape cartridges to CD-ROMs — standalone and networked in 'towers' — to the emergence of Internet data delivery via FTP and eventually the web. Baked into this latter period were many home-grown, web-based data delivery services whose cryptic names probably still resonate with data librarians of a certain age: IDLS, Equinox, QWIFS, LANDRU, ISLAND, Sherlock, and SDA. Regional training offered by DLI was often framed around one or more of these services. This patchwork of systems served as a proving ground for more ambitious national solutions to come, with several of these platforms providing subscription access to institutions across Canada.

Importantly during this period, the concept of data management arose and grew, albeit slowly. Many data librarians became involved in what was coined 'data rescue,' reflecting the reality that many government-produced data files were at risk of being lost due to ignorance, lack of funding, or neglect. More than once, Laine Ruus, a data packrat in the very best sense of the word, was asked by Statistics Canada if she had kept (managed) a copy of a data file they needed but could not find. In another example, the ICPSR regional report cited above mentions the University of Alberta Data Library rescuing 20 years of Alberta Hail Study data when that provincial government program was shuttered. These data can be found today in Borealis, the Canadian Dataverse repository.

As technology advanced, so did awareness of the importance of doing research digitally. As with the data rescue initiatives already mentioned, there was a growing understanding of how important, yet vulnerable, researcher-generated data were. In the past decade or so, the federal government and its Tri-Agency funders

issued a series of foundational policy documents outlining their stance on open science and the importance of transparency, replicability, verification, and reuse of data. Libraries as well, spearheaded by the Canadian Association of Research Libraries (CARL) and astutely led by Executive Director Susan Haigh, took an active interest in RDM. With support from CARL Library Directors, and visionary leadership from Charles (Chuck) Humphrey (University of Alberta), a roadmap for RDM in Canada emerged, culminating in the creation of CARL Portage in 2015. In 2017, I accepted the challenge of filling Chuck's rather large leadership shoes when he retired, joining Lee Wilson, then Service Manager at Portage, in continuing to develop the Canada-wide Portage Network of Experts, or NoE (a thankful nod here to DLI), which was initiated to grow and coordinate RDM capacity and training from the ground up in Canada. Together, we oversaw the transition of Portage into the Digital Research Alliance of Canada (the Alliance). The RDM team at the Alliance and the NoE, now led by Lee Wilson, continue the work of Portage through close collaboration with others in the Digital Research Infrastructure ecosystem to improve data management practices, platforms, services, and training across Canada.

Shortly after Portage was launched, I was asked to map out a graduate-level RDM syllabus for the Library School at Western University. After much searching, I ended up choosing a textbook written in the United Kingdom as a foundation for the course. While well-written and thorough, this textbook relied entirely on UK- and European-based tools, policy frameworks, and examples. And while many aspects of RDM transcend national boundaries, bringing the topic home for Canadian students would have been of great value. Others have expressed similar frustration in seeking authoritative home-grown RDM support.

Portage, and now the Alliance, have done much to address RDM training needs in Canada, working closely with the RDM NoE, and in particular the National Training Expert Group (NTEG) to create a range of webinars, templates, guides, glossaries, videos, and primers – all freely available on the alliancecan.ca website. At the same time, others in the RDM community recognized more could be done. Of particular note, Lachlan MacLeod from Dalhousie University initiated grassroots discussions about the creation of an open textbook on RDM, convening community calls and establishing a mailing list for interested participants. A core national editorial team was formed, consisting of Elizabeth (Liz) Hill, Kristi Thompson, and Emily Carlisle-Johnston, all from Western University [English] and Danielle Dennie (Concordia University) and Émilie Fortin (Université Laval) [French].

The English editorial team worked on initial concept development for the textbook, fundraising, and editing of English-language submissions. Liz Hill brings a wealth of data and RDM experience, has deep awareness of the history of data services in Canada (see article, cited below, and historical chapter included in this work), and knows/is known by just about everyone in the Canadian data ecosystem. She served as consummate people- and relationship-wrangler for the project. Kristi Thompson brings a background in computer science and quantitative analysis to the project, which along with previous editorial experience, she leveraged to review technical content in this textbook. She is known for her work on data anonymization (see the Sensitive

Data chapter in this work) and quantitative literacy, and her involvement in 'data rescue,' all grounded in strong RDM expertise. Kristi also led very successful fundraising efforts for the project. Emily Carlisle-Johnston brought essential expertise in OER, copyediting, and textbook development to the editorial team. Her knowledge of the Pressbooks open-publishing platform, her advocacy for openness throughout the project's workflow, and her previous experience leading the editorial process for OER projects while working at eCampusOntario, made Emily a perfect fit for this project.

The French editorial team was responsible for overseeing translation, reviewing French contributions, and leading the production of a complete French edition of the text. Émilie Fortin has a range of experience and a background in preservation, and in addition to her editorial work she contributed crucial material on metadata and formats to this textbook. She has been working in RDM since 2021. Danielle Dennie has a background in science librarianship as well as RDM and has held several library leadership roles. Danielle was the primary coordinator between the English and French sides of the project, liaising with the English project team and juggling copy editors and translators. Danielle and Émilie co-led outreach with the French data community and translated communications for the project.

This core national editorial team had a diverse range of skills and levels of experience, with each member contributing in distinct but complementary ways. Their collective efforts ultimately attracted over 50 members of the Canadian data community to serve as editors, authors, reviewers, fundraisers, and other contributors to this project. This larger pan-Canadian team had a shared appreciation of the value and importance of framing RDM training and resources in the Canadian context and set out to fill this need, culminating in this all-Canadian bilingual RDM textbook — Research Data Management in the Canadian Context: A Guide for Practitioners and Learners.

It is exciting to think how valuable and appreciated this work promises to be as part of an ever-growing arsenal of Canadian RDM training resources. This textbook is aimed at researchers and practitioners at all levels and from all disciplines. It has strong potential for use:

- in teaching (Library School courses, workshops, etc.)
- as a reference source (by researchers and RDM specialists, new and established)
- by administrators hoping to learn more about policy and regulatory aspects of RDM
- as a driver of change, with applications in policy discussions, development, and deployment.

The online and open nature of this work will facilitate access and ongoing improvement. The RDM landscape is constantly changing with advancements being made locally, regionally, nationally, and internationally — all with the potential to inform and augment this textbook over time.

Fundamentally, this textbook is the embodiment of a sea change in the Canadian data ecosystem. We are witnesses to and participants in the broadening of our collective national focus from solely facilitating access

to and use of existing data, to proactively expanding available content by promoting and supporting the FAIR-ification of researcher-generated data in the ways described in this work. The best practices, tips, guidance, policy discussions, and examples in this textbook will certainly bolster efforts to normalize the necessary and growing focus on FAIR. I say normalize, because we do need to make the best practices surrounding research data management a normal and expected part of researchers' mindsets and workflows — not just in response to policy imperatives, but because researchers recognize and value the benefits of data well managed, for their disciplines, for their reputations, for future reuse and verification, and for society at large. This textbook will help us, together, to reach this goal. Never underestimate the power of a dedicated community to get things done.

March 2023

Gray, S. V. & Hill, E. (2016). The Academic Data Librarian Profession in Canada: History and Future Directions. Western Libraries Publications. Paper 49. http://ir.lib.uwo.ca/wlpub/49

Humphrey, C. The CARL Portage Partnership Story. (2020). Partnership: The Canadian Journal of Library and Information Practice and Research, 15(1). https://doi.org/10.21083/partnership.v15i1.5825

About the author

Jeff Moon

Jeff Moon is the Director, Data Strategy and Services at Compute Ontario.

SECTION I

FIRST PRINCIPLES IN RESEARCH DATA MANAGEMENT

1.

THE BASICS: AN INTRODUCTION TO RESEARCH DATA MANAGEMENT

An Introduction to Research Data Management

Kristi Thompson

Learning Outcomes

By the end of this chapter you should be able to:

1. Define the terms *research data*, *Research Data Management*, and *Data Management Plan*.
2. Describe the three elements of the 2021 Tri-Agency Research Data Management Policy.
3. Understand the link between Research Data Management and research replicability.
4. List some common elements of a Data Management Plan and explain their importance.

Introduction

In 2021, Canada's three federal research funding agencies, the Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council (NSERC), and the Social Sciences and Humanities Research Council (SSHRC), released the **Tri-Agency Research Data Management Policy**. The policy's stated goal is to ensure that "research data collected through the use of public funds should be responsibly and securely managed and be, where ethical, legal and commercial obligations allow, available for reuse by others" (Government of Canada, 2021). Funding agencies in many other countries have released similar policies.

This chapter will discuss some of the fundamental questions of Research Data Management (RDM) in Canada: Where is the push towards formal RDM coming from? What is research data, in terms of this policy and in general? What are the requirements of good data management?

Canada's Three Federal Research Funding Agencies

The Natural Sciences and Engineering Research Council of Canada (NSERC), the Social Sciences and Humanities Research Council of Canada (SSHRC), and the Canadian Institutes of Health Research (CIHR), are Canada's three federal research funding agencies. They are sometimes collectively referred to as the Tri-Council or the Tri-agency; throughout this text they will often be collectively referred to as ***the agencies***. As the source of a large share of Canada's research money, they are able to set policies that significantly influence how research is conducted in Canada. In addition to the Tri-Agency Research Data Management Policy, they are responsible for the **Policy Statement on Ethical Conduct for Research Involving Humans (TCPS 2)**, the Open Access Policy on Publications, and others. Their policies are not laws. However, in addition to deciding whether or not to award funding to individual researchers, the agencies can each bar entire institutions from administering research funds, which would make every researcher at that institution ineligible to apply for funds. This gives the agencies a huge amount of power to shape how research is done in Canada.

What Are Research Data?

To understand RDM requirements, you have to understand the definition of **research data**. The term *research data* combines two key concepts: research and data. Research might be described as a systematic process of investigation, a way of finding out about things. Research transforms information into knowledge and is a part of how we discover the world. Data can be an important part of that knowledge discovery. Data are one type of information or evidence that serve as input to research. But not all information in a research project is data.

Canada's Tri-Agency FAQ (2021) states that "What is considered relevant research data is often highly contextual, and determining what counts as such should be guided by disciplinary norms" (Government of Canada, 2021b). In short, context is important; you can't really define research data without looking at how it's being generated and used. The FAQ section "How are research materials related to research data?" delves

into this: "Research materials serve as the object of an investigation, whether scientific, scholarly, literary or artistic, and are used to create research data. Research materials are transformed into data through method or practice."

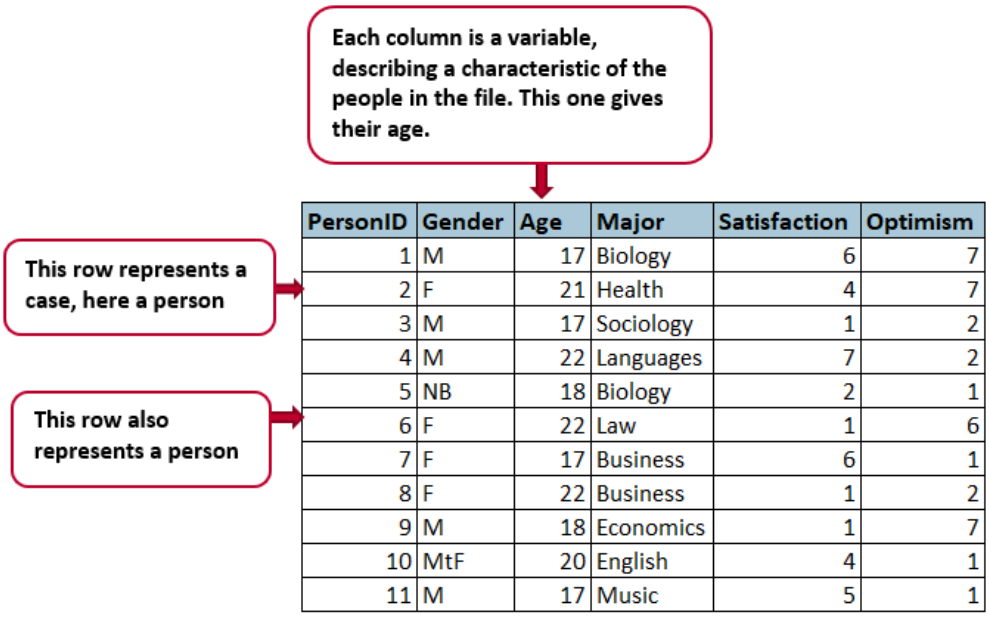
That transformation is a key part of separating general information from research data. Data are the results of taking raw information from any source (e.g., informants/survey respondents, archival or bibliographic data, social media, scientific instruments, document text) and collecting or assembling that information into a structured form to serve as an input for further research. Because of the work that goes into structuring, annotating, and organizing research data, they can also be considered a research output, along with books, articles, and other items created by researchers. Research data are a vital source of information that may not be captured in any other source. If they are published or shared, they can be referred to by other researchers and cited just like any other research output.

For example, a researcher may use a set of research articles as input for their research. If the researcher is simply reading those articles and referring to their contents through citations to support other ideas, the articles are serving as research material, but not research data. However, if the researcher takes the same set of articles, imports them into a piece of software, and reviews and annotates them in a structured way to come to some sort of formal conclusion on the group of articles as a whole, then those articles form a dataset and are considered research data.

Research data can be secondary data, meaning that the researcher did not collect or assemble the material themself. In this case, the structuring or refining to serve as input may have been done by another researcher. Or the data may come pre-structured if it's **administrative data** (say, extracted from an admissions database). But something that is a structured collection of information that is being refined into research through analysis is still considered research data.

Data Structure

A common structural format for data, used in spreadsheets and statistical files, is the rectangle, in which data are organized into rows and columns. Each row will contain one case, which is a single unit of the thing being studied (e.g., one person in a survey, or a fruit fly in an experiment). Each column will be used to store one variable or characteristic of each case, such as the age of each person (or fruit fly!) in the study.



PersonID	Gender	Age	Major	Satisfaction	Optimism
1	M	17	Biology	6	7
2	F	21	Health	4	7
3	M	17	Sociology	1	2
4	M	22	Languages	7	2
5	NB	18	Biology	2	1
6	F	22	Law	1	6
7	F	17	Business	6	1
8	F	22	Business	1	2
9	M	18	Economics	1	7
10	MtF	20	English	4	1
11	M	17	Music	5	1

Figure 1. *An image showing a rectangular data file. It's a spreadsheet with one row for each person in the dataset and a column for each characteristic.*

While we're talking about data structure, here are some simple rules for organizing rectangular, spreadsheet-style data to make it easier to manage:

- Organize the data as a single rectangle, with subjects/cases as rows and variables/features as columns; add a single row as a header at the top, with brief, descriptive names for what is in each column.
- Put just one thing in a cell and do not merge cells. Every cell should have one piece of information that corresponds to one row and one column (one case and one variable).
- Create a data dictionary — a separate document explaining what is in your rows and columns.
- Do not include calculations or functions in the original data files.
- Do not use font colour or highlighting as data.

The figure above shows what data structured this way will look like. Data in this simple format can be read by and used in any spreadsheet program or statistical package.

What Is Research Data Management?

Research data management is a general term that describes what researchers do to structure, organize, and maintain data before, during, and after doing research. In this sense, anyone who collects or uses data for the purpose of doing research is doing research data management. Creating a data file and deciding where to save it, renaming a data file, or moving it are all research data management activities. Research Data Management

(RDM), spelled with capitals, is an emerging discipline that is concerned with researching and developing ways to manage research data more effectively. The idea behind data management is to use a set of techniques to structure, organize, and document the information that is serving as input to research and to do so in a way that will allow others to understand and reproduce your research and make use of the data that went into your research.

The **research data lifecycle** is often used to illustrate the cyclical nature of research. Researchers start by planning their research. They then collect, process, and clean data to get them into shape for analysis and analyze them to form conclusions about their research. Finally, they take steps to preserve the data for the long term and make them available for others to use and study. In practice, the cycle is more complex, with many steps happening at the same time. For example, preservation of the original data needs to start as soon as the data has been collected to avoid any possibility of loss, and researchers will often process, analyze, and reprocess their data as they work with them. This is a very data-centric view of research, as the research cycle will include many other steps, from applying for funding to writing up and publishing results.

Figure 2. *Research data lifecycle.*

Reproducibility, Replicability, Traceability

Reproducibility, replicability, and traceability are three related but distinct concepts that are important to understanding the importance of good RDM. For research to be **reproducible**, it must be possible for researchers who were not part of the original research team to repeat the research using the same data,

methods, and code, and to get the same results. In practice this means data, code, and thorough documentation need to be available to external researchers.

For research to be **replicable**, researchers who were not part of the original research team need to be able to repeat the original research study on newly collected or different data and get the same or similar results. For this to be possible, the original researchers' methods need to have been documented and published, but the original data do not need to be available.

For research to be **traceable**, researchers who were not part of the original research need to be able to reproduce the analysis dataset from the original, as collected or acquired datasets. If data are traceable, everyone can be confident that no undocumented changes happened to the dataset. External researchers should also understand why every change made to the data happened, who made it, and what the decision process was. Research data are evidence — if you've ever watched CSI, this is like the chain of custody that ensures evidence in a criminal case hasn't been contaminated.

Remember those data structure tips from earlier in the chapter? Simple, standardized, widely understood formats and structures are good for reproducibility, replicability, and traceability.

Mandating specific standards for how data should be managed isn't meant to put arbitrary constraints on how people do research. The standards help to preserve research integrity by having researchers handle their data in ways that can be followed and understood and, therefore, reproduced and replicated. Research findings that cannot be repeated or reproduced are not credible. Mandated RDM also includes the goal of increased data sharing, not just so research can be reproduced directly, but so data can be reused for other projects, allowing for the creation of more research at a lower cost. The 2021 Tri-Agency Research Data Management Policy includes three requirements intended to help Canada move towards this goal.

Replicability Crisis

The replicability crisis is an ongoing issue in the physical and social sciences that calls the credibility of these sciences into question. Starting around 2010, psychologists began to repeat earlier studies in an effort to reproduce their findings and found they were unable to consistently do so. In one major effort to replicate 28 studies, close to half could not be replicated, and 32% showed effects opposite to that which had been originally reported (Klein et al., 2018). This means that people who rely on this research have been teaching, carrying out further research, and changing practices

based on results that may be incorrect. Similar issues have since been reported in other fields, such as biology, medicine, and economics. The original studies may have included bad data, incorrect analysis methods, or atypical samples, among many possible reasons for the discrepancies. If the original data aren't available and traceable, it's hard to tell.

Tri-Agency Policy: The Three Requirements

The three requirements laid out in the Tri-Agency Research Data Management Policy (Government of Canada, 2021) are:

1. Institutional Strategies. Institutions (generally post-secondary institutions and hospitals) that are eligible to administer Tri-Agency funding are required to develop formal RDM strategies, post them on their websites, and submit them to the agencies by a deadline. These strategies need to explain how the institution intends to support its researchers in doing better RDM and in coping with the next two requirements. Strategies submitted to the agencies are linked on their Institutional Strategies page.
2. Data Management Plans. The agencies will start requiring that researchers submit plans explaining how they intend to manage their data, at least for some funding opportunities. These plans will be considered when the agencies decide how to award grants.
3. Deposit. When grant recipients publish any articles or other outputs arising from research supported by the agencies, they will be required to deposit the data and code that support that research output into a digital repository. This is a fairly narrow requirement. A researcher may collect dozens of variables but write a paper that only makes direct use of a small subset of them. This subset is what they need to deposit. Also note that depositing is not the same as sharing. Data that is confidential or otherwise shouldn't be shared needs to be deposited in a secure private location.

Data Management Plans (DMPs)

A **Data Management Plan (DMP)** is a formal description of what a researcher plans to do with their data from collection to eventual disposal or deletion. DMPs have existed in some form or other since the 1960s (Smale et al., 2020), but adoption has been slow and, in many disciplines, it is still not widespread. Internationally, DMPs have become a frequent requirement of funding agencies, including in the United

Kingdom and in the United States. Tools and templates have been developed to help researchers write plans that meet funding agency requirements. The main tool used in Canada is called **DMP Assistant**. It is a web-based tool that asks users a series of questions about their data and research plans, with contextual help and guidance on how to answer those questions.

DMPs are intended to help researchers manage data across all phases of the research data lifecycle, from collection to sharing. They are often described as "living documents" that should be updated as needed while researchers work with their data. They can include a variety of different elements (Williams et al. (2017) identified 43 topics that may be required as elements of DMPs), and which elements may be required or useful can vary by discipline or by type of data. The elements of a DMP are intended to prompt researchers to consider how they will handle their data and what resources they will need before they start their research. The Tri-Agency Policy asks the researcher to submit a plan that addresses the following:

- how data will be collected, documented, formatted, protected, and preserved
- how existing datasets will be used and what new data will be created over the course of the research project
- whether and how data will be shared
- where data will be deposited.

Research funders, in Canada and internationally, want researchers to use DMPs to demonstrate that their data will be collected, stored, and preserved in a way that facilitates transparency, data sharing and reuse, and reproducibility of results. Researchers who do this will be given an edge when applying for funding to collect or use data. DMPs are also intended to have benefits for researchers, helping them think through and work with their data more effectively. In effect, DMP requirements are a form of social engineering, intended to nudge researchers into doing better research.

These benefits are largely unproven. In theory, carefully considering all the elements that DMPs incorporate should lead to better research, but theory sometimes collides with practice. "Indeed, an extensive literature review suggests there is very limited published systematic evidence that DMP use has any tangible benefit for researchers, institutions or funding bodies" (Smale et al., 2020). Given that DMPs are meant to enhance the research enterprise, it is unfortunate that relatively little thought seems to have been put into researching whether they actually achieve that goal or how they could be modified to do a better job.

We'll look quickly at some of the topics often covered in a DMP.

Data Collection

Researchers need to list the types of data they will be collecting or acquiring and what file formats the data will be saved in. From the start, researchers should consider formats that will allow for data preservation, sharing and reuse; good formats are ones that can be used in widely available software packages. Open formats are even better: they have published standards so that anyone with the training can write the software to read them. Open formats are future-proof.

Thinking about file naming conventions before starting data collection can be surprisingly important. Researchers who don't establish a system ahead of time are liable to end up with an assortment of files with names like "data.csv", "data2.csv", "finaldata.csv", "fixeddata.csv" and so on. An example of a system for naming and tracking different versions of a data collection might be "shortdescriptivename-changemade-date.ext". Including the change and date in the file name acts as a rudimentary form of **version control**, which will be discussed in more detail in chapter 10, "Supporting Reproducible Research with Active Data Curation." Version control should also include further systems to help enhance the traceability of the data, such as noting information about every change made to the data on a master documentation file or making all changes to the data using code that is updated and saved after each change.

Documentation and Metadata

Documentation is essential to both preservation and traceability. If a file is preserved as a sequence of 0s and 1s on disk, but no one knows what those numbers represent, then the file hasn't really been preserved. Documentation needs to include elements like a master study document noting where the data came from and how they were collected, giving columns in spreadsheets easily understood names, and recording detailed information about changes made to the data files.

Documentation can also include giving files and folders human-readable names and coming up with a sensible structure of folders and subfolders. One common form of additional documentation is the **README file**, which is simply a file included in each folder that lists the files present in that folder, describes the contents of each file, and explains any relationships between the files (e.g., if there are code files that were used to generate data files).

For many types of data, including health and survey files, codebooks are also important. Codebooks describe the structure and contents of data files according to some schema. For example, a survey codebook will list all the questions asked in a survey (which will be coded as variables), describe different possible response options, explain how the survey sample was chosen, and explain any additional variables created by researchers. Ideally, you should have sufficient documentation on your deposited data that someone who is knowledgeable in your field would be able to:

- understand and follow the steps you took to collect your data in the first place and the decisions you made along the way
- take your original data file and reproduce the changes you made to it to get your data into their final form
- run the analyses that produced your final publishable results.

The documentation section of a DMP should also include information explaining how the researchers will make sure they keep track of and record every change made to the data file. If there will be many people working with the data, it's especially important to have a system.

Code Files

Statistical programs, such as SPSS, Stata, and R, and general-purpose programming languages, such as Python, let you modify and analyze data by typing commands into a code file and then running them. Some programs, like SPSS, will also let you generate the commands using menu options. If any changes made to your data are done using code files, you will always be able to go back and figure out exactly how every change to your data happened.

Storage and Backup

Researchers can explain where they will be storing their data and how secure it will be in the storage and backup section. Storing only one copy of the data — on a personal hard drive that could fail or a USB stick that could be stepped on — is surprisingly common (Cheung et al., 2022). It's also a bad idea, as many have discovered. A system that ensures data are regularly backed up is a good idea. The 3-2-1 backup rule is a widely used standard: there should be three copies of each file, the copies should be on two different media, and one copy should be off-site. If data is stored somewhere with an automated backup system (such as a departmental server or a cloud service) then that reduces the need for additional copies since a copy will be in the backup system.

Preservation and Sharing

Research transparency and the preservation and sharing of research data are key goals of RDM, so it is essential to address them in a DMP. The gold standard for data sharing is posting a complete, well-documented dataset in an online archive, where it can be downloaded by anyone, with an open or Creative Commons license that explicitly allows it to be reused. Some licenses include the stipulation that data that are used in further research should be properly cited (though, even if that is not stipulated, it is good practice and professional courtesy to do so).

If data will be shared, the most important step is identifying an appropriate repository. There are many appropriate data repositories available. Many institutions (universities, colleges, hospitals, etc.) have institutional data repositories with features that ingest data to preservation formats. These institutions commit that the data will be preserved and backed up. Individual journals also host archives to make data relating to the papers they publish available. There are also disciplinary repositories that host particular types of data, such as genomic data or geospatial data.

However, open sharing in a repository is not always advisable, and for some kinds of data (such as medical data) sharing may be highly unethical. Confidentiality, commitments made to research subjects, Indigenous data sovereignty, data ownership, and intellectual property concerns can all be reasons why openly sharing a particular dataset is not an option. In cases like this, researchers may need to find alternative sharing methods. One possible alternative would be to share documentation about the data in a repository and invite potential users to contact the research team for access. Sometimes parts of a data collection can be shared while other parts are considered too sensitive. The potential users may need to commit to following certain ethical standards, or other conditions may be applied. In these cases, the data will need to be preserved in some other way, in a secure archive or on a private network. See chapter 13, "Sensitive Data," for more information.

The preservation and sharing section of a DMP needs to be explicit about how data will be preserved for the long term. It also needs to explain provisions for data sharing, including where it will be deposited, what parts of the data will be shared, and what access conditions there will be, if any. If data can't be shared openly, the DMP needs to explain why not.

Conclusion

Research Data Management is a general term for the work researchers do as they organize and maintain data during and after their research. It is also a growing field of practice that engages librarians, data professionals, and researchers with the question of how to best manage data to include research transparency, data

preservation, and data sharing so it can be criticized, studied, and used by other researchers and research consumers. Ultimately, RDM is about doing better research.

Reflective Questions

1. Pick a field of study and describe some examples of research data that might be used by researchers in that field. What might be some particular challenges of managing this data?
2. Read the Tri-Agency Research Data Management Policy. What does it tell you about how the funding agencies view RDM?
3. Find your (or a local) institution's RDM strategy. What does it tell you about how the institution views RDM?
4. Visit DMP Assistant or use the template in Appendix 1 and create a DMP for an imaginary research project.

Key Takeaways

- Research Data Management (RDM) is an umbrella term for the activities undertaken by researchers while they work with data. As a field of study, RDM asks you to engage with fundamental questions about the best way to perform research.
- Canada's three federal research funding agencies have a policy on Research Data Management that is intended to encourage researchers to make their research more transparent and to preserve and share their data.
- Data Management Plans (DMPs) are documents prepared by researchers to describe how they intend to manage their data. They cover many aspects of working with data, including

data collection, documentation, storage, sharing, and preservation.

Reference List

Cheung, M., Cooper, A., Dearborn, D., Hill, E., Johnson, E., Mitchell, M., & Thompson, K. (2022). Practices before policy: Research data management behaviours in Canada. *Partnership: The Canadian Journal of Library and Information Practice and Research, 17*(1), 1-80. https://doi.org/10.21083/partnership.v17i1.6779

Government of Canada. (2021a). *Tri-Agency research data management policy.* https://www.science.gc.ca/eic/site/063.nsf/eng/h_97610.html

Government of Canada. (2021b). *Tri-Agency research data management policy – Frequently asked questions.* https://science.gc.ca/site/science/en/interagency-research-funding/policies-and-guidelines/research-data-management/tri-agency-research-data-management-policy-frequently-asked-questions#1b

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr., R. B., Alper, S., Aveyard, M., Axt J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry D. R., Bialobrzeska, O., Binan E. D., Bocian, K., Brandt, M. J., Busching, R., ... Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443-490. https://doi.org/10.1177/2515245918810225

Smale, N. A., Unsworth, K., Denyer, G., Magatova, E., & Barr, D. (2020). A review of the history, advocacy and efficacy of data management plans. *International Journal of Digital Curation, 15*(1), 1-29. https://doi.org/10.2218/ijdc.v15i1.525

Williams, M., Bagwell, J., & Zozus, M. N. (2017). Data management plans: The missing perspective. *Journal of Biomedical Informatics, 71*, 130-142. https://doi.org/10.1016/j.jbi.2017.05.004

About the author

Kristi Thompson

Kristi Thompson is the Research Data Management Librarian at Western University, and previously held positions as data librarian at the University of Windsor and data specialist at Princeton University. She has a BA in Computer Science from Queens University and an MLIS from Western University. Kristi supports research projects, administers data archiving software, works with Western's Research Ethics boards, and is involved at a national level with developing research data infrastructure. She co-edited the book Databrarianship: the Academic Data Librarian in Theory and Practice and has published on topics ranging from data anonymization algorithms to intergenerational psychology. kthom67@uwo.ca | ORCID 0000-0002-4152-0075

2.

THE FAIR PRINCIPLES AND RESEARCH DATA MANAGEMENT

Minglu Wang and Dany Savard

Learning Outcomes

By the end of this chapter you should be able to:

1. Explain the history of the FAIR principles.
2. Understand some of the key meanings, requirements, and underlying mechanisms of the FAIR principles.
3. Be familiar with the tools and frameworks available to help improve the FAIRness of data.
4. Understand how FAIR principles are included and referenced in research policies and data availability policies.
5. Evaluate how research data repositories support FAIR principles.
6. Find communities or initiatives that are using the FAIR principles within the Research Data Management ecosystem.

Introduction

The **FAIR** principles (Findable, Accessible, Interoperable, Reusable) are guiding principles that aim to encourage **data stewards** to improve the ways in which **research data** can be found and reused by computational systems in today's growing, complex data ecosystem. In this chapter, we'll explore the scope of the principles and the tools you can use to evaluate and enhance the FAIRness of a dataset. We'll also discuss the impact of the principles and explore how they have been endorsed.

Brief History of FAIR Principles

Why Do We Need Guiding Principles for Research Data?

Research Data Management (RDM) requirements were first proposed by national research funders in European countries because of the rise of data intensive science. Requirements around **Data Management Plans (DMPs)**, data citation and data availability have since become important for the responsible conduct of research and have introduced new conditions for researchers seeking to publish or receive public funding (Hrynaszkiewicz et al., 2020). Since then, data stewards have helped researchers meet RDM requirements by advocating for data preservation, providing training on how to prepare data, and developing infrastructure to safely store data. While advancements in informational technology infrastructure have made computational analysis of large amounts of data possible, the corresponding rise in the number of data repositories and standards created to disseminate data in different disciplines and sectors has helped encourage silos and prevented data from being brought together for meaningful research. As a result, the need for broader principles that can enable responsible data sharing has become increasingly important for different members of the wider research data community.

Origins of the FAIR Guiding Principles

In 2014, at an unconference in the Netherlands called "Jointly Designing a Data FAIRport" (Data FAIRport, 2014) the foundational principles for **interoperable** research data were first discussed. The next year, a draft of the guide was expanded by a FAIR data publishing group from FORCE11 and published for public commenting and endorsement (FORCE11, 2014a). In 2016, Barend Mons and a group of contributors authored an article in *Scientific Data* describing the need to establish the FAIR guiding principles for digital assets (Wilkinson et al., 2016). These principles are designed to help humans and machines overcome barriers to discovering, accessing, reusing, and citing research data.

Since its original publication, a version of the FAIR principles has been maintained by GO FAIR. Over time, these principles have influenced researchers wishing to prepare their data for sharing, data repositories wishing to evaluate and improve their infrastructure, and others wishing to assess and enhance their policies to support a FAIR data ecosystem.

What are FAIR Guiding Principles?

FAIR Guiding Principles

The main purpose of the principles is to ensure that machines and humans can easily discover, access, interoperate, and properly reuse the vast amount of information available for scientific discovery. The principles are meant to be high-level and domain independent, meaning they are broad in scope and can be applied to different types of data across multiple disciplines. By refraining from assigning technical specifications, the FAIR guiding principles allow for different implementations of the data management norms and characteristics they propose.

The following overview of the FAIR principles is modified from the full list of principles and subpoints available at https://www.go-fair.org/fair-principles/:

Findable

Humans and computers should be able to easily find **metadata** and data.

Machine-readable metadata are essential for automatic discovery of datasets and services.

F1. (Meta)data are assigned a globally unique and **persistent identifier (PID)**.

F2. Data are described with rich metadata (defined by R1 below).

F3. Metadata clearly and explicitly include the identifier of the data they describe.

F4. (Meta)data are registered or indexed in a searchable resource.

Accessible

Once the user finds the data, they need to know how to access them and may require details around authentication and authorization.

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol.

A1.1 The protocol is open, free, and universally implementable.

A1.2 The protocol allows for an authentication and authorisation procedure, where necessary.

A2. Metadata are accessible, even when the data are no longer available.

Interoperable

The data usually need to be integrated with other data and need to interoperate with applications or workflows for analysis, storage, and processing.

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles.

I3. (Meta)data include qualified references to other (meta)data.

Reusable

The ultimate goal of FAIR is to optimize the reuse of data, so metadata and data should be well-described so that they can be replicated and/or combined in different settings.

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes.

R1.1. (Meta)data are released with a clear and accessible data usage license.

R1.2. (Meta)data are associated with detailed provenance.

R1.3. (Meta)data meet domain-relevant community standards.

In chapter 10, "Supporting Reproducible Research with Active Data Curation," you'll learn how to make data interoperable and reusable via active data curation.

Key Mechanisms of FAIR Guiding Principles: Metadata, Persistant Identifiers, and Licenses

Using appropriate metadata (information about data) is central to the FAIR principles. Similar to traditional research material (such as books and articles with bibliographic information), research data must be described in a structured way with **controlled vocabularies** that can be read by humans and machines so that data can be discovered and reused. As such, metadata are an integral part of research data outputs because they give the

user important information about a dataset's supporting documentation, identifiers, licenses, and other relevant elements. While metadata describing original research data should be rich and specific enough to allow humans and machines to understand the context and limitations of a dataset, they should also be offered by way of standardized descriptions so that the research data are more interpretable across different domains. To achieve this balance, researchers from various disciplines have endorsed well-developed metadata standards, such as those listed by the Research Data Alliance (RDA).

The other major mechanisms to guarantee findability and reusability of data are PIDs and licenses defining how data can be used. A publicly registered PID provides each dataset and its metadata with a unique and stable means of identification that can track any changes or movements online. Researchers sharing data on their own websites normally won't be able to assign such an identifier and are encouraged to instead deposit their data with a dedicated data repository to access support around the use of PIDs, such as **Digital Object Identifiers (DOI)** (i.e., https://doi.org/10.1000/182).

Many researchers have concerns about data misuse and are reluctant to share data broadly (Wiley et al., 2019, p. 5). Data users, on the other hand, are often not able to confidently reuse and reshare secondary data derived from an original research dataset due to a lack of clarity around data reuse permissions. To counter this issue, standard data licenses, such as Creative Commons licenses or Open Data Commons licenses, or custom data use agreements can encourage data reuse while protecting data creators' rights to credit and attribution. By providing information about how data that has been assigned a given license can legally and ethically be used, licensing helps define the terms of a relationship between data creators, publishers, and users for a particular dataset. You'll learn more about licensing data in chapter 12, "Planning for Open Science Workflows."

FAIR Data and Openness

Efforts to make data FAIR doesn't necessarily lead to data being shared openly without restrictions. For example, **data objects** could have PIDs and FAIR metadata but not be open or reusable because of the way they've been licensed. The FAIR Principles Working Detailed Document offers four levels of FAIRness for data objects within a repository to describe different potential degrees of access to data:

1. Each data object has a PID and offers FAIR metadata.
2. Each data object has user-defined metadata to give rich provenance information.
3. Data elements within data objects are FAIR but are not **open access** and have defined restrictions around reuse.
4. Data objects and data elements are FAIR and public with well-defined licenses (FORCE11, 2014b).

The FAIR guiding principles allow data stewards to participate in important data publishing decisions and also provide space for other principles to be invoked. For example, the CARE (Collective Benefit, Authority

to Control, Responsibility, and Ethics) Principles for Indigenous Data Governance published by the Global Indigenous Data Alliance in 2019 recognize the importance of Indigenous data sovereignty and of centring Indigenous Peoples' rights and interests in any dealings with Indigenous data. In many ways, the CARE and FAIR principles complement one another and guide researchers toward taking into account the varied participants and purposes associated with research data. **Indigenous data sovereignty** is further discussed in chapter 3.

How to Make Your Data FAIR: Tools and Guidance

FAIR Guiding Principles and Data Management Plans

Data Management Plans (DMPs) are required by certain funding opportunities according to the Canadian **Tri-Agency Research Data Management Policy** (Government of Canada, 2021). In DMPs, researchers describe methodologies and strategies that reflect the FAIR guiding principles. For example, researchers should effectively document data in early phases of a project so that high-quality and complete metadata can be generated for dissemination. Also, researchers should negotiate data sharing licenses with collaborators and obtain permissions to share data from research participants early in the data collection stage if they wish to deposit and preserve data in repositories that meet the FAIR guiding principles.

FAIRness Evaluation and Improvement Tools for Researchers

A variety of tools have been developed to help researchers understand the FAIR principles and how to implement certain practices that align with the principles. These tools range from simple checklists to customized resources designed around researchers' practices. Below is a list of FAIR assessment tools with different features for various user groups that are either currently available or under development. We recommend using these tools as you prepare to make your data FAIR.

1. How FAIR Are Your Data? Checklist (Jones & Grootveld, 2017)

Developed by a data services network in Europe, this is a simple one-page checklist based on the FAIR guiding principles with small modifications that make the concepts and terminologies more accessible for researchers. This checklist is a good introductory tool for researchers who are new to the field of RDM.

2. FAIR Data Self Assessment Tool (Australian Research Data Commons, 2022)

The FAIR data self-assessment tool was developed by the Australian Research Data Commons. By answering questions corresponding to the FAIR guiding principles, researchers can visualize the FAIRness of their practices for each principle and see overall FAIRness across the four principles. They can also compare their current ways of handling data with best practices, thus identifying potential areas of improvement.

3. FAIR Aware Tool (Data Archiving and Networked Services, 2021)

Developed by the Netherlands' Data Archiving and Networked Services, the FAIR Aware tool provides a more detailed assessment to help researchers understand and implement the FAIR principles. Although this tool asks researchers to identify their domain of research, role(s), and organization(s), the actual content of the assessment is the same for all users. Researchers are presented with 10 awareness questions concerning each of the FAIR guiding principles and then asked to rate their willingness to comply with recommended practices. Once answers are submitted, an overview report of the researcher's FAIR awareness levels is provided along with tips and resources on how to improve.

4. F-UJI Automated FAIR Data Assessment Tool (Devaraju & Huber, 2020)

The F-UJI (FAIRsFAIR Research Data Object Assessment Service) is designed to assess the FAIRness of research data objects based on comprehensive and detailed FAIRsFAIR Data Object Assessment Metrics (Devaraju et al., 2020).

Other Guidance on How to Make Data FAIR

Besides FAIRness assessment tools, international and national research data services have developed general and discipline-specific guidelines on making data FAIR. Examples include the following:

- OpenAIRE (an organization supporting the open science development in Europe) created the Guides for Researchers: How to make your data FAIR (OpenAIRE, n.d.)
- How to FAIR (Danish National Forum for Research Data Management, n.d.) developed through interviews with a broad group of researchers and librarians
- Top 10 FAIR Data & Software Things (Library Carpentry, n.d.) offers brief stand-alone guides on different topics and disciplines that can be used by members of research communities (i.e., astronomy, imaging, music, etc.)
- Sustainable and FAIR Data Sharing in the Humanities (ALLEA Working Group E-Humanities, European Federation of Academies of Sciences and Humanities, 2020), provides practical guidance for researchers looking to make digital humanities data FAIR.

In Canada, researchers at the University of Ottawa Heart Institute and the Ottawa Hospital Research Institute have developed a series of data handling courses, including one called FAIR Principles (Centre for Journalology, n.d.). Not much additional guidance on the FAIR principles is available within the Canadian context. Librarians or researchers interested in this area could consult *How to Be FAIR with Your Data: A Teaching and Training Handbook for Higher Education Institutions* (Engelhardt et al., 2022) for examples of FAIR-related training options at various higher-education institutions in Europe.

Policy Impacts of the FAIR Principles

The FAIR principles have been used by government agencies, academic institutions, research funders, scholarly societies, publishers, and a variety of other actors to underscore the cultural, economic, and social significance of research data stewardship. As a result, these principles have become foundational for organizational bodies looking to influence researchers and how they to manage and share data. Some examples of policy impacts include the European Commission citing FAIR as directly influencing the development of the European Open Science Cloud (Hill, 2019, p. 284) and the U.S. National Institutes of Health citing the application of the FAIR data principles in their Data Management and Sharing Policy (Office of The Director, National Institutes of Health, 2020).

In Canada, a key government recommendation in the Roadmap for Open Science (2020) is the implementation of the FAIR principles by federal departments and agencies. This plan aims to ensure the interoperability of scientific and research data and metadata standards for data products tied to government agencies and departments is in place by January 2025. In terms of research funding, the Tri-Agency Research Data Management Policy states that Canada's three federal research funding agencies — the Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Social Sciences and Humanities Research Council of Canada (SSHRC) — support FAIR guidance and expect researchers to share their data in accordance with FAIR principles and disciplinary standards where allowed by ethical, cultural, legal and commercial requirements (2021). In addition, Canadian academic publishers, such as Canadian Science Publishing (n.d.), have mirrored other journal publishers' efforts by describing the FAIR principles as framing the contents of their data availability policy. Complying with such policies can mean employing the above-mentioned researcher tools to ensure data are as FAIR aligned as they can be before being released. However, in addition to data preparation, these requirements are also meant to influence a researcher's thinking around the selection of a research data repository and how their choice will support FAIR alignment beyond the initial publication of their data.

FAIR Principles and Repositories

The FAIR principles represent an opportunity to recognize the current and potential value of data repositories. Wilkinson et al. (2016) underscore this idea in their work by discussing the benefits and limitations of data repositories and arguing these should evolve to respond to the discovery and reuse needs of researchers (pp. 2–4). Researchers should determine if a data repository meets their unique disciplinary RDM needs and allows them to comply with relevant ethical and legal requirements, and they should also consider whether their choice offers features that mirror FAIR guidance.

Research data repositories are special-purpose data containers designed to store research data and associated files and metadata to provide stable and long-term access to data outputs (Boyd, 2021, pp. 25–26). Repositories are critical pieces of digital infrastructure set up to encourage the discoverability of research data and help researchers publish and disseminate data. Which repository they choose will often depend on factors such as disciplinary norms, publisher or funder requirements, or data sharing guidelines. Additionally, a researcher may choose a repository based on such elements as the ease and convenience of the data deposit process, the types of files the repository accepts, the amount of data curation support they will receive, or the metadata schemas and controlled vocabularies a repository uses to describe the research data objects it stores. Consideration of these elements should lead researchers to select either a discipline-specific repository, a community-specific repository, or a generalist repository. Researchers can then explore whether their chosen repository puts the FAIR principles into practice by evaluating whether or not it offers some specific functions.

In their paper on the improvement of interoperability between types of repositories, Hahnel and Valen (2020) note that, to effectively function in alignment with the FAIR principles, a repository should do the following:

- assign PIDs (DOIs, ORCIDs, and GRIDs) to its data products and related materials
- offer its data alongside documented **application program interfaces (APIs)**
- support robust options for data curation and subscribe to web accessibility guidelines
- offer well-defined licenses that support data reuse
- describe its path to sustainability by documenting preservation and disaster recovery workflows (pp. 195–197).

This guidance around optimal repository features mirrors similar recommendations made by OpenAIRE and by the FAIR Sharing initiative (Cannon et al., 2021). Some of these elements are also represented in the TRUST Principles for digital repositories released by Lin et al. (2020).

To assess how some major Canadian and international data repositories have documented their commitment to FAIR principles, review the following examples:

- Federated Research Data Repository: https://www.frdr-dfdr.ca/docs/en/fair_principles/
- Zenodo: https://about.zenodo.org/principles/
- Figshare: https://knowledge.figshare.com/publisher/fair-figshare

Additionally, you can locate appropriate repositories by consulting the re3data directory, which is a multidisciplinary tool that lists more than 2,800 entries for data repositories that can be searched by specific criteria, such as API type and metadata standard. Another strong option is the FAIRsharing directory, which is endorsed by the Research Data Alliance and provides a multidisciplinary platform where researchers can look up entries for repositories, data standards, and data policies. Both tools are excellent options for finding disciplinary-aligned repositories.

Some larger commercial, community, or publisher-endorsed repositories may offer more flexible and specialized features that align with FAIR guidance. However, when selecting a repository, one should consider whether their choice allows them to adhere to disciplinary norms, access the support needed to meet ethical or legal requirements, and help fulfill responsibilities toward communities that have expectations around access to their data. A choice of repository based on alignment with FAIR principles should always be balanced with these equally important requirements.

Getting Involved

For those interested in supporting the implementation of FAIR principles on a large scale, the GO FAIR initiative brings together individuals, institutions, and organizations to collaborate on policy development, skills development, and technical standards/technology development. This is primarily achieved via GO FAIR Implementation Networks that bring partners together to support the creation of unique deliverables. To learn more about implementation networks or about how to join them, visit https://www.go-fair.org/implementation-networks/.

Conclusion

The FAIR principles have helped clarify how some goals of the RDM movement may be achieved. Along with other guiding principles, they have been endorsed by funders, publishers, and varied research communities, and they have helped connect and align efforts around supporting data access and reuse.

Researchers should monitor the evolution of the FAIR principles in terms of their influence on national and international research data ecosystems and how they impact data reuse in their own disciplines.

Reflective Questions

1. Use the FAIR Aware tool to conduct a self-evaluation of knowledge and skills for making data FAIR.
2. Use the FAIR principles as a framework to evaluate the FAIRness of the following sample datasets and identify suggestions to improve the FAIRness of these datasets:
    1. Don Valley Historical Mapping Project: https://doi.org/10.5683/SP2/PONAP6
    2. Soil and Plant Phytoliths from the Acacia-Commiphora Mosaics at Olduvai Gorge (Tanzania): https://doi.org/10.20383/101.0122
    3. CLOUD: Canadian Longterm Outdoor UAV Dataset: https://www.dynsyslab.org/cloud-dataset

Reflective Questions

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://ecampusontario.pressbooks.pub/canadardm/?p=102#h5p-5

Key Takeaways

- FAIR guiding principles are high-level goals to guide the continuous optimization of research data, metadata, and data publishing environments for easier data access and reuse across domains through implementation of PIDs, rich and standard metadata, and data licenses.
- Researchers can follow guidance and use tools to learn about FAIR principles, evaluate their current RDM practices, and plan for strategies to FAIRify their research data and publishing activities.
- The FAIR principles have influenced government policies, research funding policies, and publisher policies regarding data availability.
- Researchers can align their data management and sharing activities with the FAIR principles by ensuring they select data repositories that offer features that support FAIR compliance.
- Research data repository registries are important tools for identifying repositories that offer FAIR-aligned features as well as other features related to disciplinary norms or legal/ethical/community-based obligations.

Additional Readings and Resources

FAIR and CARE Principles

The Global Indigenous Data Alliance. (2019). *CARE principles for Indigenous data governance*. https://www.gida-global.org/care

GO FAIR. (n.d.). *FAIR principles*. https://www.go-fair.org/fair-principles/

Research Data Alliance. *Metadata standards catalogue*. https://rdamsc.bath.ac.uk/

The FAIR Principles and Repositories

re3data directory. https://www.re3data.org

FAIRsharing directory. https://fairsharing.org/databases/

Getting Involved

GO FAIR Implementation. https://www.go-fair.org/implementation-networks/

Reference List

ALLEA Working Group E-Humanities. (2020). *Sustainable and FAIR data sharing in the humanities: Recommendations of the ALLEA working group e-humanities*. https://doi.org/10.7486/DRI.TQ582C863

Australian Research Data Commons. (2022). *FAIR data self assessment tool*. https://ardc.edu.au/resources/aboutdata/fair-data/fair-self-assessment-tool/

Boyd, C. (2021). Understanding research data repositories as infrastructures. *Proceedings of the Association for Information Science and Technology*, *58*(1), 25–35. https://doi.org/10.1002/pra2.433

Canadian Science Publishing. (n.d.). Principles and policy on data availability. https://cdnsciencepub.com/about/policies/principles-and-policy-on-data-availability

Cannon, M., Graf, C., McNeice, K., Chan, W. M., Callaghan, S., Carnevale, I., Cranston, I., Edmunds, S. C., Everitt, N., Ganley, E., Hrynaszkiewicz, I., Khodiyar, V. K., Leary, A., Lemberger, T., MacCallum, C. J., Murray, H., Sharples, K., Soares E Silva, M., Wright, G., ... (Moderator) Sansone, S-A. (2021). *Repository features to help researchers: An invitation to a dialogue*. Zenodo. https://doi.org/10.5281/zenodo.4683794

Centre for Journalology Training. (n.d.). *FAIR principles*. https://journalologytraining.ca/courses/fair-principles/

Danish National Forum for Research Data Management. (n.d.). *How to FAIR*. https://www.howtofair.dk/

Data Archiving and Networked Services. (2021). *FAIR Aware*. https://fairaware.dans.knaw.nl/

Data FAIRport. (2014). *Data FAIRport conference: Jointly designing a data FAIRport*. https://www.datafairport.org/component/content/article/8_news/9_item1/index.html

Devaraju, A. & Huber, R. (2020). *F-UJI – An automated FAIR data assessment tool (v1.0.0)*. Zenodo. https://doi.org/10.5281/zenodo.4063720

Devaraju, A., Huber, R., Mokrane, M., Herterich, P., Cepinskas, L., de Vries, J., L'Hours, H., Davidson, J., & Angus W. (2020). *FAIRsFAIR data object assessment metrics (0.5)*. Zenodo. https://doi.org/10.5281/zenodo.6461229

Engelhardt, C., Biernacka, K., Coffey, A., Cornet, R., Danciu, A., Demchenko, Y., Downes, S., Erdmann, C., Garbuglia, F., Germer, K., Helbig, K., Hellström, M., Hettne, K., Hibbert, D., Jetten, M., Karimova, Y., Kryger Hansen, K., Kuusniemi, M. E., Letizia, V., McCutcheon, V., ... Zhou, B. (2022). D7.4 *How to be FAIR with your data. A teaching and training handbook for higher education institutions*. Zenodo. https://doi.org/10.5281/ZENODO.5665492

FORCE11. (2014a, September 1). *FAIR data publishing group*. Archived groups. https://force11.org/group/fair-data-publishing-group/

FORCE11. (2014b, September 10). *Guiding principles for findable, accessible, interoperable and re-usable data publishing version b1. 0*. https://force11.org/info/guiding-principles-for-findable-accessible-interoperable-and-re-usable-data-publishing-version-b1-0/

Government of Canada. (2020, February). *Roadmap for open science*. https://www.ic.gc.ca/eic/site/063.nsf/eng/h_97992.html

Government of Canada. (2021). *Tri-Agency research data management policy*. https://www.science.gc.ca/eic/site/063.nsf/eng/h_97610.html

Hahnel, M., & Valen, D. (2020). How to (easily) extend the FAIRness of existing repositories. *Data Intelligence*, *2*(1–2), 192–198. https://doi.org/10.1162/dint_a_00041

Hill, T. (2019). Turning FAIR into reality: Review. *Learned Publishing*, *32*(3), 283–286. https://doi.org/10.1002/leap.1234

Hrynaszkiewicz, I., Simons, N., Hussain, A., Grant, R. and Goudie, S., 2020. Developing a research data policy framework for all journals and publishers. *Data Science Journal*, *19*(1), 1-15. http://doi.org/10.5334/dsj-2020-005

Jones, S. & Grootveld, M. (2017). *How FAIR are your data?* Zenodo. https://doi.org/10.5281/ZENODO.1065990

Library Carpentry. (n.d.). *Top 10 FAIR data & software things*. Zenodo. https://doi.org/10.5281/zenodo.2555498

Lin, D., Crabtree, J., Dillo, I. Downs R. R., Edmunds R., Giaretta, D., De Giusti, M., L'Hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M. E., Mokrane, M., Navele, V., Petters, J., Sierman, B., Sokolova, D. V., Stockhause, M., & Westbrook, J. (2020). The TRUST principles for digital repositories. *Scientific Data*, *7*, 144. https://doi.org/10.1038/s41597-020-0486-7

Office of The Director, National Institutes of Health. (2020). *Final NIH policy for data management and sharing*. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html

OpenAIRE. (n.d.). *How to make your data FAIR*. https://www.openaire.eu/how-to-make-your-data-fair

Wiley, C. A. & Burnette, M. H., (2019). Assessing data management support needs of bioengineering and biomedical research faculty. *Journal of eScience Librarianship*, *8*(1), 1-19. https://doi.org/10.7191/jeslib.2019.1132

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), 160018. https://doi.org/10.1038/sdata.2016.18

About the authors

Minglu Wang

Minglu Wang is a Research Data Management (RDM) Librarian at York University. She has published book chapters, conference/working papers, and research articles closely related to academic libraries and RDM services. Minglu Wang is an active member of the Association of College & Research Libraries (ACRL), a division of the American Library Association (ALA), and she contributed to multiple years of the Association's publications of Top Trends articles and Environmental Scan white papers. She is a member of the Research Intelligence Expert Group, a part of The Digital Research Alliance of Canada (The Alliance) RDM Team, and has participated in the design and report writing of the RDM Capacity Survey of Canadian Institutions. Email: mingluwa@yorku.ca | ORCID: 0000-0002-0021-5605

Dany Savard

Dany Savard is the Associate Librarian for Collections and Research Services at the University of Toronto Mississauga Library. He has contributed to research articles on the topics of research data discovery and data repositories. He is a member of the Alliance Network of Experts' Discovery and Metadata Expert Group and is the current Chair of its Canadian Data Repositories Landscape Working Group. He holds an MLIS from Western University and a Master of Arts in Public Policy and Administration from Toronto Metropolitan University. Email: dany.savard@utoronto.ca | ORCID: 0000-0001-7472-7390

3.

INDIGENOUS DATA SOVEREIGNTY: MOVING TOWARD SELF-DETERMINATION AND A FUTURE OF GOOD DATA

Moving Toward Self-Determination and a Future of Good Data

Mikayla Redden and Dani Kwan-Lafond

Learning Outcomes

By the end of this chapter you should be able to:

1. Articulate the importance of Indigenous data sovereignty and its role in Indigenous self-determination.
2. Identify deficit-focused data and explain why these type of data are harmful.
3. Identify the differences in assumptions made by Western/dominant research culture and Indigenous research culture and understand how these assumptions affect data-related decision making in the research process.

Introduction

Indigenous Peoples is perhaps one of the broadest umbrella terms frequently applied to a contemporary global population of colonized and formerly colonized peoples who, today, are politically united because of a shared history of loss and degradation under colonization. This chapter focuses on the history and present-day iterations of **knowledge theft** and **knowledge mining** (defined in this context as collecting Indigenous knowledge without seeking permission or consulting partners in the community) from Indigenous communities, as well as the Indigenous communities' sovereignty of their own data. Knowledge mining and

data sovereignty intersect because digital data is the most common way to store and archive knowledge for use by community members and researchers.

To begin, we will present a brief history of the global political community of Indigenous Peoples, with a focus on the impact of the United Nations Declaration on the Right of Indigenous Peoples (UNDRIP) in Canada. In the Canadian context and for the purposes of this chapter, *Indigenous* is used to broadly refer to three ethnically and culturally distinct groups: First Nations, Métis, and Inuit.

The United Nations and Indigenous Self-Determination

UNDRIP was adopted in 2007 by all United Nations (UN) member states except for four settler colonial nation states: Canada, New Zealand, Australia, and the United States. These four later signed the Declaration in 2012 after considerable efforts by Indigenous Peoples in these members states and their allies. UNDRIP is an extension of the human rights system, which is most clearly articulated in the 1960 Universal Declaration of Human Rights. UNDRIP contains no new human rights, but is an articulation and affirmation of rights often denied to Indigenous Peoples (Erueti, 2022).

The access to, or denial of, rights to Indigenous Peoples is globally disparate. While settler states are partially defined by their majority rule over Indigenous communities within their borders, there are always local, historical, and culturally specific aspects to these contexts that we must consider. Local contexts will impact how access and control of data, information, and knowledge are negotiated and decided on, but are too numerous to delve into here. Instead, we point to the need to also understand policy, and the role policy frameworks can play in promoting and assuring data sovereignty, while preventing knowledge theft and knowledge mining.

UNDRIP is the most well-known and recent human rights framework that seeks to redress and uphold rights for Indigenous Peoples. But it is important to mention the International Labour Convention (ILO) 169 (1989), which most nation states in Central and South America had already signed/adopted before UNDRIP was developed. The ILO is a UN-affiliated agency with a focus on workers and working conditions in member nation states. ILO 169 itself was a revision and renaming of the Indigenous and Tribal Populations Convention 107 (1957), which arose in the wake of World War II out of a concern about discrimination and oppression faced by Indigenous Peoples.

ILO 107 and the revised ILO 169 are laws in the nation states that adopt them (Hanson, n.d-a; n.d-b). ILO 169 consists of 44 articles that set minimum standards in the areas of health care, education, and employment. It also recognizes rights to **self-determination** and calls upon nations states to protect

Indigenous Peoples from displacement (Hanson, n.d-a; n.d-b). Whereas previous human rights frameworks used individual rights as the basic unit, UNDRIP extends these rights to collective groups of Indigenous Peoples, including those living as minority groups within larger nation states (as is the case in Canada). This important global framework emphasizes not only collective rights and identities, but also self-determination and the right to free, prior, and informed consent (FPIC). It also refers to historical wrongs and offers ideas for reparative measures (Erueti, 2022). The passage of UNDRIP was aspirational, insofar as it depends on each member nation to pass legislation that makes UNDRIP law (unlike ILO 169).

A History of Indigenous Peoples and Bad Data

Engagement between Indigenous Peoples and the governments in their Anglo-colonized countries centres on administrative policies and the programming that stems from them. This is certainly true in a Canadian context, where the mandate for Indigenous Services Canada (ISC) reads, in part, "Our vision is to support and empower Indigenous peoples to independently deliver services and address the socio-economic conditions in their communities" (Indigenous Services Canada, 2022). ISC focuses on disadvantages and social disparities of Indigenous Peoples and how the colonial nation state can help them. The same can be said when looking at mandates by the United States Department of the Interior Bureau of Indian Affairs (2023) and at the National Indigenous Australians Agency's Closing the Gap framework (2019) (for a detailed analysis of these policies, see Walter et al. in the Additional Resources section). Each of these colonial organizations situate data as the basis for their policy decisions.

Across all these countries, data paint a picture of Indigenous Peoples as having poorer health, lower education levels, and lower socio-economic status, which results in, and often numerically justifies, their startling high rates of incarceration, victimization, and suicide. All of these nations had active policies to assimilate Indigenous Peoples into Anglo-colonial society by forcibly removing children from their families and communities. These data are not disputed by Indigenous folks, but the social, racial, and cultural assumptions made by those collecting the data are questioned (Walter & Andersen, 2013). These assumptions provide us with only a narrow, colonized snapshot of Indigenous realities (Walter & Suina, 2019). As a result, the policies and programs developed using these data do not reflect the needs of Indigenous Peoples. All data collected from Indigenous Peoples should be their own to control, access, interpret, and manage.

This chapter will introduce this idea, known as **Indigenous data sovereignty**, which is defined as the right of Indigenous Peoples to collect, access, analyze, interpret, manage, distribute, and reuse all data that was derived from or relates to their communities. This chapter will also discuss the frameworks and strategies that affirm Indigenous data sovereignty in the dominant research culture.

Indigenous Data: What is it? How Would It Be Different Under Indigenous Self-Determination?

Indigenous data is a broad term referring to information and knowledge about individuals, groups, organizations, ways of knowing and living, languages, cultures, land, and natural resources. It exists in many formats, including **traditional knowledge**, which is defined as information that is passed down between generations. Traditional knowledge includes languages, stories, ceremonies, dance, song, arts, hunting, trapping, gathering, food and medicine preparation and storage, spirituality, beliefs, and world views. Indigenous data also include born-digital and digitized data collected by researchers, governments, and non-governmental institutions (Walter 2018; Kukutai & Taylor, 2016; Walter et al., 2021; Walter & Suina, 2019).

Across colonized nations, Indigenous data collected by governmental and non-governmental researchers are focused on differences, disparities, disadvantages, dysfunction, and deprivation of Indigenous Peoples — abbreviated as 5D data by Walter (2016; 2018). 5D data are lacking in social and cultural context due to their collection and analysis by researchers and policymakers coming from non-Indigenous world views and comparing the data against their colonial realities. No matter the analyses conducted, the policy-forming statistics are invalid because they are produced from 5D data and, therefore, focus entirely on deficits (Walter & Suina, 2019).

Data needs vary widely among Indigenous communities, but there is a consensus that all Indigenous data should reflect the social, political, cultural, and historical realities of Indigenous lives so that it can be used to support the self-determined needs of Indigenous Peoples (Walter, 2018; Walter & Suina, 2019). These data needs are central to the global Indigenous data sovereignty movement and are affirmed by the UNDRIP.

The Indigenous data sovereignty movement advocates for self-governance, meaning that Indigenous Peoples would control all aspects of the research process, from idea conception to use of resulting data. Without Indigenous data sovereignty, there is no way to ensure that Indigenous data reflects the rich diversity in Indigenous world views, ways of knowing, priorities, cultures, and values (Walter & Suina, 2019).

Indigenous Data Self-Governance Organizations in Anglo-Colonized Nations

- Canada: First Nations Information Governance Centre
- Australia: Maiam nayri Wingara

- New Zealand: Te Mana Raraunga
- United States Data Sovereignty Network

Interacting with Indigenous Knowledge

Before getting into best practices for working with Indigenous data, you should be aware of the following assumptions that differ between Indigenous and Eurocentric research practices.

Table 1. ***Differences between Eurocentric and Indigenous research practices.***

Eurocentric assumption	Indigenous assumption
Researchers remain objective and unbiased.	**Research is NOT objective and unbiased.** It can't be. Researchers are connected to all living things — this includes the human or non-human subjects of their research. Emotions are connected to cognition. When we think, we use reason, which is tied to our emotions, making research subjective.
Research is planned and led by the researcher(s).	**Research is community based.** Community members always shape the research question. No matter the topic, research allow us to gather knowledge that works toward one common goal: to create social action. Knowledge paired with action leads to social change.
The researcher is largely unaffected by the research process.	**Personal growth** of the researcher is an important result (because research is subjective).
No piece or member of the sample is more valuable than the others (outside of a case study).	**The eldest community members most likely carry the most valuable knowledge.** If elders are not involved in the research process, it is not based in traditional knowledge. One caveat to keep in mind here is that not all elder community members are "Elders." There are younger community members who carry traditional knowledge or language. Therefore, the terms "Traditional Teacher," "Knowledge Keeper," or "Language Keeper" are more descriptive than simply referring to all traditional peoples as "Elders."

In addition to these four assumptions, Indigenous Peoples consider the following four Rs during the entire research cycle, including publication: Relationality, Respect, Reciprocity, and Responsibility.

- **Relationality** is the centre of everything in Indigenous world views and knowledge systems (Wilson, 2008). Relationships inform all of our experiences; as Littletree, Belarde-Lewis, and Duarte (2020) put it, they are "at the heart of what it means to be Indigenous." As we engage with the world, it is our relationships that ensure we are accountable to our relations in every one of our interactions. Our relations include the land, our ancestors, and future generations. We are our relationships and are, in fact, made up of relationships between four realms: the intellectual, spiritual, emotional, and physical

parts of ourselves (Archibald, 2008). It is essential that researchers interested in Indigenous ways of knowing understand that all data, books, articles, stories, art, and other outputs began as relationships (Meyer, 2008). These things are a result of Peoplehood, defined as communal knowing. An example of this is shared histories and languages, ceremonies, celebrations, and life cycles (Holm, Pearson, & Chavis, 2003). Indigenous ways of knowing are where those relationships turn into actions. Some examples of this are asking questions, watching dance, listening to relations, dreaming, telling stories, experiencing life events, making art, and intergenerational activities like planting seeds and nurturing them as they grow into something we harvest in the fall. The Expressions of these ways of knowing are tangible items like documents, songs, tools, traditional dress, written and oral stories, books, food, paintings, carvings, and pottery. These Expressions are often held by knowledge organizations, such as libraries, museums, schools, sacred organizations, and Indigenous nations (Kidwell, 1993). The relationality at the centre of these items is often missed by non-Indigenous researchers and knowledge organizations. For a deeper understanding of relationality, see Littletree's (2018) conceptual model. Also see Holm, Pearson, and Chavis (2003); Archibald (2008); Wilson (2008); Meyer (2008); and Kidwell (1993), whose work informed the model.

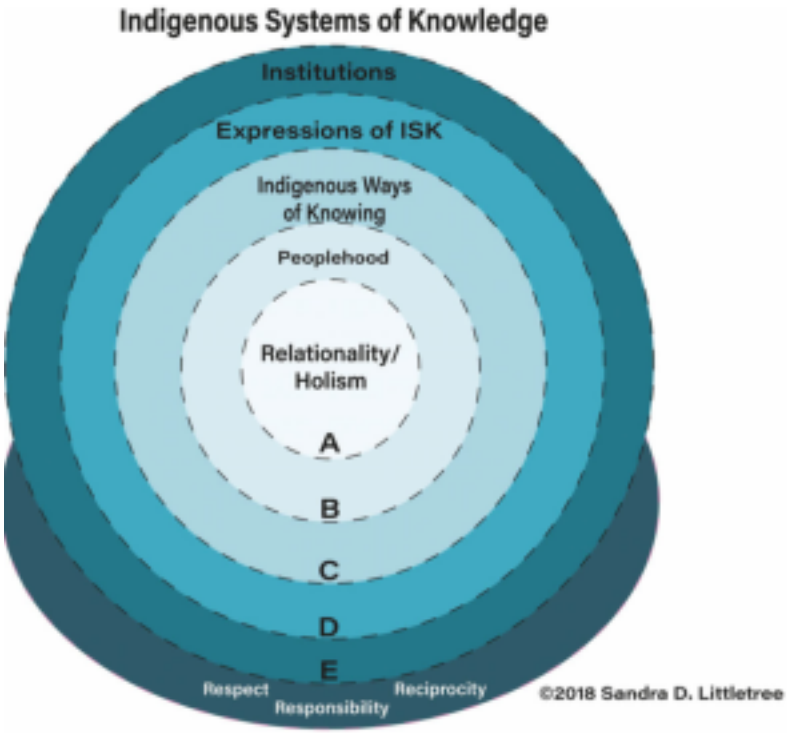


Figure 1. *Indigenous Systems of Knowledge by Sandra D. Littletree. All Rights Reserved. Used with permission.*

Respect, reciprocity, and responsibility support relationality.

- **Respect** for land, cultural protocols, history, language, and intellectual, spiritual, emotional, and physical health. Do not make assumptions about the knowledge you are working with. Use an educated, but open-minded approach. The knowledge you are inquiring about may be associated with painful historical events and elicit a great deal of trauma.
- **Reciprocity** for the information you are receiving. Be open to give and receive information. There is a long history of knowledge mining from Indigenous communities by settlers. Reciprocity does not solely refer to monetary compensation, although it is important to financially compensate individuals and communities for their time and information. Reciprocity also includes supporting communities in recovering traditions and important cultural expressions.
- **Responsibility** to obtain informed consent and nurture any relationships you have built for life — well past the end of the research project. Indigenous world views tell us that time is non-linear; it is circular. The community you are working with must guide the process and make decisions about their knowledge and information at every step.

For an in-depth look at these assumptions and considerations, see *Research Is Ceremony: Indigenous Research Methods* (2008) by Shawn Wilson and "Centering Relationality: A Conceptual Model to Advance Indigenous Knowledge Organization Practices" (2020) by Sandra Littletree, Miranda Belarde-Lewis, and Marisa Elena Duarte.

First Nations Data Self-Governance in Canada

In 1994, the federal government excluded First Nations people who live on-reserve from national population surveys (First Nations Information Governance Centre, 2022a; 2002b). In response to the data gap this created, First Nations advocates and academics formed what would later become the First Nations Information Governance Centre. Two years later, the Assembly of First Nations formed the National Steering Committee (NSC), which was tasked with developing the First Nations and Inuit Regional Longitudinal Health Survey, an early iteration of what is now known as the First Nations Regional Health Survey (RHS). The first Survey report was published in 1997 (First Nations Centre, 1997). The RHS is the only national health survey that is governed by Indigenous Peoples and based on both Indigenous and Western understandings of health and well-being. It was later reviewed by a group at Harvard University who determined that it was, "unique in First Nations ownership of the research process, its explicit incorporation of First Nations values into the research design and in the intensive collaborative engagement of First Nations people and their representatives at each stage of the research process" (Harvard Project on American Indian Economic Development, 2006).

In 1998 the NSC established a set of principles called the First Nations Principles of OCAP® to ensure that First Nations people were stewards of their own information in the same way they are stewards of their own

lands. The NSC later became the First Nations Information Governance Committee and, later, an incorporated nonprofit called the First Nations Information Governance Centre (FNIGC).

OCAP® is an acronym for ownership, control, access, and possession. These four principles govern how First Nations data and information should be collected, protected, used, and shared. OCAP® was created because Western laws do not recognize the community rights of Indigenous Peoples to control their information. The principles are reflective of Indigenous world views on stewardship and collective rights. Historically, Indigenous Peoples have not been consulted about information collected about them, nor who collects it, how they store it, or who else has access to it. As a result of this lack of self-governance, data collection has lacked relevance to the priorities and concerns of Indigenous Peoples.

The principles affirm the rights and self-determination of Indigenous communities to own, control, access, and possess information about their peoples and asks any researchers interested in conducting research with an Indigenous community to learn the principles before they begin. The principles can benefit anyone who works with (or hopes to work with) Indigenous research, data, information, or cultural knowledge and supports Indigenous Peoples' path to data sovereignty (FNIGC, 2022). FNIGC and Algonquin College have developed an online course to train researchers in the principles and history of OCAP®.

Watch the trailer for the course here: https://youtu.be/y32aUFVfCM0

The First Nations Principles of OCAP®

1. Ownership: Communities or groups collectively own their own knowledge, data, and information in the same way that individuals own their own personal information.
2. Control: Communities have control over all stages of research, from collection to storage and everything in between. Communities have control and decision-making power over all aspects of research and information that impacts them.
3. Access: Communities should be able to access their collective information and data, no matter its location. Communities should be able to manage and make decisions regarding the access to and control of their information.
4. Possession: This is like Ownership, but more concrete. It is the physical control of data, the mechanism that asserts and protects ownership of information. It may also be thought of as stewardship.

A Non-Exhaustive List of Strategies for Conducting

Research That Respects OCAP®

(Adapted from Schnarch (2005), National Aboriginal Health Organization (2005), and First Nations Information Governance Centre (2016))

- Prepare for it to take more time. You will need to get permission from community decision-makers like the Chief and Council, advisory committees, and Knowledge Keepers in addition to your research ethics board, individual participants, funding agencies, etc. Community consent is as important as the informed consent of individual participants. Research must be suspended if the community does not consent.
- Negotiate the research relationship and create a written agreement that affirms your rights and responsibilities as well as those of the community and all other partners in the research process. Be sure that all parties understand, agree, and receive a copy of the document.
- Seek funding sources that have policies that affirm Indigenous self-determination and sovereignty.
- Provide explanations and seek feedback for all aspects of the project. This can include your purpose, the anticipated benefits and risks of the project, the methods you plan to use, how you recruit your participants, how you plan to report your findings, and what you plan to do with the resulting data.
- Respect the privacy, cultural and community protocols, well-being, and individual and collective rights of Indigenous Peoples. Follow stringent ethical guidelines. Develop a code of research ethics or guidelines specific to the project. Be sure to consider that each community may have distinct interpretations and comfort levels with OCAP® and other self-determination frameworks.
- Support the interests of the community and maximize the benefits of the work. This includes building on successful Indigenous initiatives and providing opportunities for further capacity building.
- Submit all communications, summaries, and reports of your research to the community in the appropriate language prior to publication.
- Ensure that Indigenous communities have access to their data, not just the reports and resulting publications.

Critical Analysis of OCAP®

Critics might say that a necessary precursor to Indigenous controlled data and research is capacity development. They may argue that there is a lack of expertise within the community, which could lead to risks and consequences. Some would encourage First Nations, Inuit, and Métis Peoples with existing credentials in higher education to become involved, or even encourage folks from the community without existing credentials to obtain some that are related to research.

Both of these solutions could benefit individuals, but do they support nation and community building in addition to career building? Obtaining higher education credentials often requires leaving the community, which can alienate them from their communities. Not so long ago, choosing to go to university forced First Nations, Inuit, and Métis Peoples to relinquish their Indian identities and status and assimilate into white settler society. The real beneficiaries in this situation are the institutions where individuals go to study under or work for.

Opportunities to work as a full-time researcher within communities are very rare. The ability to walk in two worlds (i.e., balance Indigenous values and manage community responsibilities while advancing academic careers) is challenging, often forcing folks to make a difficult choice between the two. Furthermore, suggesting that communities cannot conduct ethical and beneficial research on their own is harmful at best. Research does not have to be specialized, use complex methodologies, or be full of scientific jargon to be beneficial.

Looking to the Future with OCAP®

At this point in time, the Principles of OCAP® are the strongest tool First Nations people in Canada and their allies have in asserting their sovereignty over data. The principles have the capacity to challenge bad data and research practices and encourage good ones. Still, there are challenges we face in moving forward in a meaningful way.

- For Research Ethics Boards: Assess all research applications going forward for OCAP® principles, or another appropriate framework, so that all research is compliant. But what about assessment of ongoing and historical research against OCAP®? After all, substantial harm has been caused to Indigenous communities via exploitative research practices. Does Truth and Reconciliation not remain a stated commitment of many educational institutions and governmental bodies?
- For policy writers: Address the ownership, control, access, and possession of data and research for all policies, and review previous policies. Some examples of existing policies that affect community-based research are institutional fire and smoking policies; data storage and dissemination policies; and intellectual property policies.
- For researchers: Be flexible, willing to compromise, and able to challenge your own assumptions about the ownership, control, access, and possession of the work that you may see as "yours." Remember that true community-based research aims to create positive and Indigenous-determined social action.
- For data management professionals: Consider the community a research project is focused on before you develop a **Data Management Plan**. Defer to the community to assess their understanding of and comfort level with data self-determination no matter the set of principles you are working from. Always approach projects with Relationality, Respect, Reciprocity, and Responsibility at the forefront of your mind.

Other Frameworks for Indigenous Self-Determination and Good Research Practices

- Canada: National Inuit Strategy on Research (2018), Inuit Tapiriit Kanatami
    - Note: to the best knowledge of the authors, a framework from the Métis Nations does not exist currently.
- Australia: Communique (2018), Maiam nayri Wingara
- New Zealand: Principles of Maori Data Sovereignty (2018), Te Mana Raraunga
- Global: CARE Principles for Indigenous Data Governance (2019) Global Indigenous Data Alliance

Conclusion

This chapter has focused on the history and present-day iterations of knowledge theft and knowledge mining, including a history of the global political community of Indigenous Peoples, especially the impact of the United Nations Declaration on the Right of Indigenous Peoples (UNDRIP) in Canada. We have contextualized the importance of Indigenous data sovereignty within this history, and have shared best practices for working with Indigenous data in order to challenge historically bad data practices. These best practices include recommendations from the Principles of OCAP®, the strongest tool First Nations people in Canada and their allies have in asserting their sovereignty over data.

Reflective Questions

1. Which assumption or consideration of reframing your research practices to incorporate Indigenous ways of knowing did you find most challenging? Why?

2. Can you identify any strategies you could deploy in your own research to be respectful of the Principles of OCAP® or other principles for self-determination?
3. Consider asking your institution to provide researchers the opportunity to complete the Fundamentals of OCAP® online training course or host the FNIGC to lead a workshop before you submit your next application to your research ethics board. If funding is not currently available, consider the following resources:

    - Watch this brief video: Understanding the First Nations Principles of OCAP®: Our road map to information governance from First Nations Information Governance Centre (2014)
    - Watch this conference presentation: First Nations data sovereignty and twenty five years of OCAP® with Aaron Franks, presented at the 2022 Canada Open Data Summit
    - Explore this webpage: The First Nations Principles of OCAP®
    - Print this brochure to keep around your workplace: The First Nations Principles of OCAP® from First Nations Information Governance Centre (2022)
    - Read this document: Exploration of the impact of Canada's information management regime on First Nations data sovereignty from First Nations Information Governance Centre (2022)
    - Read this document: Ownership, Control, Access, and Possession (OCAP®): The path to First Nations information governance from FNIGC (2014)
    - Print this infographic: Indigenous Peoples' rights in data from Global Indigenous Data Alliance (GIDA) (2022)
    - Explore this presentation: Indigenous data sovereignty and governance from GIDA (2022)

Key Takeaways

- Data gathered by colonial nation states "others" Indigenous Peoples by comparing them against an Anglo-colonized reality, which lacks social and cultural context, and focuses on

social disadvantages and disparities. This makes the policies it informs invalid. Indigenous data sovereignty is crucial if the goal is to form valid and useful policies and programs.
- Researchers informed by Indigenous ways of knowing make different assumptions than those informed by Western ways of knowing. Indigenous researchers also ensure that their work is community-based, by centering relationality, respect, reciprocity, and responsibility.
- Good Indigenous data is self-determined, meaning that Indigenous Peoples own it, control it, determine who has access to it, and oversee its storage.

Additional Readings and Resources

Lovett, R. Lee, V., Kukutai, T., Cormack, D., Rainie, S. C., & Walker, J. (2019). Good data practices for Indigenous data sovereignty and governance. In A. Daly, S. K. Devitt, & M. Mann (Eds.), Good data (pp. 26-36). Institute of Network Cultures: Amsterdam.

Toombs, E., Drawson, A. S., Chambers, L., Bobinski, T. L. R., Dixon, J., & Mushquash, C. J. (2019). Moving towards an Indigenous research process: A reflexive approach to empirical work with First Nations communities in Canada. *The International Indigenous Policy Journal*, *10*(1). https://doi.org/10.18584/iipj.2019.10.1.6

Tuhiwai Smith, L. (2012). *Decolonizing methodologies: Research and Indigenous peoples.* University of Otago Press: Dunedin, New Zealand.

Reference List

Archibald, J.-A. (2008). *Indigenous storywork: Educating the heart, mind, body, and spirit.* Vancouver: UBC Press.

Erueti, A. (2022). *The UN declaration on the rights of Indigenous Peoples: A new interpretative approach.* Toronto: Oxford University Press.

First Nations Centre. (1997). *First Nations and Inuit Regional Health Surveys, 1997.* https://fnigc.ca/wp-content/uploads/2020/09/71d4e0eb1219747e7762df4f6a133a3d_rhs_1997_synthesis_report.pdf

First Nations Information Governance Centre. (2022a). *Our history.* https://fnigc.ca/about-fnigc/our-history/#slide-1

First Nations Information Governance Centre. (2022b). *The First Nations Principles of OCAP.* https://fnigc.ca/ocap-training/

First Nations Information Governance Centre. (2016). Pathways to First Nations' data and information sovereignty. In T. Kukutai, & J. Taylor (Eds.), *Indigenous data sovereignty: Toward an agenda,* (pp. 139-156). Canberra, Australia: ANU Press.

First Nations Information Governance Centre. (2014, July 22). *Understanding the First Nations Principles of OCAP: Our road map to information governance* [Video]. Youtube. https://youtu.be/y32aUFVfCM0

Hanson, E. (n.d-a.). *ILO convention 107.* UBC Indigenous Foundations. Retrieved April 17, 2022, from https://indigenousfoundations.arts.ubc.ca/ilo_convention_107/

Hanson, E. (n.d-b.). *ILO convention 169.* UBC Indigenous Foundations. Retrieved April 17, 2022, from https://indigenousfoundations.arts.ubc.ca/ilo_convention_169/

Harvard Project on American Indian Economic Development. (2006). *Review of the First Nations regional longitudinal health survey (RHS) 2002/2003.* https://fnigc.ca/wp-content/uploads/2020/09/67736a68b4f311bfbf07b0a4906c069a_rhs_harvard_independent_review.pdf

Holm, T., Pearson, J. D., & Chavis, B. (2003). Peoplehood: A model for the extension of sovereignty in American Indian studies. *Wicazo Sa Review, 18*(1), 7–24. https://doi.org/10.1353/wic.2003.0004

Kidwell, C. S. (1993). Systems of Knowledge. In A. M. Josephy & F. E. Hoxie (Eds.), *America in 1492: The world of the Indian peoples before the arrival of Columbus,* (pp. 369–403). Vintage Books.

Indigenous Services Canada. (2022). *Mandate.* https://www.sac-isc.gc.ca/eng/1539284416739/1539284508506

Littletree, S., Belarde-Lewis, M., & Duarte, M. (2020). Centering relationality: A conceptual model to advance Indigenous knowledge organization practices. *Knowledge Organization, 47*(5), 410-426. https://doi.org/10.5771/0943-7444-2020-5-410

Meyer, M. A. (2008). Indigenous and authentic: Hawaiian epistemology and the triangulation of meaning. In N. K. Denzin, Y. S. Lincoln, & L. T. Smith (Eds.), *Handbook of critical and indigenous methodologies,* (pp. 217–232). Sage.

National Aboriginal Health Organization. (2005). *Ownership, control, access, and possession (OCAP) or self-determination applied to research: A critical analysis of contemporary First Nations research and some options for First Nations communities.* https://ruor.uottawa.ca/bitstream/10393/30539/1/OCAP_Critical_Analysis_2005.pdf

National Indigenous Australians Agency (2022). *Closing the gap.* https://www.niaa.gov.au/Indigenous-affairs/closing-gap

Schnarch, B. (2005). A critical analysis of contemporary First Nations research and some options for First Nations communities. *Journal of Aboriginal Health, 1*(1), 80-95. https://jps.library.utoronto.ca/index.php/ijih/article/view/28934

United Nations Declaration on the Rights of Indigenous Peoples. (2017). https://www.un.org/development/desa/indigenouspeoples/wp-content/uploads/sites/19/2018/11/UNDRIP_E_web.pdf

United States Department of the Interior Bureau of Indian Affairs (2023). *Mission statement.* https://www.bia.gov/bia

Walter, M. (2016). Data politics and Indigenous representation in Australian statistics. In T. Kukutai & J. Taylor (Eds.), *Indigenous data sovereignty: Toward an agenda*, (pp. 79-87). Canberra, Australia: ANU Press.

Walter, M. (2018). The voice of indigenous data: Beyond the markers of disadvantage. *Griffith Review*, *60*, 256–263.

Walter, M., & Andersen, C. (2013). *Indigenous statistics: A quantitative research methodology*. Routledge: Walnut Creek.

Walter, M., Kukutai, T., Russo Carroll, S., & Rodriguez-Lonebear, D. (2021). *Indigenous data sovereignty and policy.* Taylor & Francis: Milton.

Walter, M., & Suina, M. (2019). Indigenous data, Indigenous methodologies, and Indigenous data sovereignty. *International Journal of Social Research Methodology, 22*(3), 233-243. https://doi.org/10.1080/13645579.2018.1531228

Wilson, S. (2008). *Research is ceremony: Indigenous research methods.* Fernwood Publishing: Halifax.

About the authors

Mikayla Redden

I am a mixed-race woman; Anishinaabe with Anglo settler heritage. I am a granddaughter, daughter, sister, auntie, helper, and learner. I live and work on the Tkaronto Purchase but was born and raised on Treaty 20. Though I am a member of Curve Lake First Nation, I was not raised in the community. My great-grandfather is John 'Jack' Jacobs. Jack was married to my great-grandmother, Edith Marsden of Scugog First Nation. Jack enfranchised himself and his children under section 214 of the Indian Act in March of 1935. This means that they relinquished their Indian identities and assimilated into white settler society. Our family settled in nearby Burleigh Falls, Ontario, finding community with a local Métis settlement. The branch of the family I come from eventually moved to Keene, Ontario. I have the privilege of walking in two worlds; learning from my relations on and off-reserve, both urban and rural, traditional and contemporary, and am able to apply pieces of that knowledge to my work life, as an academic librarian.

Dani Kwan-Lafond

I am mixed-race woman, born in Treaty 4 territory, and I am a member of many communities through family and kin, including Asian, French, Jewish, and Anishnaabe communities. I teach courses focused on social inequity, race, and Indigenous-Settler relations. I do not self-identify as Indigenous and the focus of my work is on settler colonial policies and the ideologies that maintain inequity, as well as land-based learning, Indigenization, and experiential learning. I live and work, and make community, on the historical and present-day lands of the Anishnabek nation, also home to Haudenosaunee Confederacy and other Indigenous peoples, as well as to many newcomers.

SECTION II

A CANADIAN CONTEXT FOR RESEARCH DATA MANAGEMENT

4.

CANADIAN RESEARCH DATA MANAGEMENT: HISTORY AND LANDSCAPE

Eugene Barsky; Elizabeth Hill; Tatiana Zaraiskaya; Minglu Wang; and Lucia Costanzo

Learning Outcomes

By the end of this chapter you should be able to:

1. Describe the history and background of Research Data Management in Canada.
2. Identify the Canadian groups and individuals involved in Research Data Management.
3. Understand regional developments in Research Data Management.
4. Comprehend the technological tools and data repositories used collaboratively by Canadian researchers.

Introduction

Canada and many other developed countries are establishing **Research Data Management** requirements across a range of scholarly disciplines. Barriers to data management, data preservation, and data sharing, which you'll learn about in future chapters, are being addressed through the recommendation and use of community standards, such as established **metadata**, data documentation, and disciplinary repositories.

As you've now learned, the Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Social Sciences and Humanities Research Council of Canada (SSHRC) are Canada's federal research funding agencies. In March 2021, **the agencies** released the Tri-Agency Research Data Management Policy to gradually begin **Data Management Plan**

(DMP) requirements with selected grant programs. Through these programs, the agencies actively encourage research institutions to provide their researchers with an environment that enables robust **research data** stewardship and curation practices and to deliver support for managing and depositing research data in secure, curated, and accessible repositories. But even before this policy was released, visionary leaders and organizations, especially Canadian academic libraries, were carrying out grassroots data management awareness-raising initiatives and efforts.

Over the past decade, academic libraries in Canada have been working collaboratively to deliver RDM support to their communities (Steeleworthy, 2014; Liss, 2018). Collaborations between academic libraries and the broader research community address the central challenges of infrastructure, services, and training through initiatives such as the Portage Network (Portage) and Research Data Canada (RDC). Both these entities are now part of the Digital Research Alliance of Canada (Alliance).

In this chapter, we provide a brief history and overview of Canadian RDM, which began with grassroots initiatives before evolving into larger national efforts. The chapter updates and expands on previous work from a few years ago (Barsky et al., 2017).

A Brief History of Research Data Management in Canada

Since the end of the 20th century, academic libraries have discussed and advocated for centralized data archiving and data discovery services and improved access to research data in Canada. However, in a country with a relatively small and geographically dispersed population, centralization is challenging. In the early stages of RDM, Canadian academic librarians succeeded in strengthening the **social sciences** and, especially, government data collections available to researchers for **secondary analysis** purposes. The academic libraries also contributed to the development of a national RDM community of practice. By leveraging the close ties between researchers and data librarians and specialists, the network of **data stewards** was not only able to contribute collaboratively to the development of RDM tools and infrastructures, but was also able to make new resources available to local researchers through data education, consultation, and data deposit services.

Providing Access to Statistics Canada Data

The Data Liberation Initiative (DLI), a subscription-based service providing access to Statistics Canada data, is an excellent early example of how data management collaboration can help build and maintain data delivery infrastructure and train data reference experts. The DLI program began in 1996 as a result of consultations between Statistics Canada, the Canadian Association of Research Libraries (CARL), and the Humanities

and Social Sciences Federation of Canada (Boyko & Watkins, 2011). The DLI was founded in response to both the high costs of Statistics Canada's public use microdata files and the lack of data infrastructure at Canadian universities to provide access to these data (Humphrey, 2005). Due to budget cuts in the 1980s, the public use microdata files were priced on a full cost recovery basis, so only the most well-funded researchers could afford them.

The DLI collection includes thousands of data files for hundreds of survey series. Its size and the demand from researchers have directly contributed to the growth of the library data infrastructure needed to manage and preserve access to these data. When the DLI was formed, there was little expertise in many libraries to support data services. However, Statistics Canada required a point of contact within the library who would be responsible for distributing data to end-users. So libraries had to develop staff expertise through DLI training activities (Humphrey, 2005). Training programs under the DLI have led to the development of skilled library professionals and a national academic data community. The need to support the DLI program also led to the development of local initiatives to provide or improve data delivery to data specialists and users. These data delivery systems include <odesi> and Abacus in Ontario and BC, as well as systems in the Western provinces and Quebec (Gray and Hill, 2016). Sources cited for this chapter provide further in-depth reading.

National Research Data Strategy in the Early 2000s

In the 2000s, Hackett (2001) identified a wide range of issues related to Canadian research data acquisition, preservation, and access. Difficulty locating and accessing previously collected Canadian data was a key issue. This difficulty was due to high costs, a lack of a central resource directory or depository service, and a lack of a national body to set standards and provide guidance, funding, and infrastructure (Hackett, 2001). There were some exceptions. In the disciplines of physical sciences and genetics, there was already an international culture of data sharing through disciplinary repositories. The importance of data sharing to scientific practice in these disciplines led to the establishment of some Canadian repositories that did not need a policy. Examples of domain repositories include the Polar Data Catalogue (a project of the Canadian Cryospheric Information Network), the Canadian Astronomy Data Centre (an initiative of the Canadian Advanced Network for Astronomical Research), and CBRAIN (an initiative of the McGill Centre for Integrative Neuroscience, MCIN). However, the lack of interdisciplinary coordinated data curation and metadata standards still remains a problem.

For the last twenty years, the federal government has consulted with various research communities, including the National Library of Canada and the National Archives of Canada (now Library and Archives Canada), about the benefits and challenges of RDM. In 2005, the Canadian government released the National Consultation on Access to Scientific Research Data (NCASR) report. This was the cumulative work of an expert task force of more than seventy Canadian leaders in the fields of research, administration, and libraries,

among others (Strong & Leach, 2005). The report included a recommendation for the development of a national steering body to form a national data archive and coordinate data management. It also included a recommendation for project funding across sectors in Canada. However, the approach ultimately failed to gain political support (Humphrey, 2012a).

Without a national steering body or resources from the federal government, academic libraries had to forge an alternative path. They built institutional and cross-institutional repositories for disseminating and archiving data, particularly long-tail research data, which is the large number of relatively small datasets produced across many disciplines (Heidorn, 2008). These long-tail datasets are diverse and are often difficult to manage (Cooper et al., 2021). Libraries had expertise in archiving and preserving research output and a history of engagement in solutions for access to and dissemination of licensed data through their work with the DLI program. They were recognized as being well positioned to take on the challenge of managing long-tail datasets.

A Grassroots Approach to Canadian RDM Infrastructure Beginning in the 2010s

In 2008, a Research Data Strategy Working Group was formed to implement the recommendations made by the NCASR. It was a task force appointed by the National Research Council of Canada with over seventy Canadian leaders in scientific research. At the same time, CARL, a group representing Canada's largest university libraries and two federal institutions, had started participating in various national conversations about the future of Canadian digital research infrastructure. CARL gradually made the case for RDM, high-performance computing (represented by Compute Canada), and high-speed research network (represented by Canada's National Research and Education Network (CANARIE)) to be considered equally important pillars for such an infrastructure (Humphrey, 2012b).

In 2011, CARL and the Research Data Strategy Working Group held a Research Data Summit, which resulted in the formation of RDC in 2012. Since 2014, the project has been supported by CANARIE, a not-for-profit organization whose mission is to operate the national backbone network of Canada's research and education network. RDC has helped form committees and launch technical projects, and it has partnered with international organizations to advance research data infrastructure and expertise. RDC coordinated the National Data Services Framework (NDSF) Summit, first held in 2017 and again in 2019–2022. The NDSF Summit brought together RDM groups and experts, such as funding agency representatives, disciplinary data repository curators, and data librarians, from around the country. They discussed and raised awareness on the importance of prioritizing a nationally coordinated RDM infrastructure and services for the future of Canadian digital research infrastructure (Attendees of the NDSF Summit, 2019).

As part of CARL's efforts to enhance library readiness for research data support services, an RDM course was offered to libraries in early 2013. In the wake of the course, a forum called the Canadian Community of Practice for Research Data Management was created for ongoing dialogue related to RDM activities in Canada.

CARL directors created more formal relationships with the organizations providing Canadian libraries with research computing infrastructure, namely CANARIE, Compute Canada (high-performance computing), Canadian University Council of Chief Information Officers (CUCCIO), and the National Science Library. A one-year pilot project, known as project ARC, was launched in 2014 to foster a community of practice for research data in Canada. The pilot resulted in the creation of a network of experts, including academic librarians, system and code developers, and data service providers. Project ARC was a success and became the Portage Network in 2015, with the mission of providing stewardship for Canadian researchers through a network of experts across the country. As of April 1, 2021, Portage became part of the Alliance. The RDC subsequently amalgamated with the Alliance in the spring of 2022. Currently, the Alliance provides an integrated digital research infrastructure and service for all academic researchers across Canada.

National RDM Policy in the Late 2010s and Early 2020s

By 2016, following in the steps of other countries, Canada's federal research funding agencies began developing an RDM policy by releasing a "Statement of Principles on Digital Data Management." This statement proposed expectations for researchers, research communities, research institutions, and funders to collaborate on building a robust and open environment for Canadian research data.

In 2018, the agencies announced a draft RDM policy and started a public consultation. The agencies received over one hundred submissions of feedback from a variety of experts on Indigenous research, monitoring and compliance, and each of the three pillars of implementation detailed in the policy: RDM strategy, DMP, and data deposit. In March 2021, the agencies formally announced their **Tri-Agency Research Data Management Policy**: promoting excellence in RDM within the Canadian research community, while recognizing the diverse context of disciplinary scientific inquiry, legal and ethical constraints, institutional capacities, and Indigenous communities' **self-determination** and engagement. As a result of this long-anticipated announcement, the policy established an RDM support mandate within research institutions.

The policy requires each Canadian institution to submit an RDM strategy so research funders can assess readiness across institutions. Developing an RDM strategy allows institutions to think through local gaps and develop solutions, and it encourages collaboration with other institutions. The release of the Tri-Agency RDM Policy coincided with the establishment of the Alliance, a national not-for-profit organization whose goal is to harmonize and improve access to digital tools and services for Canadian researchers. A key vision of

the Alliance is to build a network of collaborative national RDM services in three areas: advanced research computing, research data management, and research software.

National Collaboration: From Portage Network to the Alliance

Origin and Current Organization

Portage was launched by CARL in 2015 in response to Canada's Action Plan on Open Government and was a precursor to the Alliance. Portage began as a community-based national network of RDM services and support that leveraged the existing national and regional networks of Canadian academic libraries. It was envisioned by dedicated RDM advocates and leaders (Humphrey, 2012b). The initial concept of the network was discussed during an informal meeting at a CARL conference in 2013.

In 2014, CARL launched a one-year community of practice pilot project, called project ARC. Building on the success of the pilot, a library-based RDM network of experts (NOE) was framed, and operational models and governance were established over the following two years (September 2015–August 2017) (Humphrey, Shearer, and Whitehead, 2016). Since then, the NOE has developed and made available numerous RDM-related training resources, guidelines, and templates aligned with the Canadian funders' requirements to support the research community and help data stewards. The NOE strengthened the connections among existing regional data repository infrastructures that used the Dataverse software, which ultimately led to the formal partnerships and the launch of the national service Borealis Dataverse Repository (Borealis). It also coordinated the development of a Data Management Plan Assistant (**DMP Assistant**) web-based application, a repository known as the Federated Research Data Repository (FRDR), and Lunaris, a data discovery platform.

After joining the Alliance in April 2021, the Portage NOE community became part of the Alliance RDM team. The future governance and operations of the NOE is currently under discussion. The NOE has grown to over 140 experts from 60 institutions across Canada. It collaborates with a broad range of interested parties and partners locally, nationally, and internationally to develop services and infrastructure so academic researchers can access the support they need for RDM (Humphrey, 2020). At the time of writing, the NOE includes the following nine active groups:

1. Curation Expert Group (CEG)
2. Data Management Planning Expert Group (DMPEG)
3. Data Repositories Expert Group (DREG)

4. Dataverse North Expert Group (Dataverse North)
5. Discovery and Metadata Expert Group (DMEG)
6. Preservation Expert Group (PEG)
7. Research Intelligence Expert Group (RIEG)
8. Sensitive Data Expert Group (SDEG)
9. National Training Expert Group (NTEG)

The efforts of the RDM community of experts have continued to advance through the efforts of the Alliance RDM team to develop shared resources, expertise, and training materials. The outputs and publications of each expert group are openly available on the Alliance website. Below are highlights of the major accomplishments of the community.

Infrastructures, Services, and Tools

Canada's current network of local and regional collaborations makes it easier and more efficient to foster national data management infrastructure, services, and tools. Data specialists and librarians from Canadian academic institutions and staff from the Alliance RDM have contributed to the development and ongoing support of the RDM infrastructures and tools mentioned in this chapter. For example, the Dataverse North Working Group was formed to bring the Dataverse repository providers and librarians in Canada together to coordinate and discuss local and national training, support services, outreach strategies, promotions, and infrastructure development and needs.

An even bigger, multi-functional data management infrastructure, FRDR, was developed with the Alliance RDM as its service provider and Compute Canada as its hardware and infrastructure host. It also had support from several expert groups, including the DMEG, PEG, and CEG. Today, FRDR provides a wide range of RDM services to Canadian institutions, organizations, and researchers, including data discovery, storage, preservation, and curation. All Canadian researchers are eligible to deposit open data in FRDR and obtain a **Digital Object Identifier (DOI)** to uniquely identify their dataset and generate a permanent web address. FRDR also has a large data ingest capacity and dedicated curation support.

FRDR originally included functionality to index data from other Canadian data repositories and make their data discoverable. However, in 2022 the decision was made to develop this capability as a separate service, named Lunaris. Lunaris is a bilingual platform that provides a single place to search for data from FRDR and other sources. Lunaris does not host data, it instead provides links to external repositories where users can go to download data.

Preservation of research data is essential to ensure that it remains accessible and usable in the long term. However, Canada still lacks a robust research data preservation plan or strategy. PEG was created to improve

Portage's capability in developing infrastructure and best practices for preserving research data and metadata. This includes working with relevant parnters on software development projects that add platforms and preservation services to the RDM infrastructure in Canada. PEG has been collaborating with other expert groups to increase awareness of preservation issues, liaising with FRDR and Borealis repositories on preservation functionality in repositories, and working with FRDR, SciNet, Scholars Portal, and University of Toronto Libraries on a preservation pipeline project to facilitate researcher access to a robust long-term **digital preservation** environments.

Initially, the online DMP Assistant tool was hosted and overseen by the University of Alberta, but later responsibility for the tool has moved to the Alliance RDM. The Tri-Agency RDM Policy highlights the importance of Data Management Plans in the research process and defines a DMP as one of three core pillars. Canada's three federal research funding agencies also announced that a DMP would soon become a requirement and not a recommendation for all Canadian researchers seeking public funding. Before this announcement, the use of DMPs was already a standard requirement for American and European research funding applications. Developed in partnership with the agencies, the DMP Assistant offers step-by-step advice for developing a Data Management Plan. In addition, the NOEs developed several bilingual documents, including guides describing how to:

- create an effective Data Management Plan,
- customize the DMP content and appearance.

There are also a number of discipline-specific DMP exemplars and templates highlighting best practices in DMPs for various disciplines within the training resources area of the Alliance website.

Best Practices, Standards, and Guidance

As a national collaborative network of experts, the Alliance RDM fostered a coordinated framework of existing, disperse infrastructure and online tools: DMP Assistant, Scholars Portal Dataverse (rebranded to Borealis, the Canadian Dataverse Repository in 2022), FRDR, and Lunaris. It also developed guidelines and recommendations on best RDM practices in close partnership with the three federal research funding agencies. The guidelines and documentation developed by the Alliance RDM working groups can be found on Zenodo and include:

- A Guide to Curating Dataverse Datasets, developed by the Dataverse Curation Guide Working Group. This guide outlines best practices for preparing datasets for publication in the Dataverse repository.
- A Dataverse North Metadata Best Practices Guide, developed and continuously updated by the Dataverse North Working Group. This guide provides an overview of metadata best practices and offers

examples from various disciplines, including geospatial data.

- Appraisal Guidance for the Preservation of Research Data, developed by the Appraisal for Preservation Working Group. The guide addresses the needs of data creators and curators to evaluate and select research data for long term access.
- Sensitive Data Toolkit for Researchers, published in 2020 and continuously updated by the SDEG. The 3-part guide includes a glossary of terms related to **sensitive data**, a data risk matrix, and a sample consent language. We've listed and provided a link to each part of the guide in the in following textbox. The guide has been widely adopted by Canadian institutions.

Sensitive Data Toolkit for Researchers Part 1: Glossary of Terms for Sensitive Data used for Research Purposes

Sensitive Data Toolkit for Researchers Part 2: Human Participant Research Data Risk Matrix

Sensitive Data Toolkit for Researchers Part 3: Research Data Management Language for Informed Consent

Network and Community Building

Besides offering RDM infrastructure and best practices, the Alliance RDM aimed to break down social, cultural, and technological barriers associated with an RDM ecosystem (Humphrey 2012b). The Alliance RDM has, in fact, cultivated a variety of networks and communities in recent years.

Members of the Alliance RDM DREG were involved in the development of the DataCite Canada Consortium, which was launched in January 2020 with Alliance RDM as the operating lead, Canadian Research Knowledge Network as the administrative lead, and funding from the Alliance. More than fifty consortium institutions worked together to develop a governance and funding structure and to offer DOI minting services and metadata registration through DataCite to all of their members. The DataCite Canada consortium is a significant achievement for Canadian institutions. It allows us to collaboratively manage the national pool of DOIs for a variety of research repositories and other digital assets while having a stable, shared, and collaborative pricing scenario for various tiers of research institutions in Canada. Also, it allows a community of practice to resolve technical issues and initiate innovative DOI projects within Canada.

To help Canadian research data repositories align their practices with global standards, the DREG adjudicated Alliance RDM funding for a cohort of CoreTrustSeal (CTS) certification applicants. A total of

12 repositories made up the first cohort of applicants, including several Borealis institutional repositories seeking improvement of current practices. CTS certification has a lengthy process before it is successful. For the benefit of applicants, DREG organizes and oversees the writing and reading groups and assists applicants with the peer-review process.

CEG is dedicated to identifying, evaluating, and promoting best practices in curating data. This includes techniques, methods, and tools that can better prepare data and metadata, improve data quality, and ultimately facilitate data dissemination and reuse. It also fills in the need for training and supporting a new generation of data curators. Community building and networking are key aspects of the expert group's approaches. In 2019, CEG hosted the first Canadian Data Curation Forum, in partnership with McMaster University and with funding from SSHRC. A key goal of this forum was to establish a national community of practice among data stewards, librarians, data service providers, and system developers. The Forum's program included a variety of keynote talks, discussions, and workshops with the objectives of facilitating communication and collaboration around data curation practices and standards and developing skill and training resources. The Forum was a huge success and achieved its goal of establishing a network of data curators who have met regularly with the CEG since then to discuss and update each other on data curation, current issues, and development.

Research and Training

To keep up with a constantly changing environment, the Alliance RDM built a research intelligence group and a training team to monitor gaps in RDM areas and to provide timely training to the community and its broader groups.

RIEG prioritizes ongoing surveillance of RDM-related topics and mandates. RIEG guides the development of best RDM practices in Canada and informs relevant communities about existing and arising issues in related policies and practices. It maintains an RDM Roadmap of Research Priorities to identify gaps in RDM knowledge, skills, services, and policies. RIEG also conducts independent studies and surveys and analyzes the results to provide evidence-based recommendations to Alliance RDM. In 2016, it established the Canadian RDM Survey Consortium and developed a common survey instrument. Fifteen universities have since used the instrument to survey researchers in their institutions to understand their RDM practices and attitudes. In 2019, RIEG conducted two surveys on Canadian institutions, measuring their RDM capacity and strategy development status, before the Tri-Agency RDM Policy was announced. The survey results provided evidence of existing RDM initiatives and services and voiced the institutions' priorities and needs for further RDM support areas.

As RDM continues to evolve, it is crucial that researchers, data professionals, and others involved with RDM have the information and training they need to stay up to date with the latest developments and best

practices. The development of RDM training resources has been one of the core activities of the Alliance RDM. Since 2017, the Alliance RDM NTEG has developed RDM training material. The NTEG oversees a range of specific projects that collaboratively develop and deliver training and resources to support RDM skill development across Canada. Immediately following the announcement of the Tri-Agency RDM Policy, NTEG coordinated a series of well-attended workshops on the most important aspects of the policy. The workshops helped researchers and others understand the policy requirements and raise awareness of existing tools and resources that could support them in developing DMP and institutional RDM strategies.

Data Repository Services in Canadian Libraries

Just as a network of experts, training, and support has been established nationally, various university libraries have also developed a Canadian data repository service. Most notably, the Dataverse repository has been a key resource. The Dataverse repository is an **open source** software, developed by Harvard's Institute for Quantitative Social Science, to store, share, cite, preserve, discover, and analyze research data. Its open source nature enables institutions to host their own installations of the Dataverse software and offer a customized solution tailored to their own community needs.

There has been an evolution from local and regional installations of Dataverse software in Canada, including Scholars Portal Dataverse and other institutions and regions, to a national service called Borealis: Scholars Portal Dataverse first began offering the service outside of the Ontario Council of University Libraries in 2019, an official national service was offered in 2020 with agreements with the four regional academic library consortia, and the new brand Borealis was launched in 2022. The shared national installation also provides the opportunity for local branding and for providing shared training resources to users. During this transition, a Dataverse North expert group developed training resources, provided support and outreach, and developed promotion strategies. This is an important factor, as Canadian universities often prefer to store data on locally hosted servers.

In the Dataverse platform, data can be deposited into Dataverse collections that are part of a larger network. A Dataverse collection is a container for datasets (research data, code, documentation, and metadata) and can be set up for an individual researcher, department, journal, or organization. As an example, a researcher can deposit data into their institutional Dataverse collection, which is a part of the larger Borealis repository. Researchers and their collaborators can create their own accounts and deposit their data into an institutional collection (defined by their affiliation) or into research project collections, if available. Librarians and data stewards can also curate data contributions and handle data submissions on behalf of researchers. The Dataverse software is quite flexible in this regard. It is possible to apply institutional or project branding to Dataverse collections and sub-collections.

The Dataverse repository software also provides data analysis functionality in the browser; users do not need to download the data files in order to interact with them. The tabular data files that are uploaded to the system can be analyzed using the integrated web-based data analysis and visualization tool. Dataverse software can also be integrated with other library resources for improved discovery. For instance, since all partners of UBC Abacus Dataverse (libraries at the University of Victoria, University of Northern British Columbia, and Simon Fraser University) use ProQuest Summon as a discovery search engine, the Dataverse collections corresponding to their libraries are accessible through the specific Open Archives Initiative (OAI) protocol feeds. Each OAI feed includes all data from partner institutions and appropriate licensed data for that school. Through improved discovery (especially the assignment of DOIs for research datasets), curated data could be easily accessed and reused by researchers (e.g., in ORCID, Google, DataCite, Google Data Search, Crossref, and other services), thereby enhancing citations and improving research metrics for individuals and institutions.

Dataverse repository software has proven to be a flexible platform that can support many models for library RDM services in Canada. It offers a range of features that may improve data discoverability and access. It also provides excellent data management for preservation. However, Dataverse software is not a fully featured digital preservation system (although the national Borealis repository does support bit-level digital preservation, which is explained in the chapter, "Digital Preservation of Research Data," and in the Borealis Preservation Plan. The repository is format-agnostic and accepts all types of files, not just tabular data.

The Ontario Council of University Libraries sponsored work by Artefactual to develop a technical integration between the Dataverse software and Archivematica, a robust, open source tool for processing digital objects for preservation and access. This preservation processing tool could be used in conjunction with the established Borealis service or any Dataverse installation (with Archivematica version 1.8+ and Dataverse software version 4.8.6+).

Support for RDM in Canada has been a national focus. Historically and currently, regions and communities have faced issues related to support and infrastructure based on their own networks, regional or provincial funding and participation in consortium decisions by region.

Indigenous Data Sovereignty

Many of the initiatives and developments that we have mentioned in this chapter, and others that will be referenced throughout this textbook, have occurred without considering Indigenous Peoples and their data or redressing historical injustices. In fact, there has been a long history of mistreatment and neglect of Indigenous communities in Canadian research. While the Tri-Agency RDM Policy now explicitly addresses Indigenous data considerations, and Indigenous data experts are also included in the Sensitive Data Expert

Group, we encourage the Alliance RDM team to address these issues more comprehensively in the near future.

First 'Nations advocates and academics have responded to these gaps. For example, the First Nations Information Governance Centre (FNIGC) was incorporated as a nonprofit in 2010 to serve First Nations in data sovereignty, with work encompassing research, training, capacity building, and data collection. Their work dates back to 1996, when the Assembly of First Nations formed a National Steering Committee with the mandate of creating a national First Nations Health Survey (the First Nations Regional Longitudinal Health Survey), following Canada's decision to exclude the on-reserve population from major longitudinal data collection projects. In 1998, the committee established the principles of **OCAP®** (standing for ownership, control, access and possession) as a tool and standard for collecting and managing First Nations data. For more on OCAP® see the chapter, "Indigenous Data Sovereignty."

Regional Efforts

Across Canada, institutions have taken individual approaches on developing and expanding RDM services depending on their size, available resources (human resources and infrastructure), and research focus. College and university librarians and specialists are key members of the institutional RDM working groups and committees. They are involved in developing institutional RDM policies and strategies.

Many institutions across Canada have participated in surveys of RDM practices and needs that were based on a common survey instrument developed by librarians at the University of Toronto in 2015. The survey instrument was subsequently adapted with some modifications by many institutions across the country. This survey led to a richer understanding of disciplinary RDM practices and of local and national RDM needs, and it helped researchers become aware of RDM best practices (Cheung, et al., 2022) .

Courses on RDM have been taught at library schools across Canada. As described earlier, regions have adapted Dataverse repository software locally and, in many cases, nationally. All regions had representation on the Alliance RDM committees. Some schools responded to the need to provide RDM support with the development of RDM librarian positions or library roles. Below we describe regional initiatives highlighting unique services and areas of focus.

RDM in the Atlantic region

CAAL/CBPA (the Council of Atlantic Academic Libraries/Conseil des bibliothèques postsecondaires de l'Atlantique, formerly CAUL-CBUA) is the network of public university and college libraries in the Atlantic region. CAAL/CBPA has focused on building and coordinating digital preservation activities in the region.

The Digital Preservation and Stewardship Committee (DPSC) was formed in 2013. It later expanded its work on building and developing RDM services on a broad scale to align its work with the national vision. The most recent initiative involves the 2020 CAAL/CBPA Innovation Grant that enables a series of RDM workshops to be delivered and streamed across Atlantic institutions, with DPSC members taking the lead in organizing and conducting the workshops. The events are called Atlantic RDM days, and they are conducted in English and French. These workshops are important to colleges and universities that do not have the resources to support RDM at the institutional level but must still comply with the Tri-Agency RDM Policy and promote RDM best practices within their research community.

In 2015, Dalhousie University was one of the first Atlantic research institutions to start building an RDM team, which included many partners across the institution (Office of Research Services, Academic Technology Services, and Dalhousie University Libraries). Dalhousie University was one of the first Canadian institutions to develop and publish an RDM strategy, as required by the Tri-Agency RDM Policy. Dalhousie University now offers an RDM course entitled "Managing Research Data."

Several Atlantic research institutions have joined the national Borealis repository to provide data archiving services to their local research community. Others have agreed to maintain their own instances of the Dataverse repository installed on local institutional servers. This is due to the availability of local institutional resources to maintain and keep the repository up to date. For instance, since 2018, UNB Libraries at the University of New Brunswick have hosted a local Dataverse repository. This institutional data repository is hosted and maintained independently by the UNB Libraries through the collaborative work of the Library Systems team and the Libraries' RDM Services Committee. Like other Canadian institutions, all research universities in the Atlantic region have access to the national data archiving infrastructure, FRDR, available through the Alliance's website.

RDM in Quebec

Since the 1960s, academic libraries in Quebec have collaborated under le Bureau de la Coopération Interuniversitaire (BCI), formerly known as la Conférence des Recteurs et des Principaux des Universités du Québec (CREPUQ). In 1967, le Comité de coordination des bibliothèques was created. A few years later, it became le Sous-comité des bibliothèques (Roy et Bégin, 1969).

Dataverse internationalization took place in two phases: the first phase began in 2015, and the second phase began a few years later (Bilodeau, 2018). Marie-Hélène Vézina, a senior librarian from l'Université de Montréal with experience in digital project development, teamed with Scholars Portal staff, with support from the broader Dataverse community, including Harvard's Institute for Quantitative Social Science, to internationalize Dataverse software. Although some translation work had been done in the past, nothing had been done to support multilingualism. The developed code became part of the central Dataverse software

codebase, which allowed a bilingual (French and English) installation to be deployed by Scholars Portal. L'Université de Montréal contributed the French translation. The Scholars Portal and BCI institutions finalized and signed a formal agreement in spring 2019, and the first institutional Dataverse collections from BCI institutions were made available to researchers in summer 2022 (Vézina, 2022).

At l'Université de Montréal, the first dedicated RDM librarian position was established. Soon after, a second RDM librarian position was opened at l'Université Laval. McGill University set up an RDM research support position, and three smaller institutions shared an RDM research support position, namely l'Institut national de la recherche scientifique (INRS), l'École nationale d'administration publique (ENAP), and la Télé-université, Université du Québec (TÉLUQ).

Other institutions have allocated part-time resources to RDM. Institutional Dataverse collections are being launched in Borealis. The focus will likely be on keeping pace with growing needs in the years to come.

RDM in Ontario

In Ontario, there are 23 public universities and 24 colleges. Since the 1960s, the libraries at these universities have been successfully collaborating through the Ontario Council of University Libraries (OCUL). In its early years, OCUL was involved in traditional library services, such as consortial licensing of journals and facilitating effective resource sharing. In those early years, several institutions developed their own data repository systems, including Carleton University's Social Science Data Archive, founded in 1965 in the Sociology and Anthropology Department; Western University's Data Resources Library, launched in the late 1970s, which worked with the Social Science Computing Laboratory to disseminate and archive several faculty research projects; and the University of Toronto's Map and Data Library, established in 1988, with services that included the acquisition and preservation of datasets produced by the University of Toronto researchers.

In 2002, OCUL formed Scholars Portal, a shared technology infrastructure that hosts and provides access to OCUL's growing digital collections. As data services came to greater prominence, Ontario libraries saw an opportunity to collaborate under the OCUL umbrella in order to improve services, reduce duplication of effort, and better manage limited resources. Over the last decade, OCUL has undertaken several successful data infrastructure projects, including the development of the collaborative <odesi>, a social science data portal, and Scholars GeoPortal, a geospatial data portal. While both of these data portals do contain some research data, they are intended as curated collections of published datasets from authoritative sources, such as government statistical agencies. As such, they are not conducive to the widespread inclusion of member libraries' institutional research data outputs. These systems are primarily focused on discovery and access rather than long-term data preservation (Moon, 2014).

For this reason, other solutions were needed in Ontario as well as in Canada to address the growing demand for library research data repositories. In 2011, Scholars Portal joined the UBC Library pilot and installed a Dataverse repository, an open source software and offered it to the OCUL community as a pilot program. The pilot was intended to address a community-identified need for an Ontario-based repository service that would allow for easy-to-use, web-based self-deposit by researchers. Dataverse software was chosen for the pilot due to its support for research data, including the **Data Documentation Initiative (DDI)** built-in metadata. Scholars Portal staff developed documentation and training materials to inform and train staff at OCUL libraries about the benefits of incorporating Dataverse software into the suite of services offered for data management and deposit of research data. As a result, the Scholars Portal Dataverse repository, now branded Borealis, has allowed some OCUL libraries to launch RDM services without needing to have the technical infrastructure and staffing to support repositories of their own. Models for the service vary from library to library, ranging from self-serve deposit to library-mediated curation. Today, the service has grown dramatically. Many more institutions across Canada have joined or migrated their research data content to Borealis, making it a national hub for research data archiving. The support for the use of Borealis is largely provided by local library staff and is independent of the infrastructure hosted and supported by Scholars Portal.

The OCUL data community, which was initially formed to address data access for Statistics Canada DLI data, has evolved into a forum for support of RDM. Experts from Ontario academic institutions have been key members of the Alliance RDM community and working groups.

RDM in the Prairie Provinces

Institutions in the Prairie provinces have been very influential in the national RDM collaborations over the last decade. In early 2015, the University of Alberta Libraries implemented the first Canadian instance of an open source online tool to help researchers write DMPs. A UK-based DMPOnline code was used at that time, and UBC and the University of Alberta were the first Canadian institutions to adapt the Canadian version.

Almost immediately, the project was adapted by other Canadian institutions within the CARL Portage framework and was branded as DMP Builder. Later in the tool's lifecycle, it was rebranded again and became DMP Assistant, which included English and French options to better serve the francophone academic community. Over 50 Canadian institutions now use DMP Assistant with custom institutionally specific guidelines developed by the Alliance RDM NOE. It has been almost a decade since the University of Alberta Libraries sponsored DMP Assistant for the Canadian RDM community, who greatly appreciate their work.

Since late 2015, the University of Saskatchewan (USask) Research Computing has been implementing a similar initiative in partnership with the Office of the Vice-President Research. As a result of Compute

Canada's seed funding, the USask team was chosen to create a national data discovery interface for research data in Canada. The USask-based team is still chiefly responsible for the software development and operation of the Lunaris platform, now under the Alliance umbrella. They adapted the open source code base from the UBC Library Open Collections as their main discovery interface back in 2016 and the Geodisy open source code base, also developed by the UBC Library, as their map-based data discovery interface. Using the open source Archivematica software, the USask-based team has also developed an excellent collaboration with the Globus Connect platform to work with big data and preserve research data digitally at scale.

RDM in British Columbia

British Columbia institutions have long been engaged with RDM, with the University of British Columbia (UBC), Simon Fraser University (SFU), and the University of Victoria (UVic) taking the lead in this work. The UBC Library is one of the largest university libraries in Canada and has been conducting ad hoc RDM activities since the early 1970s. In 2008, to help smaller regional schools, UBC entered into an arrangement to make the Abacus data repository available to other universities in the province. At the time of writing, four major university research libraries in British Columbia (Simon Fraser University, University of Victoria, University of Northern British Columbia, and the University of British Columbia) are using the UBC instance of the Dataverse repository as a licensed data repository.

Data is provided to users from each institution according to their data licenses using the Canadian Access Federation, an organization that manages digital identities in higher education and research through a trust framework for access control. The UBC Library data team provides basic and advanced training on the Dataverse repository to groups, departments, and labs on the UBC campus and to its partners in other university libraries and research institutions. After the training, these groups should be managing their own data within the appropriate Dataverse assigned to them. UBC Library School (now known as iSchool) was also one of the earliest Canadian institutions to offer a Research Data Management graduate course.

The SFU and UVic Libraries have also contributed greatly to the RDM landscape in Canada. Early in the 2010s, the SFU Library developed Radar, its own Islandora-based research data repository (now depreciated and replaced by FRDR), and it became the Canadian leader on zero-knowledge encryption of sensitive data. The UVic Libraries have also successfully experimented with RDM services and have accommodated unique license needs for research teams, such as, for example, the well-used Canada Health Infoway datasets.

RDM in Northern Canada

Northern Canada consists of three territories: the Northwest Territories, Nunavut, and Yukon. The two research institutions located in Northern Canada are Yukon University and Aurora College. As part of the

institutional RDM strategy (mandated by the Tri-Agency RDM Policy), Yukon University librarians and the Research Services Office work together to build an institutional repository hosted by BC ELN Arca – a collaborative initiative for digital repositories in BC based on Islandora software, primarily aimed at smaller institutions and colleges. Research outputs deposited by Yukon University researchers into BC ELN Arca will be harvested by Lunaris.

In October 2022, librarians from Aurora College in Inuvik participated in an institutional strategy panel organized by the Alliance. They shared their unique experience addressing RDM issues at the small-size Northern institution. Some institutions from Northern Canada, including Yukon University, work together with universities and colleges from British Columbia to develop their institutional RDM strategies in line with the Tri-Agency RDM Policy. They collaborate as an ad hoc group to create action plans and share visions for RDM services in small institutions.

Conclusion

It is an exciting time for RDM in Canada, and it took years of dedicated work and sophisticated, multi-provincial collaboration to get to this point. Libraries are seeing new opportunities to engage with their communities and with one another. With these new opportunities inevitably come challenges, such as costly digital infrastructure that must be managed on an ongoing basis. We believe that Portage and the formation of the Alliance have the greatest potential to meet some significant unmet needs, but they will need sustainable funding in order to be successful.

The development of open source tools, infrastructure, and support services for RDM is crucial if Canadian scholars are to successfully integrate these new activities into their workflows. Academic libraries have a history of supporting data access, dissemination, and preservation, and they have an established mandate to participate in the preservation of the research outputs of their community (e.g., in institutional repositories). Libraries can provide leadership in the adoption of best practices and open standards. They can also partner with other groups in the development of infrastructure and tools. The Canadian library community has been actively encouraging research data sharing since the 1960s and is well-positioned to play a leadership role going forward.

Reflective Questions

1. What new knowledge have you gained about the Canadian data community?
2. How do you think the Canadian academic library data community compares to other areas of academic librarianship?
3. Given the current international open science movement, what challenges do you see in research data management today?
4. Which parties or organizations are best positioned to provide RDM support to Canadian researchers?

Key Takeaways

- The development of data services, awareness, infrastructures, tools, and RDM culture in general has evolved over several decades locally, regionally, and nationally.
- Data librarians, data specialists, library consortia, government funding agencies, and governance bodies play key roles in identifying needs and developing services in RDM.
- To promote best data management practices in support of RDM services, government, institutions, service providers, and the research community need to continue to partner at every stage of the research lifecycle.
- A number of tools and technical infrastructures are available to support RDM, and these will evolve to support ongoing and new needs.

Acknowledgement

The initial draft of the RDM in Quebec section was contributed by Ève Paquette-Bigras, academic librarian at the Université de Montréal. The authors are grateful to Ève for providing the background and context for the RDM achievements in that province.

Additional Readings and Resources

Doiron, J., Neilson, M., & Nicholson, R. (2020). *Data management planning in Canada*. White paper for NDRIO. https://alliancecan.ca/en/document/261

Government of Canada. (2021). *Tri-Agency research data management policy.* https://www.science.gc.ca/eic/site/063.nsf/eng/h_97610.html

Lavoie, B., & Dempsey, L. (2004). Thirteen ways of looking at...digital preservation. *D-Lib Magazine*, *10*(7–8). https://doi.org/10.1045/july2004-lavoie

Moon, J. (2021). Update on Portage and the Digital Research Alliance of Canada Alliance. https://alliancecan.ca/en/latest/news/update-portage-and-digital-research-alliance-canada

Read, K., McDonald, G., Mackay, B., & Barsky, E. (2014). A commitment to First Nations data governance: A primer for health librarians. *Journal of the Canadian Health Libraries Association / Journal de l'Association des bibliothèques de la santé du Canada*, *35*(1), 11–15. https://doi.org/10.5596/c14-003

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J-. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). Comment: The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*, 1–9. https://doi.org/10.1038/sdata.2016.18

Reference List

Attendees of the NDSF Summit. (2019). *Kanata declaration (Version 2.0).* National Data Services Framework Summit 2019 (NDSF 2019), Ottawa, Canada. Zenodo. https://doi.org/10.5281/zenodo.3234815

Barsky, E., Laliberté L., Leahey, A., and Trimble, L. (2017). Collaborative research data curation services: A view from Canada. In L.R. Johnston (Ed.), *Curating research data, volume one: Practical strategies for your digital repository*. Association of College and Research Libraries. https://dx.doi.org/10.14288/1.0340778

Bilodeau, G. (2018). *Gestion des données de recherche (GDR): Écosystème Canadien – un bref survol*. https://www.ulaval.ca/sites/default/files/recherche-creation/documents/conduite%20responsable/gestion-donnees-recherche-bilodeau.pdf

Boyko, E., & Watkins, W. (2011). The Canadian data liberation initiative: An idea worth considering. *International Household Survey Network*, *IHSN Working Paper* (006). www.ihsn.org/sites/default/files/resources/IHSN-WP006.pdf

Cheung, M., Cooper, A., Dearborn, D., Hill, E., Johnson, E., Mitchell, M., & Thompson, K. (2022). Practices before policy: Research data management behaviours in Canada. *Partnership: The Canadian Journal of Library and Information Practice and Research*, *17*(1), 1-80. https://doi.org/10.21083/partnership.v17i1.6779

Cooper, A., Steeleworthy, M., Paquette-Bigras, È., Clary, E., MacPherson, E., Gillis, L., Wilson, L., & Brodeur, J. (2021). *Dataverse curation guide*. Zenodo. https://doi.org/10.5281/zenodo.5579820

Gray, S. V., & Hill, E. (2016). *The academic data librarian profession in Canada: History and future directions.* Western Libraries Publications. Paper 49. http://ir.lib.uwo.ca/wlpub/49

Hackett, Y. (2001). A national research data management strategy for Canada: The work of the National Data Archive Consultation Working Group. *IASSIST Quarterly, 25*(3), 13-16. https://doi.org/10.29173/iq91

Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends*, *57*(2), 280–299. https://doi.org/10.1353/lib.0.0036

Humphrey, C. (2005). Collaborative training in statistical and data library services: Lessons from the Canadian data liberation initiative. *Resource Sharing and Information Networks*, *18*(1–2), 167–181. https://doi.org/10.1300/J121v18n01_13

Humphrey, C. (2012a, December 5). Canada's long tale of data. *Preserving Research Data in Canada.* http://preservingresearchdataincanada.net/2012/12/05/hello-world

Humphrey, C. (2012b, December 13). Research data management infrastructure II. *Preserving research data in Canada*. https://preservingresearchdataincanada.net/2012/12/13/research-data-management-infrastructure-ii/

Humphrey, C. (2020). The CARL Portage partnership story. *Partnership: The Canadian Journal of Library and Information Practice and Research*, 15(1), 1–7. https://doi.org/10.21083/partnership.v15i1.5825

Humphrey, C., Shearer, K., & Whitehead, M. (2016). Towards a collaborative national research data management network. *International Journal of Digital Curation*, *11*(1), 195–207. https://doi.org/10.2218/ijdc.v11i1.411

Liss, S. N. (2018, September 5). Addressing gaps in Canadian research data management: A comprehensive guide of the Portage Network. *University Affairs*. https://www.universityaffairs.ca/magazine/sponsored-content/addressing-gaps-in-canadian-research-data-management/

Moon, J. (2014). Developing a research data management service – a case study. *Partnership: The Canadian Journal of Library and Information Practice and Research*, *9*(1), 1–14. https://doi.org/10.21083/partnership.v9i1.2988

Roy, J. et Bégin, J.-O. (1969). *Enquête relative à un plan de coordination*. Montréal : Comité de coordination des bibliothèques de la CREPUQ.

Steeleworthy, M. (2014). Research data management and the Canadian academic library: An organizational consideration of data management and data stewardship. *Partnership: The Canadian Journal of Library and Information Practice and Research*, *9*(1), 1–11. https://doi.org/10.21083/partnership.v9i1.2990

Strong, D. F., & Leach, P. B. (2005). *National consultation on access to scientific research data*. Final Report. Government of Canada. https://publications.gc.ca/site/eng/272526/publication.html

Vézina, M.-H. (2022). *Métadonnées bibliographiques des thèses et mémoires du dépôt institutionnel de l'Université de Montréal [Canada]* [dataset]. Borealis. https://doi.org/10.5683/SP3/SJJACL

About the authors

Eugene Barsky

Eugene Barsky is the Research Data Librarian at the University of British Columbia. Eugene works with the UBC researchers curating and managing research data, from planning to deposit to preservation. Eugene participates in building the Canadian Federated Research Data Repository service (FRDR), and he collaborates with Digital Research Alliance of Canada (the Alliance) and the European Union (OpenAIRE). He is the PI for the national Geodisy project funded by the Alliance. His recent peer-recognition includes the Canadian Association of Research Libraries, American Society for Engineering Education, and Special Library Association awards. He had published more than 30 peer-reviewed papers and presented at more than

70 conferences. Eugene is an adjunct professor at the iSchool at UBC where he teaches research data management, and is one of the founders of the Portage Network of Experts in Canada. Email: eugene.barsky@ubc.ca | ORCID: 0000-0002-5119-2271

Elizabeth Hill

Elizabeth Hill is the Data Librarian at Western University in London Ontario. She provides access and data literacy instruction to data sources at Western. She has an external advisor role with Statistics Canada. Elizabeth is active in various data communities and working groups in participant and leadership roles. Her research interests include supporting researchers, and she has published on topics related to data delivery systems and data librarianship in Canada. ORCID: 0000-0002-9715-238X

Tatiana Zaraiskaya

Tatiana Zaraiskaya is a STEM Librarian at the University of New Brunswick Libraries, where she is also responsible for RDM. Tatiana has been a member of the RIEG Portage Network (The Alliance) team since 2016, has participated in several surveys by RIEG, and was one of the leaders of the RDM Survey of Queen's University and the UNB. She is a co-author of multiple conferences and other scholarly publications related to RDM and an author of the DMP Portage template. Tatiana holds a PhD in Biophysics from the University of Guelph and an MLIS from Western Ontario University. Email: t.zaraiskaya@unb.ca | Google Scholar: https://scholar.google.com/citations?user=BB6c8XQAAAAJ&hl=en | ORCID: 0000-0001-9294-6052

Minglu Wang

Minglu Wang is a Research Data Management (RDM) Librarian at York University. She has published book chapters, conference/working papers, and research articles closely related to academic libraries and RDM services. Minglu Wang is an active member of the Association of College & Research Libraries (ACRL), a division of the American Library Association (ALA), and she contributed to multiple years of the Association's publications of Top Trends articles and Environmental Scan white papers. She is a member of the Research Intelligence Expert Group, a part of The Digital Research Alliance of Canada (The Alliance) RDM Team, and has participated in the design and report writing of the RDM Capacity Survey of Canadian Institutions. Email: mingluwa@yorku.ca | ORCID: 0000-0002-0021-5605

Lucia Costanzo

Lucia Costanzo is the Research Data Management (RDM) Librarian at the University of Guelph. She recently completed a secondment at the Digital Research Alliance of Canada (the Alliance) as the Research

Intelligence and Assessment Coordinator. As part of this role, Lucia coordinated the activities of the Research Intelligence Expert Group, which included informing and advising the Alliance RDM Team and Alliance management on emerging developments and directions, both nationally and internationally, in RDM and broader Digital Research Infrastructure ecosystems. Before the secondment, Lucia actively supported, enabled, and contributed to the learning and research process on campus for over twenty years at the University of Guelph. Email: lcostanz@uoguelph.ca | ORCID: 0000-0003-4785-660X

5.

RESEARCH DATA SHARING AND REUSE IN CANADA: PRACTICE AND POLICY

Meghan Goodchild; Shahira Khair; Amber Leahey; Kaitlin Newson; and Lee Wilson

Learning Outcomes

By the end of this chapter you should be able to:

1. Understand practices, policies, and services guiding research data sharing and reuse in Canada.
2. Identify elements of Canadian digital research infrastructure, including data storage options like data repositories and long-term preservation platforms, as well as services that support access and use of these infrastructures.
3. Using case studies, identify supports and barriers to data sharing and reuse throughout the research data lifecycle, along with areas in need of future development.

Introduction

Canadian researchers from all disciplines and sectors are producing unprecedented amounts of data (Baker et al., 2019). With the advancement of open science and open data policies from journal publishers, research funders, disciplinary groups, and institutions, researchers are becoming more acutely aware of the need to manage their data in accordance with related policies on data deposit and sharing. These policies support broader goals around research transparency, reproducibility, and reuse (the Alliance Research Data Management Working Group [Alliance RDM WG], 2020). (See chapter 12, "Planning for Open Science Workflows," for an overview of open science and open data.)

A major value proposition for data sharing and reuse is the acceleration of scientific progress and prevention of unnecessary expensive data collection. Data sharing also enables research results to be reproduced, which can improve the integrity and trustworthiness of published findings. When a researcher's data is easy to discover and access, this increases the visibility and impact of their research. Additionally, sharing data, research environments, and tools enables and enhances collaboration, leading to greater **interoperability** and research efficiencies.

In order to maximize the benefits of data sharing and reuse, research data outputs should be guided by the **FAIR principles** of Findability, Accessibility, Interoperability, and Reusability, discussed in chapter 2, (Wilkinson et al., 2016), and supported by a foundation of a TRUST-ed (Transparency, Responsibility, User focus, Sustainability and Technology) digital research infrastructure and support services (Lin et al., 2020). Therefore, data sharing is an integral component of conducting high-quality research, requiring ongoing **Research Data Management (RDM)** practices. RDM services in Canada are emerging across disciplines, institutionally, and at the regional and national levels to support researchers in data sharing and reuse.

In this chapter, you'll learn about policies and practices, digital research infrastructure, and tools and services for research data sharing and reuse in Canada. We'll review policies and practices, Canadian infrastructure, tools, and services that support the **research data lifecycle**; and support services around data curation and preservation. Then we'll consider case studies that highlight data sharing and reuse practices and highlight disciplinary challenges.

Policies and Practices in Canada

Research Funding Agencies

Funding agencies and governments around the world have recognized the need for national RDM policies to support access to publicly funded data. Funding agency mandates that require data sharing influence researcher behaviour and the demand for RDM infrastructure and services (the Alliance RDM WG, 2020). The Canadian Tri-Agency RDM Policy (2021) is driving a culture change for data deposit and sharing, as it outlines requirements for researchers to "deposit into a digital repository all research data, metadata and code that directly support the research conclusions in journal publications and **pre-prints** that arise from agency-supported research" (Government of Canada, 2021), implementation forthcoming. Grant recipients are expected to provide access to their data in accordance with the FAIR principles and disciplinary standards while respecting ethical, cultural, legal, and commercial requirements. Indigenous data sovereignty (discussed in depth in chapter 3) recognizes the inherent rights of Indigenous communities to govern the collection, ownership, and use of their data and may result in distinct practices regarding the sharing of research data.

Funder Policies

Local and Regional

Canadian research institutions may set their own requirements for data management and sharing, according to internal policies governing research practices and intellectual property. They must also publicly post a strategy indicating how RDM practices will be supported (Government of Canada, 2022).

National

- Tri-Agency RDM Policy (2021)
  - Grant applicants must include a **Data Management Plan** for certain funding applications (phased implementation beginning in spring 2022)
  - Grant recipients should deposit into a digital repository all **research data**, **metadata**, and code that directly support the research conclusions in journal publications and pre-prints that arise from agency-supported research. Deposit will be expected at the time of publication (implementation forthcoming).
  - Although sharing data is not required, the Agencies expect researchers to provide appropriate access to the data where ethical, cultural, legal, and commercial requirements allow and in accordance with the FAIR principles and the standards of their disciplines. Whenever possible, these data, metadata, and code should be linked to the publication with a **persistent identifier (PID)**.
- Tri-Agency Statement of Principles on Digital Data Management (2016)
  - Data should be collected and stored using software and formats that ensure secure storage, preservation of, and access to the data beyond the duration of the research project.
- Tri-Agency Open Access Policy on Publications (2015)
  - Researchers funded by the Canadian Institute of Health Research (CIHR) should deposit specific types of data (e.g., bioinformatics) into an appropriate public database.
- SSHRC Research Data Archiving Policy (1990)
  - Research data must be preserved and made available for use within two years of project completion (Government of Canada, 2016).

International

Many public research funders in other countries that support Canadian research require researchers to share underlying datasets of published research, including:

- U.S. funders including National Institutes of Health (NIH) and National Science Foundation (NSF)
- UK Research and Innovation funders
- European Commission Horizon 2020

Several private research funding sources have their own data sharing expectations (e.g., Wellcome Trust, Bill & Melinda Gates Foundation).

Other Policies and Practices

Journal publishers have also been driving the adoption of RDM practices. When they require a data availability statement, research data is more likely to be shared online. When policies are less stringent, such as recommending data archiving, there's only a slight increase in archiving rates over having no policy (Vines et al., 2013). Data sharing and availability differ by discipline. For example, the fields of biological science, earth science, medical science, and physical science have a higher rate of data sharing (Stuart et al., 2018), whereas data were less readily available in materials for energy and catalysis, psychology, optics and photonics, and forestry (Tedersoo et al., 2021).

Over the past 20 years, data sharing rates have improved (Tedersoo et al., 2021), but studies show that results are not always fully reproducible from shared data, often because of inadequate documentation and metadata (Rieseberg et al., 2021). There have been increasing efforts to mitigate this issue. For example, the *Journal of Molecular Ecology* encourages authors to use the **open access** database GEOME to establish permanent links between genetic data and geographic and ecological metadata to make the data deposited FAIR (Rieseberg et al., 2021). The Public Library of Science (2022) announced an "Accessible Data" pilot feature for certain articles to emphasize links to datasets in specific repositories, in order to increase sharing and discovery of research data and to highlight the benefits of open science models. The *American Journal of Political Science*, in partnership with the Odum Institute for Research in Social Science, provides data curation and verification services to ensure that datasets reproduce the results of corresponding articles (Jacoby et al., 2017). Therefore, policies alone are not sufficient but may require technical and discipline-specific solutions to ensure that the data shared can be accessible and reusable.

Infrastructure, Tools, and Services

A range of infrastructures are necessary to support the production, sharing, and reuse of data over its lifecycle. These work together to make data FAIR beyond the timespan of a research project.

There are three types of research data storage: **active**, **repository**, and **archival**. Figure 1 outlines active storage during the research phase, repository storage during the access and publishing phase, and archival storage during the preservation phase, which requires additional processing to support long-term accessibility.

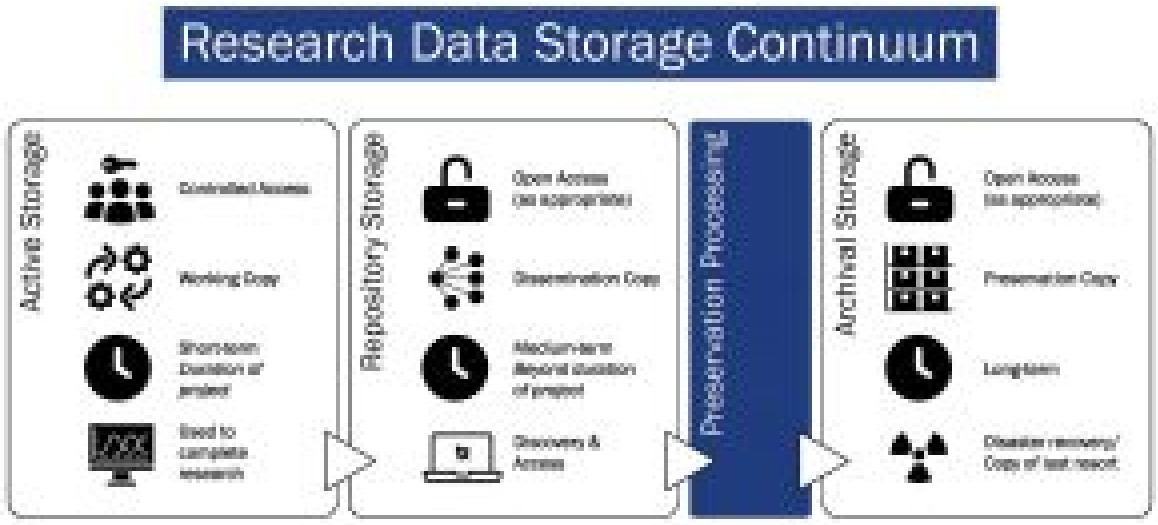


Figure 1. *Research data storage spectrum (the Alliance RDM WG, 2020). © All Rights Reserved, reused with permission.*

Table 1 details active, repository, and archival storage and provides examples used in Canada. Table 2 outlines different research infrastructures that facilitate sharing, reuse, and access.

Table 1: ***Types of research data storage.***

Type	Attributes	Examples
Active Storage	• supports data that need to change or be acted on frequently, from constant (every second) to periodic (every week)	• computing and analysis storage (e.g., regional and national high-performance computing) • institutional enterprise and personal storage (e.g., hard drives) • commercial cloud storage (e.g., Microsoft Azure, OneDrive, Google Cloud, Amazon Web Services) • hosted project file storage (e.g., Open Science Framework, Code Ocean)
Repository Storage	• supports stewardship and maintenance of data and metadata or other objects, including code, that represent the authoritative	• repository platforms (e.g., CKAN, InvenioRDM, The Dataverse Project, HUBzero) • hosted services (e.g., GitHub, Zenodo,

Type	Attributes	Examples
	copy in the scholarly record • main functions include ingestion, curation, retention, and access (the Alliance RDM WG, 2020) • access typically mediated by software platforms, including portals or research gateways	Federated Research Data Repository (FRDR), Borealis, institutional or disciplinary repositories)
Archival Storage	• supports long-term preservation; may not be the primary access point for reusability but can be relied upon for access and reuse • regional library associations may offer this infrastructure to member institutions	• institutional archival storage • storage used by academic library services (e.g., Ontario Council of University Libraries' (OCUL) Ontario Library Research Cloud (OLRC), offered nationally; and the Council of Prairie and Pacific University Libraries' (COPPUL) WestVault)

Table 2: ***Research data infrastructures in Canada.***

Type	Attributes	Examples
Multi-Disciplinary Repositories	• use of disciplinary repositories encouraged when available. • when not available, may use institutional or generalist repositories that can accommodate multiple file types and use cases	• see Table 3 for Canadian repositories • international platforms and hosted services (e.g., Mendeley Data, Figshare, Dryad, Zenodo, Harvard Dataverse Repository)
Disciplinary Repositories and Infrastructures	• focus on specific types of data (e.g., genomics) and may use specialized standards • may act as a knowledge base, providing curation, extraction, organization, annotation, and linking to bodies of literature or data • may be project-specific portals to collect and share research data for knowledge exchange and mobilization purposes; may include links to repositories or alternative storage options	• see Table 3 for Canadian repositories • large-scale research projects, including Linked Infrastructure for Networked Cultural Scholarship (LINCS), Ocean Networks Canada, Genome Canada, Data Access Support Hub (DASH), Linked Parliamentary Data Project (LiPaD)
Preservation Services and Tools	• support long-term care and preservation of digital objects of research value	• Archivematica — repository integrations (e.g., FRDR, Borealis) • consortial preservation services (e.g.,

Type	Attributes	Examples
	• use specialized software to prepare research data for long-term **integrity checking** preservation using techniques such as file **normalization**, integrity checking, and **data packaging**	COPPUL's Archivematica-as-a-service, OCUL's Permafrost service) • national preservation software (e.g., DuraCloud, hosted to support digital preservation for OLRC subscribers)
Research Data and Software Replication Services and Tools	• allow others to display, manipulate, and interpret data to support reuse and reproducibility • used so others can replicate the data (e.g., for collection, analysis, visualization)	• software and code replication platforms and services (e.g., Code Ocean, Syzygy, Jupyter Hub, GitHub) • tools that facilitate reproducible code and computing environments (e.g., Jupyter Notebooks, Docker)
Data Discovery Services	• connect metadata and data using a common schema, format, and structure to help researchers discover and reuse data • improve discovery across repositories with varying standards and levels of interoperability	• international and Canadian research data search services (e.g., Lunaris, Open Data Canada, OpenAIRE, Google Dataset Search, DataCite Commons, Data Citation Index) • domain-specific services (e.g., iReceptor Commons, Global Biodiversity Information Facility, Canadian Open Neuroscience Platform)
Interoperability and Standards	• support one of four types of interoperability: technical, semantic, organizational, and legal (Corcho et al., 2021).	• PIDs (e.g., Digital Object Identifiers for data, ORCID iDs for researchers, ROR for organizations, RAiD for research projects) • metadata standards (e.g., Dublin Core, Data Documentation Initiative, DataCite Schema, Data Catalog Vocabulary) • ontologies and subject classification (e.g., Canadian Subject Headings, ISO standards, W3C vocabularies) • data licensing (e.g., Creative Commons, Open Government Licence) • software licensing (e.g., MIT, GNU, Apache) • open protocol and exchange standards (e.g., OAI-PMH, SWORD)

Canadian Data Repositories

Data repositories are key to research infrastructure in Canada. National and institutional data repositories are emerging to support researchers with deposit, sharing, and long-term preservation of data to provide open, equitable, and connected RDM services, circumventing expanding commercial interests and reducing reliance on customized solutions, such as research project websites that often require maintenance and resources for the long term. Through federal, provincial, and institutional funding, Canadian repositories are available to researchers at no additional cost and may have a longer lifespan than the research project. Table 3 outlines types of data repositories in Canada, many of which can be discovered through global registries such as the Registry of Research Data Repositories (re3data), FAIRSharing, and OpenDOAR.

Table 3: ***Data repositories in Canada.***

Type	Attributes	Examples
Multi-Disciplinary Repositories	• support data across disciplines • may provide curation services • may aggregate data across datasets	• institutional (e.g., UNB Dataverse, University of Prince Edward Island's Data Repository) • national (e.g., Borealis, FRDR)
Disciplinary Repositories	• support data related to specific disciplines • may provide curation services • may aggregate data across datasets	• disciplinary (e.g., Polar Data Catalogue, Barcode of Life Data System, Canadian Integrated Ocean Observing System)
Government Repositories	• for data collected or compiled by government departments • domain focused (i.e., not generic open data sites)	• BC Data Conservation Centre • World Ozone and Ultraviolet Radiation Data Centre • National Climate Data Archive • NRCan Earth Observation Data Management System

Type	Attributes	Examples
Knowledge Bases	• extract, gather, and curate data in a subject area • rely on core datasets to link together a growing body of information	• Avibase • DrugBank • BioGRID
Academic Data Repositories	• developed and/or supported by universities to host licensed and open data collections • may also include government data	• library data services (e.g., Odesi, Abacus Data Network, Scholars GeoPortal, Geoindex)

Support Services

To produce datasets with a high potential for reuse, researchers must use good curation practices as data are cleaned, documented, interrelated, stored, and shared. A range of services provide support for researchers in developing these RDM practices and are detailed in Table 4.

The 2021 Researcher Needs Assessment survey conducted by the Digital Research Alliance of Canada (the Alliance) found researchers have varying levels of access to and awareness of support for research workflows, at local, provincial, and national levels, with the greatest access at the local level (Pérez-Jvostov et al., 2021).

- Internal supports: The first point of support for many researchers is internal to their own research groups. For instance, many research groups employ data managers to support team members with data management and publication. Researchers usually discover and select tools and services based on peer reference (Pérez-Jvostov et al., 2021).
- Higher education institutions: These provide formal services and support through offices of research, academic libraries, and research computing services (Pérez-Jvostov et al., 2021). The Tri-Agency's requirement for institutional RDM strategies will help to coalesce cross-campus support.
- Shared support models: These can improve efficiency while meeting the demands of researchers and increasing access and equity. They are often coordinated by regional or national consortia. Case Study 1 is an example of a national community of practice as a support network for institutional repository administrators.
- Disciplinary services and supports: These serve the needs of specific research communities and are often

advanced nationally and internationally through research organizations and publishers. They are vital for the adoption of standard practices and tools in related disciplines since they are tailored to specific research workflows.

Table 4: ***Support services in Canada.***

Category	Services
Data Management Planning (DMP)	The Alliance supports the infrastructure and oversees the development of the DMP Assistant tool. Academic libraries and offices of research work together to support local researchers in developing DMPs in compliance with the Tri-Agency's RDM Policy.
Data Discovery and Access	Academic libraries support data discovery and access through reference services and database licensing. Some of these services are shared among institutions (e.g., Odesi, Abacus Dataverse repository). National and provincial organizations enable access and use of population data for research. Due to the sensitive subject matter, support often requires entering into an agreement with the service provider (e.g., CRDCN and StatCan Data Centres, ICES, Population Data BC). The Alliance supports a national discovery service Lunaris to increase exposure to Canadian data repositories and datasets. Exploratory work supports access to common datasets on high-performance computing infrastructures (e.g., bioinformatics datasets).
Computing and Storage	Local research computing and IT departments may offer services to researchers to support data management computing and active storage infrastructure. Increasing support for data management on active storage is important for the Alliance and its national federation of partners, and researchers can receive support through the Alliance's national help desk.

Category	Services
Data Curation and Publication	A range of workflows and related guidance have been developed to assist curators, including: • Dataverse Curation Guide • Data Curation Network CURATE(D) Model and Curation Primers • DCC guides and checklists • Canadian Data Curation Commons (beta) To support open publishing, academic libraries provide curation support to researchers depositing data and other scholarly objects to institutional repositories or other digital asset management systems. Borealis enables local services through a distributed support model, where infrastructure is centrally hosted, but researchers receive support for curation from a local administrator (see Case Study 1 below). The Alliance provides curation support to researchers using the nationally accessible FRDR. They also support researchers in developing and deploying research gateways on advanced research computing infrastructure. Other repositories act as trusted resources for stewarding research data and provide services supporting their platforms (e.g., Canadian Astronomy Data Centre, Ocean Networks Canada, Polar Data Catalogue). Commercial publishers, including Springer Nature and Elsevier, offer services supporting dataset curation and publication. Others have partnerships with third-party repositories to support authors in publishing datasets underlying their publications (e.g., the partnership between Wiley and Dryad).
Training	Researchers receive training from services developed within their disciplinary communities and institutions (Pérez-Jvostov et al., 2021), often led by peer researchers and support specialists who act as "**data stewards**" who develop activities that advance awareness, understanding, development, and adoption of RDM tools, best practices, and resources. Key events in Canada include: • workshops and summer bootcamps • Train the Trainer courses and resources • online training modules
Emerging Service Areas	Services supporting data sharing and reuse in Canada are developing in response to the needs of researchers related to RDM. Emerging service areas include support for the following: • **digital preservation** (see more in chapter 11) • sensitive data curation (see more in chapter 13) • research software curation • Indigenous data sovereignty

Case Study 1: Developing a National Dataverse Repository Service and Community in Canada

Background

The Dataverse Project is open-source research data repository software that allows users to share, cite, explore, and analyze research data. It is developed at the Institute for Quantitative Social Science at Harvard University, with collaborators from all over the world. Borealis, the Canadian Dataverse Repository, is based on Dataverse software and began as a regional research data repository for the Ontario Council of University Libraries, growing over the past ten years into a national, bilingual service with over 60 institutions subscribing. The infrastructure is hosted at the University of Toronto, with data files stored securely on the Ontario Library Research Cloud. Borealis offers one repository option for researchers who may not have a disciplinary repository available to them and who may benefit from the flexible data sharing features (e.g., open to restricted), exploration tools in the browser, and preservation-friendly actions and storage.

Analysis

Although Borealis is centrally hosted, academic libraries and institutions manage their collections, thereby supporting local researchers in depositing and sharing datasets. Since local capacity varies across institutions and regions (Goddard et al., 2018), cultivating a community of practice is crucial for capacity building across institutions and for collaborative development of resources and training materials to support researchers. Alongside efforts to develop technical infrastructure, the Borealis team has been collaborating with the Alliance's Dataverse North Expert Groupon community-building initiatives, including creating bilingual resources, developing outreach and training materials for admins and users, hosting monthly community meetings, and maintaining an email list to openly share knowledge, expertise, and researcher needs (Goodchild & Huck, 2022).

Discussion

Creating spaces and supports for the community to flourish is essential for the viability of Borealis. Feedback provides insight to set priorities for technical and service developments, and community involvement in the development of user guides, admin guides, and other projects ensures that resources meet the needs of the research community. The overall goal of the

community — to encourage successful sharing and reuse of research data — aligns with national efforts to strengthen digital research infrastructure and the RDM community in Canada (the Alliance RDM WG, 2020).

Considerations for Data Sharing

Data sharing requires planning. At the project outset, as part of a Data Management Plan, researchers must consider software and tools needed to collect, analyze, and document data; appropriate storage and backup procedures; how data will be deposited and (if possible) shared; and how they will manage data to ensure ethical and legal requirements are met.

Disciplinary differences, including attitude and culture, can influence data sharing and reuse. Certain research fields have traditions of data sharing and reuse and have adopted standards and tools to support this work. Especially within the humanities, where outputs do not always fit within traditional definitions for research data, researchers may consider different approaches to encourage sharing. Services and tools are often developed with disciplinary needs in mind and may be difficult to adopt or repurpose for other disciplines or contexts. Although general tools and services can help, they often lack disciplinary context that would make reuse and adoption possible. Other disciplinary considerations include the following:

- file formats (open vs. proprietary, standard tools and software within the discipline)
- metadata standards for documentation and dataset discovery
- active data storage, data transfer tools, and repository storage to support disciplinary needs (e.g., big data, sensitive data)
- repository selection based on features and user community
- availability of data curation support:
    - data quality review
    - data documentation for reuse
    - data transformation (e.g., cleanup, anonymization, de-identification)
- terms of access and licensing for reuse
- data exploration and visualization tools
- the benefits of sharing different types of data

The following case studies examine research projects or disciplinary considerations in the fields of **digital humanities** (Case Study 2), health sciences (Case Study 3), and natural sciences (Case Study 4), highlighting issues faced by researchers and exploring solutions and lessons learned.

Case Study 2: Digital Humanities

Background

Queen's University Library hosts the Diniacopoulos Collection Virtual Exhibit, the culmination of a research project that presents virtual reality movies and scaled 3D models of Greek and Egyptian archaeological artifacts from the Classics Department collection. The virtual exhibit is built in WordPress using Object2VR software to create an interactive experience for users to rotate and examine objects in virtual 3D in the browser.

Analysis

Researchers wanted to share and preserve the data from this project for future use as the field of virtual reality continues to grow. Web-based viewers and content management systems require ongoing maintenance of software and tools with an unknown lifespan, revealing considerations around sustainability and long-term access. Challenges were encountered in selecting a repository, given the size of the dataset (60 GB), the large number of files (6500+), and the complex folder structure, and that there are few options and best practices for this field. Additionally, including documentation and disciplinary metadata was crucial to ensure that the data could be reused and understood outside of the original context.

Discussion

The research team deposited the dataset into the Queen's Dataverse Collection (Jones, Bevan, & Monette, 2017), part of Borealis, to benefit from the support of the Queen's University Library, as well as features such as comprehensive metadata fields and the ability to assign a **Digital Object Identifier (DOI)** that could be linked from the virtual exhibit. The Borealis team supported the deposit of large zip archive folders for each artefact. Debate continues around the understanding of research data in the humanities, and further investigation is needed through dataset metrics and citations to determine whether there are challenges around the reuse of this contextual data and whether improved tools and platforms would better manage, share, and preserve these types of digital humanities projects.

Case Study 3: Sharing Sensitive Data

Background

Sensitive data refers to any data that may cause harm if made public. Typically, this refers to data collected about human subjects and can include sensitive, confidential, or personal information related to an individual's health, ethnicity, political views, and/or geographic location, to name a few. Research data involving human subjects must be managed in accordance with Research Ethics Board (REB) guidelines and approval. Many institutions provide security standards and protection guidelines for managing sensitive and confidential data.

In Canada, research funded by the three federal research funding agencies (**the agencies**) involving human subjects is guided by the Tri-Council Policy Statement (TCPS 2): Ethical Conduct for Research Involving Humans (GoC Panel, 2023). Researchers must adhere to the policy, which covers issues related to consent, privacy, fairness, and equity, in relation to different types of human research, including clinical trials, genetic research, and research involving First Nations, Inuit and Métis. Research involving Indigenous peoples may not be subject to **TCPS 2** guidelines depending on circumstances and terms agreed to or governing the research data that is considered under the control of the participants or community groups (see chapter 3, "Indigenous Data Sovereignty"; review the OCAP® principles for a model dealing with data on First Nations). The handling and use of sensitive data may be governed by other legal and ethical frameworks of the research program (e.g., CIHR, SSHRC) or institution, or at the provincial (e.g., FIPPA, PHIDA) or federal (e.g., Privacy Act, PIPEDA) level.

In 2021, the Tri-Council provided guidelines for researchers titled Guidance on Depositing Existing Data in Public Repositories (Government of Canada Panel on Research Ethics, 2021), which state that researchers may deposit and share data in a repository if they have received consent from participants to do so and/or if they receive approval from an REB. Researchers must be in compliance with the TCPS 2 before deposit and sharing and must seek REB approval before collection or reuse of human subject research.

Analysis

Infrastructure and support services for sensitive data storage, deposit, and sharing continues to be a major gap in Canada. The complexity around sensitive data requires intersection with a

number of units on campus, including REB guidelines, contracts and legal support, RDM practices, and infrastructure and workflows to manage sensitive data throughout the lifecycle.

For health sciences research, there are several avenues to publish or share sensitive data with various considerations. De-identification or anonymizing the dataset involves removing identifiable data from a dataset. However, some datasets cannot be de-identified without compromising the usefulness of the data. Data may be shared through closed-access portals with data sharing/transfer agreements. One potential downside of this arrangement is the administrative overhead and potential need for a custom portal.

Discussion/Conclusions

There are ongoing efforts to improve tools, infrastructure, workflows, and resources around sensitive data management and sharing. Software, such as Research Electronic Data Capture (REDCap), has grown in popularity as a tool to capture data for clinical research and create databases and projects that are compliant with legal guidelines, secure, and easy to use (Patridge & Bardyn, 2018). The Alliance's Sensitive Data Repository Project has led to the creation of a zero-knowledge encryption tool to facilitate secure deposit and controlled access to sensitive data within the FRDR platform. For the next phase of the project, the Alliance RDM team is leading the collaborative participation of institutions in policy framework development, which aims to clarify and streamline the workflow of sensitive data deposit and sharing. The Alliance's Sensitive Data Expert Group has released resource documents to provide guidance around RDM practices in the context of research ethics frameworks, including the Sensitive Data Tool Kit:

- Part 1: Glossary of Terms for Sensitive Data used for Research Purposes
- Part 2: Human Participant Research Data Risk Matrix
- Part 3: Research Data Management Language for Informed Consent

Researchers need continued leadership for national solutions to ensure equitable access to support, tools, and infrastructure for sensitive data management and sharing.

Case Study 4: Supporting Canada's Large Data Producers:

SuperDARN and the Federated Research Data Repository

Background

The Super Dual Auroral Radar Network (SuperDARN) is a network of three dozen scientific radars deployed around the world by universities and government laboratories in ten countries. SuperDARN Canada (headquartered at the University of Saskatchewan) operates five radars in Canada, which produce valuable data that researchers can use to understand space weather, radio communication, and physics of the Earth's upper atmosphere. However, due to the high-quality captures and rapid collection rates of the radars, SuperDARN is generating data at a massive scale, and storing this data in a manner that is secure, discoverable, and accessible is challenging. In 2018, SuperDARN Canada began meeting with the team at the FRDR.

Analysis

The size, scale, scope of the data, and complexity of SuperDARN's organizational framework as an international research partnership presented numerous challenges. SuperDARN's data collection began in 1993, and the data exist in both raw and processed forms. SuperDARN Canada and the FRDR team had to consider which format of the data (approximately 80 TB of raw data or approximately 10 TB of processed data for each algorithmic version) would be best to publish and, of the processed data, which algorithm generation to choose — the older algorithm that has widespread use or the newer one. **Versioning** datasets to update the outdated algorithm would mean doubling the size of the collection.

Data are collected across time, regions, and instruments via radar installations operating in the Northern and Southern Hemispheres, so the teams had to consider how to subdivide the data into publishable units that would be best suited for discovery, reuse, and usage tracking and reporting. They also had to consider the size of a dataset and number of files and take into account web browser limitations. And while the files are small, depending on how the data were organized, datasets had the potential to be many terabytes.

Since raw and processed data were available only as binary file types, the FRDR curation team could not perform quality checks on the files. The complexity of data also meant that without extensive documentation, the datasets would be useful only to a small number of users involved with the research.

Discussion/Conclusions

Format

The team decided to publish the raw form of the data dating back to 1993.

Curation

The FRDR curation team worked with SuperDARN Canada to review the datasets and develop **README files** that capture detailed descriptive and technical metadata required for reuse of the data by the broader researcher community. Links to associated publications and documentation were added, and datasets were linked to analysis and visualization software developed by SuperDARN.

Lessons Learned

In addition to the solutions discussed above, the following lessons were learned from this project:

- Consultation on data publication needs can take time and is an ongoing process. It took several years from the initial conversation to when the first datasets were ingested; and beyond publication, FRDR and SuperDARN Canada still meet periodically.
- Consistent communication is important, particularly when decisions require longer timeframes; setting regular meetings and documenting discussions and decisions ensures that everybody remains on the same page and that threads do not get lost.
- Sustainability and future planning are key. When working with SuperDARN, FRDR needed to think about the data publication needs associated with the collection and the commitment going forward.

Future of Data Sharing in Canada

There are a number of developments that could better support Canadian researchers with maximizing the benefits of data sharing. A few suggested areas are provided below, including improving access and inclusion, enhancing research platforms that support the lifecycle of data, developing tools and technologies to

automate curation workflows, and improving **integration** and interoperability between systems and platforms.

Access and Inclusion

Systemic barriers to the inclusion of all researchers and disciplines in accessing and using data sharing tools and services must be removed to support more equitable adoption of data sharing policies and practices. New ways of thinking about data sharing are needed to transform infrastructures that support all types of research data; both in terms of formats and standards, but also the related conceptual models and workflows.

As data sharing workflows mature, attention must be paid to ensuring equitable publishing models are created. Given the high cost of storage, particularly for large datasets, we need to balance sustainability and equity.

Examples

- greater customization of data repositories and flexible tools and standards
- web accessibility standards within software and platforms
- open access agreements between research institutions, publishers, and repositories

Research Lifecycle Platforms

Typical repository workflows for uploading or downloading data require moving data across platforms and between storage locations. Transferring data in this manner is inefficient and costly, and it may be impossible or infeasible for large datasets due to cost, transfer times, or infrastructure limitations. Additionally, certain datasets rely on specialized software or computational environments for analysis. Research platforms and underlying storage clusters that accommodate the full lifecycle of data are needed, where data can be analyzed and curated and an authoritative version shared.

Examples

- easy-to-use tools for deploying datasets between different storage layers (e.g., moving data to and from repository and active storage)
- all-in-one cloud platforms for data analysis, curation, and sharing

Curation Automation

Making data available is not enough to advance open science, which requires significant resources of time and money to make datasets FAIR. New tools and technologies could reduce this investment and support researchers and curators in producing high-quality outputs.

Examples

- artificial intelligence algorithms that generate high-quality metadata from data
- software for automated data linkage within and across datasets
- software that guides researchers in documenting their datasets, with built-in standards and taxonomies
- software that checks for reproducibility and dataset quality

Integration and Interoperability

As demonstrated by the range of policies, tools, and services supporting the sharing of research data, there is significant momentum to progress these infrastructures. However, many are provided and developed in relative isolation, with only a few pieces of middleware or policy frameworks to connect them. As these infrastructures are developed, interoperability (e.g., connecting policy to platform, platform to service, service to policy, etc.) and integration into the research and publishing workflows will be a central focus to improve ease of use and increase adoption of data sharing practices.

Examples

- policy frameworks for sharing data across jurisdictional boundaries
- Incorporating Data Management Plans into research and sharing infrastructure
- connecting datasets into a broader network of research outputs

Conclusion

Canadian infrastructure, tools, and services that support sharing research data are important, particularly given policies requiring access to publicly funded data. A researcher's area of study and ethical considerations impact the way data is shared and influence developments of policy and infrastructure that could advance data sharing in Canada.

Reflective Questions

1. What are the challenges of sharing research data?
2. What are the types of data storage? Provide an example of each.
3. What are considerations around data sharing? How do disciplinary differences play a role in sharing?
4. What kinds of data services (local, domain-specific, or national) could be developed to address the challenges and barriers identified in this chapter?

Key Takeaways

- Funders and publishers may set requirements that promote research data sharing; however, policies alone are not sufficient to ensure results are reproducible. Technical and discipline-specific solutions may be required to ensure that data is accessible and reusable.
- Storage options, infrastructures, and data repositories in Canada support the production, sharing, and reuse of research data over its lifecycle. Research data storage can be broken down into three types: active, repository, and archival. Canadian research institutions often provide storage infrastructures to their researchers, though availability depends on institutional capacity.
- Support services exist for Canadian researchers developing RDM practices, publishing data, or planning for reuse of data, including their own research groups, higher-education institutions, and services unique to the needs of specific research communities.
- Researchers should consider disciplinary differences and context around data sharing. Certain fields have a tradition of open data sharing and reuse. While some disciplines have adopted standards and tools to support this work, others may need support and tools to address areas such as metadata, file size, data type, and requirements around data sensitivities.

- Data sharing and reuse is supported through integration and interoperability between systems and platforms, including platforms that support the lifecycle and technologies that facilitate curation workflows.

Additional Readings and Resources

Barsky, E., Laliberté, L. W., Leahey, A., and Trimble, L. (2017). Chapter 3. Collaborative Research Data Curation Services: A View from Canada. In L. R. Johnston, *Curating research data, volume one: Practical strategies for your digital repository* (79-101). Association of College and Research Libraries. https://dx.doi.org/10.14288/1.0340778

Cheung, M., Cooper, A., Dearborn, D., Hill, E., Johnson, E., Mitchell, M., & Thompson, K. (2022). Practices before policy: Research data management behaviours in Canada. *Partnership: The Canadian Journal of Library and Information Practice and Research*, *17*(1), 1–80. https://doi.org/10.21083/partnership.v17i1.6779

The First Nations Information Governance Centre. (2014, May). *Ownership, control, access and possession (OCAP™): The path to First Nations information governance*. https://achh.ca/wp-content/uploads/2018/07/OCAP_FNIGC.pdf

Garnett, A., Leahey, A., Savard, D., Towell. B., & Wilson, L. (2017). Open metadata for research data discovery in Canada. *Journal of Library Metadata*, *17*(3-4), 201-217. https://doi.org/10.1080/19386389.2018.1443698

Thompson, K., & Kellam, L. M. (Eds.). (2016). Introduction to databrarianship: The academic data librarian in theory and practice. In L. M Kellam & K. Thompson (Eds.), *Databrarianship: The academic data librarian in theory and practice*. Association of College and Research Libraries. https://scholar.uwindsor.ca/cgi/viewcontent.cgi?article=1047&context=leddylibrarypub

Rice, R., & Southall, J. (2016). *The data librarian's handbook*. Facet Publishing.

Reference List

The Alliance Research Data Management Working Group. (2020). *The current state of research data management in Canada*. https://alliancecan.ca/sites/default/files/2022-03/rdm_current_state_report-1_1.pdf

Baker, D., Bourne-Tyson, D., Gerlitz, L., Haigh, S., Khair, S., Leggott, M., Moon, J., Ridsdale, C., Tourangeau, R., & Whitehead, M. (2019). *Research data management in Canada: A backgrounder*. Zenodo. https://doi.org/10.5281/zenodo.3574685

Corcho, O., Eriksson, M., Kurowski, K, Ojsteršek, M., Choirat, C. van de Sanden, M., & Coppens, F. (2021). *EOSC interoperability framework: Report from the EOSC executive board working groups FAIR and architecture*. Publications Office of the European Union. https://data.europa.eu/doi/10.2777/620649

Goddard, L., Barsky, E., Cooper, A., Darnell, A., Davis, C., Doiron, J., & Taylor, S. (2018). *Dataverse north working group: Year 1 recommendations*. UBC Faculty Research and Publications. https://doi.org/10.14288/1.0386773

Goodchild, M., & Huck, J. (2022, March). *Building a shared open research data repository community in Canada*. Open Science Framework. osf.io/n8wzt

Government of Canada. (2022). *Published institutional research data management strategies*. https://www.ic.gc.ca/eic/site/063.nsf/eng/h_98428.html

Government of Canada. (2021). *Tri-Agency research data management policy*. https://www.science.gc.ca/eic/site/063.nsf/eng/h_97610.html

Government of Canada. (2016). *Research data archiving policy*. Social Sciences and Humanities Research Council. https://www.sshrc-crsh.gc.ca/about-au_sujet/policies-politiques/statements-enonces/edata-donnees_electroniques-eng.aspx

Government of Canada Panel on Research Ethics. (2021, May 25). *Guidance on depositing existing data in public repositories*. https://ethics.gc.ca/eng/depositing_depots.html

Government of Canada Panel on Research Ethics. (2023, January 11). *Tri-Council policy statement: Ethical conduct for research involving humans – TCPS 2 (2022)*. https://ethics.gc.ca/eng/policy-politique_tcps2-eptc2_2022.html

Jacoby, W. G., Lafferty-Hess, S., & Christian, T-M. (2017). *Should journals be responsible for reproducibility?* Inside Higher Ed Blog. https://www.insidehighered.com/blogs/rethinking-research/should-journals-be-responsible-reproducibility

Jones, K., Bevan, G., & Monette, M. (2017). The Diniacopoulos ceramics display, Department of Classics – 2016. https://doi.org/10.5683/SP/T7ZJAF, Borealis, V2, UNF:6:B4MvStefgmeu6++JjcdOQg== [fileUNF]

Lin, D., Crabtree, J., Dillo, I. et al. (2020) The TRUST Principles for digital repositories. *Sci Data*, *7*, 144. https://doi.org/10.1038/s41597-020-0486-7

Patridge, E. F., & Bardyn, T. P. (2018). Research electronic data capture (REDCap). *JMLA*, *106*(1), 142–144. https://doi.org/10.5195/jmla.2018.319

Pérez-Jvostov, F., Iron, K., Khair, S., Sahrakorpi, S., & Zhang, Q. (2021). *Researcher needs assessment: Summary of what we heard*. Digital Research Alliance of Canada. https://alliancecan.ca/sites/default/files/2022-03/needsassessment_alliance_20220126.pdf

Public Library of Science. (2022, March 29). *PLOS launches new feature to promote data sharing and access*. The Official PLOS Blog. https://theplosblog.plos.org/2022/03/plos-launches-new-feature-to-promote-data-sharing-and-access/

Rieseberg, L., Warschefsky, E., O'Boyle, B., Taberlet, P., Ortiz-Barrientos, D., Kane, N. C., & Sibbett, B. (2021). Editorial 2021. *Molecular Ecology*, *30*(1), 1-25. https://doi.org/10.1111/mec.15759

Stuart, D., Baynes, G., Hrynaszkiewicz, I., Allin, K., Penny, D., Lucraft, M., & Astell, M. (2018). *Whitepaper: Practical challenges for researchers in data sharing. figshare*. https://doi.org/10.6084/m9.figshare.5975011

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., & Sepp, T. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific data*, *8*, 192. https://doi.org/10.1038/s41597-021-00981-0

Vines, T. H., Andrew, R. L., Bock, D. G., Franklin, M. T., Gilbert, K. J., Kane, N. C., Moore, J-S., Moyers, B. T., Renaut, S., Rennison, D. J., Veen, T., & Yeaman, S. (2013). Mandated data archiving greatly improves access to research data. The *FASEB Journal*, *27*(4), 1304-1308. https://doi.org/10.1096/fj.12-218164

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., ... Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

About the authors

Meghan Goodchild

Meghan Goodchild is the Research Data Management Librarian at Queen's University and Scholars Portal of the Ontario Council of University Libraries. At Queen's University Library, Meghan is the lead contact for research data management and collaborates with campus partners to improve workflows and services supporting the research data lifecycle. At Scholars Portal, Meghan provides leadership on the team supporting Borealis, the Canadian Dataverse Repository. She holds a PhD in Music Theory and a Masters of Information Studies from McGill University.

Shahira Khair

Shahira Khair (she/her) is a Librarian at the University of Victoria Libraries, with portfolios in organizational analysis and data management. Prior to joining UVic, she worked with national organizations advancing digital initiatives in research and higher ed, including the Canadian Association of Research Libraries and the Digital Research Alliance of Canada. She holds a Masters of Science in Biology and a Masters of Information Studies from the University of Ottawa.

Amber Leahey

Amber Leahey is a Data & GIS Librarian and the Service Director for Borealis, the Canadian Dataverse Repository, a secure, bilingual, national data repository provided in partnership with academic libraries and research institutions across Canada. In her role she supports libraries, institutions, and researchers with data management, sharing, preservation, and reuse, through ongoing development of data and research services at Scholars Portal and the University of Toronto Libraries. She holds a Master of Library and Information Studies from the University of Toronto.

Kaitlin Newson

Kaitlin Newson is a Digital Research Consultant with ACENET at the University of Prince Edward Island. Formerly, Kaitlin was the Digital Projects Librarian with Scholars Portal, a service of the Ontario Council of University Libraries, where she supported digital infrastructure for research data management, scholarly publishing, and cloud storage services for Canadian university libraries. She holds a Master of Information from the University of Toronto.

Lee Wilson

Lee Wilson is the Director of Research Data Management at the Digital Research Alliance of Canada. In this role, Lee oversees the national Research Data Management team at the Alliance, as well as service provision and development through partnerships with a variety of Canadian institutions and organizations. Lee previously held the position of Manager for RDM Platforms and Services at the Alliance, worked as a Research Consultant for Data Management in Atlantic Canada with ACENET, and as a part of the data management team for the Marine Environmental Observation Prediction and Response Network, supporting researchers working with oceans data. He holds a Master of Library and Information Studies from Dalhousie University.

6.

THE RDM MATURITY ASSESSMENT MODEL IN CANADA (MAMIC)

Jane Fry; Jennifer Abel; Dylanne Dearborn; Alison Farrell; and Chantal Ripp

Learning Outcomes

By the end of this chapter you should be able to:

1. Explain what a maturity assessment model is.
2. Understand the value of completing a Research Data Management maturity assessment.
3. Understand why and how a made-in-Canada maturity assessment model was developed.
4. Be able to use the Maturity Assessment Model in Canada to assess Research Data Management service maturity at a Canadian research institution.
5. Be able to support evidence-based decision making with the results gathered from the completed Maturity Assessment Model in Canada.

Introduction

By now you know that **Research Data Management (RDM)** involves a range of practices and services such as data management planning, curation, discovery, and preservation. So research institutions — universities, colleges, hospitals — thinking about RDM should consider all services, resources, and personnel that support RDM for every research project, particularly when an institution is formalizing services, as many Canadian institutions were at the time of writing (spring 2022) in response to the **Tri-Agency Research Data Management Policy**.

But how can people at a research institution determine whether all areas of the **research data lifecycle** are being supported and who is responsible for what? To support Canadian research institutions in undertaking this important step, the authors of this chapter came together in the summer of 2021 to develop the RDM Maturity Assessment Model in Canada, or **MAMIC** (Fry et al., 2021), to help RDM partners understand the services and resources available to support data management at their institution.

In this chapter, we'll examine why Canadian institutions may want to conduct an **RDM maturity assessment** for their institution, in particular, looking at the institutional RDM strategy requirement which was implemented by Canada's three federal research funding agencies (**the agencies**); the development of the MAMIC; and how to complete the MAMIC and use its results. Finally, we'll highlight the importance of community efforts in creating the tool.

Access the MAMIC here: English, French

The Need: How to Assess an Institution's RDM Services

In the spring of 2021, the Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Social Sciences and Humanities Research Council of Canada (SSHRC) released their long-anticipated Tri-Agency RDM Policy. This policy supports Canadian research excellence by ensuring that researchers engage in sound RDM and data stewardship practices, and that their institutions support them in these practices. The Agencies expect high standards of excellence — that research is performed ethically, funds are used wisely, experiments and studies are replicable, and research results are as accessible as possible (Government of Canada, 2021). To demonstrate this, institutions must create and publish an RDM strategy that sets out their commitment to RDM principles and how they will support their researchers in adopting them (see section 3.1 of the policy).

Since one type of strategy will not fit all situations, each institution should consider its particular circumstances, such as its size, research intensity, and existing RDM capacity. But how does an institution determine what its RDM capacity is or what its strategy should be? To help determine this, in 2018, the Digital Research Alliance of Canada (formerly the Portage Network) released an Institutional RDM Strategy Development Template, which was updated in November 2021. The template outlines a five-stage process to inform and shape the creation of an RDM strategy that meets local needs and resource capacities. We'll focus

on the second stage of the process, which encourages institutions to assess their state of RDM using assessment models and tools.

What Is a Maturity Assessment Model? And Why Does Canada Need One?

Maturity assessment models and tools evaluate an institution's maturity and readiness for RDM service provision and help determine the level of sophistication of a service or product. A common feature of these models is the use of a scale to represent an organization's maturity in specific capabilities — in other words, how reliably the organization performs the process (Rans & Whyte, 2017). The maturity rubric allows a user to quantify capabilities and enable continuous process improvement.

Internationally, RDM is well-established for enabling research excellence, and several maturity models have already been developed. However, when Canadian institutions began using these models to evaluate the state of RDM on their campuses, they found that these tools did not align with the Canadian RDM landscape; for example, Canadian institutions are not required to have RDM policies, unlike institutions in some other countries.

In 2021, after the release of the Tri-Agency RDM Policy, members of the national RDM community began informally discussing how institutions should go about creating their RDM strategies. The National Training Expert Group (NTEG)[1], a group of information professionals, researchers, and practitioners, decided to create a series of webinars and workshops to be presented in October 2021 to bring representatives of different institutions together to discuss their strategy development work. While planning for the fall series, several members noted that there was no Canadian maturity assessment model that institutions could use in the second stage of their strategy development. NTEG decided that a Canadian-focused maturity assessment model tool could be a useful starting place for discussions about institutional RDM capacity with strategies in alignment with the Tri-Agency RDM Policy. In April 2021, a smaller group set to work on developing what would become the first version of the MAMIC, in time for the October workshop. We — the authors of this chapter — along with Shahira Khair of the University of Victoria, were the members of that group.

1. NTEG is part of the Network of Experts that is affiliated with the Digital Research Alliance of Canada.

How the MAMIC Was Created

Environmental Scan of Maturity Assessment Models

As a first step, we examined several international assessment tools. While the available models were excellent, they included sections not applicable to Canada, and there were gaps — things we needed to include in the Canadian model, one of them being the requirement by the agencies for an RDM Institutional Strategy. After our review, we focused on aspects of the three most popular models to help develop the MAMIC — the Research Infrastructure Self Evaluation Framework (RISE) tool, published in 2017 by the Digital Curation Centre (DCC); the Evaluate your RDM Offering tool by SPARC Europe; and the Data Management Framework of the Australian National Data Service (ANDS)[2]. We based our Canadian model primarily on RISE, with elements inspired by the SPARC model and the ANDS frameworks.

Overview of the MAMIC

In our Canadian tool, we detail the reason for, intention of, and definition of the MAMIC, and provide a section on how to complete it. There are four tables to be filled out by the research partners:

- Institutional Policies and Processes
- IT Infrastructure
- Support Services
- Financial Support

Each table has five columns:

- element being assessed
- definition of that element
- its maturity level (how advanced the institution's RDM is)
- its scale (who may access the service or support)
- any required explanation as to the rating given to that element

2. Since the time of the development of the MAMIC, ANDS has been folded into the Australian Research Data Commons (ARDC), https://ardc.edu.au/

Below each table, there's space for the date of completion as well as the name(s) and role(s) of the person(s) filling it out, since users need to know who those research partners are in order to address questions or concerns about how the table was completed. The MAMIC is also intended to be used in the future, so it's important to know who filled in the previous version.

We also wanted to ensure that the terms used were well defined, so we included a page of definitions of maturity and scale levels specific to each table, along with hints to help fill them out.

Initial Version of the MAMIC

After receiving feedback from members of the RDM community, a draft was completed, the MAMIC was translated into French (where it is the **MEMAC,** or **Modèle d'évaluation de la maturité de la GDR au Canada)** and introduced to the attendees at the Institutional Strategies Workshop in late October 2021.

Note: There are certain areas not covered in this initial version, so changes will need to be made in future versions. For example, a future version should contain considerations for Indigenous data sovereignty. Another idea is to explore different ways to present the tool, such as developing an online tool that could allow users to produce different types of charts, as the SPARC tool does.

These revisions will be useful for those who plan to do this type of assessment on a regular basis as part of reviewing and revising their institutional RDM strategy or as part of ongoing service improvements. It may also be useful to apply the MAMIC at a national scale to highlight and address gaps and to showcase where institutions may be able to rely on national resources.

Using the MAMIC

The MAMIC can be used to determine whether RDM resources and services exist, as well as who is responsible for these different supports so that institutions can support researchers in effective data management and also be aware of what is needed to supplement their current offerings. Using the model involves coordination between interested parties across campus, such as the library, research office, ethics office, and IT department.

Categories and Measures

Before the work begins, research partners filling in the MAMIC should discuss the process so everyone understands how scales and measures will be applied and to ensure that key decisions about how to do the

work are documented. Each category (Institutional Policies and Processes, IT Infrastructure, Support Services, and Financial Support) can be assessed in its own table — see Appendix 2 for an example of a completed Institutional Policies and Processes category — using three different measures:

Measure 1: **Maturity Level** of the element at the institution is rated on a 5-level scale ranging from "does not exist" OR "do not know" to "robust and focuses on continuous evaluation." Note: the first level of this scale is 0 *not* 1 because sometimes an institution cannot provide a service or support, or does not feel they need it. The category "does not exist" is not meant to indicate a level of maturity, but rather an acknowledgement that this element is not available to researchers at an institution.

Measure 2: **Scale** is used to identify who can access the service or support. The element may not be applicable to certain users or is not available to all populations. This allows an institution to see whether its services are being offered in an equitable and appropriate manner or if there are accessibility issues.

Measure 3: **Comments** are perhaps the most important measure because this section identifies specific strengths and weaknesses and provides an avenue for discussion. This is also a place where regional, national, consortial, or other tools that complement the institution's RDM maturity can be noted. It can be difficult to determine the maturity level or scale of an element if there are multiple initiatives within that element (e.g., multiple units offering similar data management services), so the comments section can be used for explanations of such instances.

Filling in the MAMIC

The data collected in the MAMIC are for that particular institution's use only; none will be collected by any other organization, and those who fill in the MAMIC are the ones to decide how to collect and use their data.

While the MAMIC can be completed by an individual, we recommend that a group of interested research partners be involved. These people should come from the areas being assessed. For example, IT Infrastructure should be assessed by representatives of IT; Financial Support should be assessed by representatives of the areas which provide RDM services and support (e.g., libraries, IT, research services).

After the MAMIC was made public, the RDM community shared four examples of the MAMIC completion process with us, and each looked similar. Three of the institutions had a small working group composed of librarians, IT representatives, research office members, and either a researcher or an industry partner. At one of the institutions, however, the data librarian filled in the entire document, reaching out to colleagues in the office of research services to help fill in gaps. This method was less effective and a tremendous amount of work, by comparison. In each case, the results of the MAMIC were taken to a larger RDM committee for discussion.

Benefits of the MAMIC

When developing RDM strategies and supports, partners must reflect on the state and scope of RDM services and supports at their institution and on future needs and desires. A maturity assessment model, like the MAMIC, can help identify gaps, strengths, weaknesses, challenges, and opportunities that exist in the research data landscape. This helps the institution decide where resources and efforts should be directed so they can have supports in place to ensure the success of researchers.

Effective use of this tool creates a complete and representative assessment, but this process requires collaboration and input from a variety of research partners; so a benefit of using the MAMIC is that these types of opportunities for discussion can open the door for relationship building. This supports the institutional RDM landscape and presents opportunities for dialogue, collegiality, and partnership outside of RDM. For example, it may open up lines of communication between IT services and the library's internal IT unit to allow for greater integration of library services and IT resources.

Bringing together research partners by using a shared tool can illustrate the complexity of RDM and the breadth of efforts across an institution. This can help break down silos and distinguish areas of expertise within the institution, draw connections and interactions, and highlight areas for collaboration and discussions about the institutional strategy and priorities, resource allocation, or budget considerations. Using the same tool over time can also be helpful for benchmarking, to track institutional developments and progress.

On a larger scale, the MAMIC may facilitate conversations between Canadian institutions. Noting where external resources are available or are being developed can help institutions decide where to invest locally. Also, identifying gaps across institutions may offer an opportunity to forge new national initiatives. This can reduce the duplication of effort to solve each gap at an institutional level, which can be time consuming, costly, and require dedicated staff support.

Conclusion

This chapter has presented the MAMIC in two ways: as a tool that RDM practitioners and institutions can use in current or future RDM work, and as a useful example of how Canadian RDM community members can create tools to help everyone work more effectively and efficiently. We identified a need and set out to fill it using the skills and techniques we use elsewhere in our work: conducting environmental scans and literature reviews, developing materials for user groups, gathering user feedback, and working as a team. We also used the resources and people available to us — in particular, the national RDM Network of Experts community and the RDM team at the Digital Research Alliance of Canada — to help develop and disseminate the tool.

Reflective Questions

Choose a category of the MAMIC to reflect on, and then complete the following:

- Consider what research partners should be involved in order to get an accurate picture of RDM supports offered in this category at an institution. How would you encourage participation from them?
- List four ways the MAMIC can help assess the level of RDM support at an institution.

Key Takeaways

- A maturity assessment model is a tool to determine the level of sophistication of a service or product.
- Maturity assessment models specific to RDM have been developed by different international organizations and have been used for years to assess RDM support services.
- The MAMIC was developed to reflect the needs of Canadian institutions that are creating institutional RDM strategies.
- Completing the MAMIC allows research partners to engage in discussion and evaluation about the state of RDM at their institution, to understand the breadth of RDM offerings and support, and to collaborate across divisions.
- There are a variety of ways in which completing the MAMIC could be used to help in institutional decisions and discussions around RDM. This can enable research partners to move their institution forward by making evidence-based decisions about how RDM services and resources could develop in the future.

Additional Readings and Resources

Australian Research Data Commons (ARDC). https://ardc.edu.au/

Digital Research Alliance of Canada – Putting the Policy into Practice Webinar Series, October 2021

- Session 1: Introduction to the Tri-Agency RDM Policy and Data Management Plans: English, French
- Session 2: Data Deposit: English, French
- Session 3: Institutional Strategies: English, French
- Session 4: Panel on Institutional Strategies: English, French

Digital Research Alliance of Canada – Research Data Management. https://alliancecan.ca/services/research-data-management

Fry, J., Doiron, J., Létourneau, D., Perrier, L., Perry, C., & Watkins, W. (2017, January 31). *Research data management training landscape in Canada: A white paper* [R]. http://dx.doi.org/10.14288/1.0372048

Institutional RDM Strategy Template Revision Working Group. (2021). Institutional research data management strategy development template (3.0). Zenodo. https://doi.org/10.5281/zenodo.5745906

Jacob, B., Whyte, A., Meyer, A., D'haenens, S., Hartmann, N. K., & Weiß, N. (2019, October 2). *Using RISE, an international perspective* [lightning talk].15th International Digital Curation Conference (IDCC), Dublin, Ireland. https://doi.org/10.5281/zenodo.3565440

Jones, S., Pryor, G., & Whyte, A. (2012). Developing research data management capability: The view from a national support service. In: R. Moore, K. Ashley, & S. Ross (Eds.), *Proceedings of the 9th International Conference on Preservation of Digital Objects (iPRES)*, (pp. 142-149). Toronto: University of Toronto Faculty of Information. https://phaidra.univie.ac.at/detail/o:293775

Jones, S., Rans, J., Sisu, D., & Whyte, A. (2014). Reshaping the DCC institutional engagement programme. *International Journal of Digital Curation*, *9*(2), 47-64. https://doi.org/10.2218/ijdc.v9i2.334

Kouper, I., Fear, K., Ishida, M., Kollen, C., & Williams, S. C. (2017). Research data services maturity in academic libraries [B]. http://dx.doi.org/10.14288/1.0343479

National Research Data Management Network of Experts. https://alliancecan.ca/services/research-data-management/network-experts

Perry, C., Fry, J., & Doiron, J. (2017, June 1). Portaging the landscape: Developing and delivering a national RDM training infrastructure in Canada [presentation]. IASSIST 2017. https://doi.org/10.5281/zenodo.4551708

SPARC Europe. *How open are you?* https://sparceurope.org/what-we-do/open-access/sparc-europe-open-access-resources/open-research-checklist-institutions/

UC3. (September 12, 2016). *Building a user-friendly RDM maturity model*. https://uc3.cdlib.org/2016/09/12/building-a-user-friendly-rdm-maturity-model/

Reference List

ANDS. (2018, March 23). *Creating a data management framework*. https://web.archive.org/web/20220309174711/https://www.ands.org.au/__data/assets/pdf_file/0005/737276/Creatinga-data-management-framework.pdf

Fry, J., Dearborn, D., Farrell, A., Khair, S., & Ripp, C. (2021, November 30). RDM maturity assessment model in Canada (MAMIC) (1.0). Zenodo. https://doi.org/10.5281/zenodo.5745493

Government of Canada. (2021, March 15). *Tri-Agency research data management policy*. https://www.science.gc.ca/eic/site/063.nsf/eng/h_97610.html

Rans, J., & Whyte, A. (2017). Using RISE, the research infrastructure self evaluation framework v.1.1. Edinburgh: Digital Curation Centre. www.dcc.ac.uk/guidance/how-guides/RISE

SPARC Europe. *Evaluate your RDM offering*. https://sparceurope.org/evaluate-your-rdm-offering/

About the authors

Jane Fry

Jane Fry is the Data Services Librarian at Carleton's MacOdrum Library where Research Data Management is one of her responsibilities. She is also the lead on Carleton's Institutional RDM Strategy Working Group. As well, she chaired the Training Expert Group under Portage for four years and remains a member of that group today.

Jennifer Abel

Jennifer Abel is the Research Data Management Specialist in the University of Calgary's Research Services Office, and is coordinating the development of UCalgary's RDM Strategy. She previously worked as the Training Coordinator for both the Portage Network and the Digital Research Alliance of Canada's RDM team. She holds a PhD in Linguistics and an MLIS from the University of British Columbia.

Dylanne Dearborn

Dylanne Dearborn is the Research Data Management Coordinator and the Research Data Librarian for Sciences & Engineering at the University of Toronto, in the Map and Data Library. She chaired the Research Intelligence Expert Group with Portage for three years and remains a member of that group today.

Alison Farrell

Alison Farrell is a Research Data Management and Public Services Librarian at the Health Sciences Library at Memorial University of Newfoundland and is a member of the RDM Institutional Strategy working group at Memorial. She was the co-chair for the Portage Institutional Strategies working group, whose mandate was to provide educational materials on developing institutional strategies.

Chantal Ripp

Chantal Ripp is a Research Librarian in the Interdisciplinary Data Team at the University of Ottawa. She is a member of the University's RDM Advisory Group responsible for developing an institutional strategy. She is also a member of a number of other local and national committees, including the DLI Professional Development Committee and the Digital Research Alliance Data Management Plan Expert Group (DMPEG).

SECTION III

WORKING WITH DATA

7.

DATA CLEANING DURING THE RESEARCH DATA MANAGEMENT PROCESS

Lucia Costanzo

Learning Outcomes

By the end of this chapter you should be able to:

1. Describe why it is important to clean your data.
2. Recall the common data cleaning tasks.
3. Implement common data cleaning tasks using OpenRefine.

What Is Data Cleaning?

You may have heard of the 80/20 dilemma: Most researchers spend 80% of their time finding, cleaning, and reorganizing huge amounts of data and only 20% of their time on actual data analysis.

When starting a research project, you will use either primary data generated from your own experiment or secondary data from another researcher's experiment. Once you obtain data to answer your research question(s), you'll need time to explore and understand it. The data may be in a format which will not allow for easy analysis. During the data cleaning phase, you'll use **Research Data Management (RDM)** practices. The data cleaning process can be time consuming and tedious but is crucial to ensure accurate and high-quality research.

Data cleaning may seem to be an obvious step, but it is where most researchers struggle. George Fuechsel, an IBM programmer and instructor, coined the phrase "garbage in, garbage out" (Lidwell et al, 2010) to remind

his students that a computer processes what it is given — whether the information is good or bad. The same applies to researchers; no matter how good your methods are, the analysis relies on the quality of the data. That is, the results and conclusions of a research study will be as reliable as the data that you used.

Using data that have been cleaned ensures you won't waste time on unnecessary analysis.

Six Core Data Cleaning and Preparation Activities

Data cleaning and preparation can be distilled into six core activities: discovering, structuring, cleaning, enriching, validating, and publishing. These are conducted throughout the research project to keep data organized. Let's take a closer look at these activities.

1. Discovering Data

The important step of discovering what's in your data is often referred to as **Exploratory Data Analysis (EDA)**. The concept of EDA was developed in the late 1970s by American mathematician John Tukey. According to a memoir, "Tukey often likened EDA to detective work. The role of the data analyst is to listen to the data in as many ways as possible until a plausible 'story' of the data is apparent" (Behrens, 1997). EDA is an approach used to better understand the data through quantitative and graphical methods.

Quantitative methods summarize variable characteristics by using measures of central tendency, including mean, median, and mode. The most common is mean. Measures of spread indicate how far from the center one is likely to find data points. Variance, standard deviation, range, and interquartile range are all measures of spread. Quantitatively, the shape of the distribution can be evaluated using skewness, which is a measure of asymmetry. Histograms, boxplots, and sometimes stem-and-leaf plots are used for quick visual inspections of each variable for central tendency, spread, modality, shape, and outliers.

Exploring data through EDA techniques supports discovery of underlying patterns and anomalies, helps frame hypotheses, and verifies assumptions related to analysis. Now let's take a closer look at structuring the data.

2. Structuring Data

Depending on the research question(s), you may need to set up the data in different ways for different types of analyses. Repeated measures data, where each experimental unit or subject is measured at several points in time or at different conditions, can be used to illustrate this.

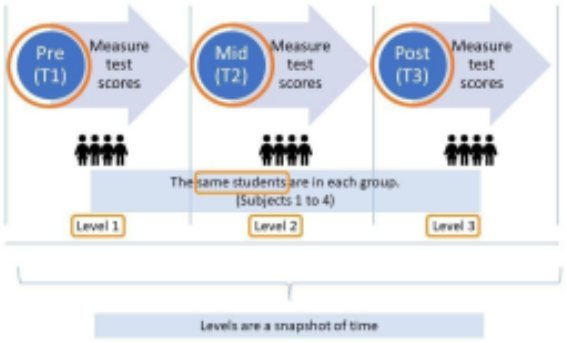


Figure 1. *Investigating the effect of a morning breakfast program.*

In Figure 1, researchers might be investigating the effect of a morning breakfast program on Grade 6 students and want to collect test scores at three time points, such as the pre- (T1), mid- (T2), and post-morning (T3) periods of the breakfast program. Note that the same students are in each group, with each student being measured at different points in time. Each measurement is a snapshot in time during the study. There are two different ways to structure repeated measures data: long and wide formats.

Table 1 shows data structured in long format, with each student in the study represented by three rows of data, one for each time point for which test scores were collected. Looking at the first row, Student One at Timepoint 1 (before the breakfast program) scored 50 on the test. In the second row, Student One at Timepoint 2 (midway through the breakfast program) scored higher on the test, at 65. And in the third row, Student One at Timepoint 3 (after the program) scored 80.

ID	TIME	SCORE
1	1	50
1	2	65
1	3	80
2	1	70
2	2	75
2	3	90
:	:	:

Table 1. *Data structured in long format.*

The wide format, shown in Table 2, uses one row for each observation or participant, and each measurement or response is in a separate column. In wide format, each student's repeated test scores are in a row, and each test result for the student is in a column. Looking at the first row, Student One scored 50 on the test before the breakfast program, then scored 65 on the test midway through the breakfast program, and achieved 80 on the test after the program.

So, a long data format uses multiple rows for each observation or participant, while wide data formats use one row per observation. How you choose to structure your data (long or wide) will depend on the model or statistical analysis you're undertaking. It is possible you may need to structure your data in both long and wide data formats to achieve your analysis goals.

ID	TEST1	TEST2	TEST3
1	50	65	80
2	70	75	90
:	:	:	:

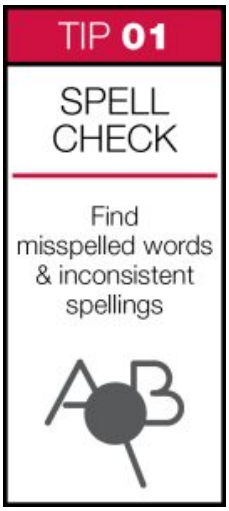
Table 2. *Data structured in wide format.*

Structuring is an important core data cleaning and preparation activity that focuses on reshaping data for a particular statistical analysis. Data can contain irregularities and inconsistencies, which can impact the accuracy of the researcher's models. Let's take a closer look at cleaning the data, so that your analysis can provide accurate results.

3. Cleaning Data

Data cleaning is central to ensuring you have high-quality data for analysis. The following nine tips address a range of commonly encountered data cleaning issues using practical examples.

Tip 1: Spell Check



Finding misspelled words and inconsistent spellings is one of the most important data cleaning tasks. You can use a spell checker to identify and correct spelling or data entry errors.

Spell checkers can also be used to standardize names. For example, if a dataset contained entries for "University of Guelph" and "UOG" and "U of G" and "Guelph University" (Table 3), each spelling would be counted as a different school. It doesn't matter which spelling you use, just make sure it's standard throughout the dataset.

Tip 01 Exercise: Spell Check

Go through Table 3 and standardize the name as "University of Guelph" in the SCHOOL column.

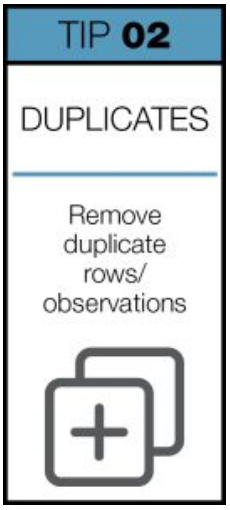
ID	AGE	SCHOOL	GRADE
1	17	Universtiy of Guelph	88
2	21	UOG	60
3	18	University of Guelph	80
4	19	University of Guelph	75
8	18	University of Guelph	72
12	21	University of Guelph	60
13	18	University of Guelph	80
14	19	Guelph University	77
15	18	University of Guelph	49
16	21	U of G	60
17	18	University of Guelph	88
19	19	Guelph University	73
20	18	University of Guelph	72

Table 3. *Data requiring spell checking.*

View Solutions for answers. Data files for the exercises in this chapter are available in the Borealis archive for this text.

Tip 2: Duplicates

Sometimes data have been manually entered or generated using methods that could cause duplications of rows. Check rows to determine if data have been duplicated and need to be deleted. If each row has an identification number, it should be unique for each observation. In this example, there are two observations with an ID number of 3 (Table 4) and both have the same values, so one of these rows of data should be deleted.



Tip 02: Exercise: Duplicates

Go through Table 4 and delete the duplicate observations.

HINT: If using Excel, look for and use the 'Duplicate Values' feature.

ID	AGE	SCHOOL	GRADE
1	17	Universtiy of Guelph	88
2	21	UOG	60
3	18	University of Guelph	80
4	19	University of Guelph	75
3	18	University of Guelph	80
12	21	University of Guelph	60
13	18	University of Guelph	80
14	19	Guelph University	77
15	18	University of Guelph	49
16	21	U of G	60
17	18	University of Guelph	88
19	19	Guelph University	73
20	18	University of Guelph	72

Table 4. *Data with duplicate rows that should be deleted.*

View Solutions for answers.

Tip 3: Find and Replace

With some carefully crafted replacements, it's possible to get data fairly clean and into a good format by looking for patterns and repetition in a file. In this example, we're counting the number of bird sightings in Guelph. Looking at the LOCATION column, we need to replace the abbreviations "St" and "ST" with the full spelling of "street" (Table 5). This can be done using the basic Find and Replace function.

Tip 03 Exercise: Find and Replace

Go through Table 5. Find and replace all instances of “ST” and “st” with “Street” in the LOCATION column.

HINT: Use caution with global Find and Replace functions. In the example shown below, instances of ‘St’ or ‘st’ that do NOT indicate ‘Street’ (e.g. ‘Steffler’ and ‘First’) will be erroneously replaced. Avoid such unwanted changes by strategically including a leading space in the string you are searching for (so, ‘spaceSt’). Experiment with the ‘match case’ feature as well, if available. Always keep a backup of your unchanged data in case things go awry.

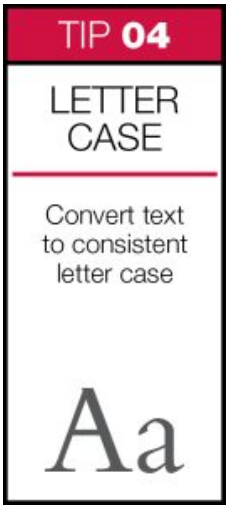
ID	BIRD	LOCATION	TOTAL
1	17	Quebec St	6
2	21	Cork Street	5
3	18	Moffatt St	8
4	19	Victoria Street	5
5	18	Steffler St	8
6	21	Extra St	0
7	18	Doyle St	2
8	19	Oxford Street	7
9	18	Dublin St	4
10	21	First Street	6
11	18	Sixth Street	1
12	19	North ST	3
13	18	Lake St	2

Table 5. *Data with inconsistent labelling.*

View Solutions for answers.

Tip 4: Letter Case

Text may be lowercase, uppercase (all capital letters), or proper case (only the first letter of each word capitalized). Text can be converted to lowercase for email addresses, to uppercase for province abbreviations, and to proper case for names. In this example, the text case is not consistent in the table. Sometimes names and emails are a mix of uppercase, lowercase, and proper case (Table 6).



Tip 04 Exercise: Letter Case

Convert text in the NAME column in Table 6 to proper case. Then convert text in the EMAIL column to lowercase.

HINT: If using Excel, look for UPPER, LOWER, and PROPER functions.

ID	AGE	NAME	EMAIL
1	17	James Smith	JSMITH@GMAIL.COM
2	21	Michael Smith	MSMITH@GMAIL.COM
3	18	Robert SMITH	SMITHR@AOL.COM
4	19	Maria Garcia	mgarcia@hotmail.com
8	18	David SMITH	DAVIDSMITH@GMAIL.COM
12	21	Maria Rodriguez	mariaR@gmail.com
13	18	Mary SMITH	MARYSMITH@GMAIL.COM
14	19	Maria Hernandez	hernandez@outlook.com
15	18	Maria Martinez	mmartinez@mail.com
16	21	James Johnson	james@gmail.com
17	18	Lee Hartman	hartman@mail.com
19	19	Patricia SMITH	SMITHP@MAIL.CA
20	18	Ben SMITH	BENSMITH@MAIL.COM

Table 6. *Data with inconsistent letter case.*

View Solutions for answers.

Tip 5: Spaces and Non-Printing Characters

Spaces and non-printing characters can cause unexpected results when you run any type of sort, filter, and/or search function. Leading, trailing, and multiple embedded spaces or non-printing characters are invisible. They can sneak in when you import data from web pages, Word documents, or PDFs.

Tip 6: Numbers and Signs

There are two issues to watch for:

1. data may include text
2. negative signs may not be standardized

You may obtain a dataset with variables defined as strings (these may include numbers, letters, or symbols). Numeric functions, such as addition or subtraction, cannot be used on string variables, so in order to run any sort of quantitative data analysis, you'll need to convert values in string format to numeric values. Looking at the table of bird sightings (Table 7), there is a column indicating whether the bird is a juvenile. To run quantitative data analysis, you'll need to convert string values of "no" to the numeric value of 0 and string values of "yes" to the numeric value of 1. Leave the original JUVENILE column for reference and create another column with the numeric values. The original column is used to verify the transformation of the new column and is deleted once the transformation is confirmed to be correct. For this example, the new column, JUVENILE _NUM, contains the numeric values of the string values from the JUVENILE column (Tip 06 Exercise).

Numbers can be formatted in different ways, especially with finance data. For example, negative values can be represented with a hyphen or placed inside parentheses or are sometimes highlighted in red. Not all these negative values will be correctly read by a computer, particularly the colour option. When cleaning data, choose and apply a clear and consistent approach to formatting all negative values. A common choice is to use a negative sign.

Tip 06 Exercise: Numbers and Signs

Create a new column named JUVENILE_NUM as part of Table 7. Record a value of 0 in the JUVENILE_NUM column when "no" appears in the JUVENILE column. Record a value of 1 in the JUVENILE_NUM column when "yes" appears in the JUVENILE column.

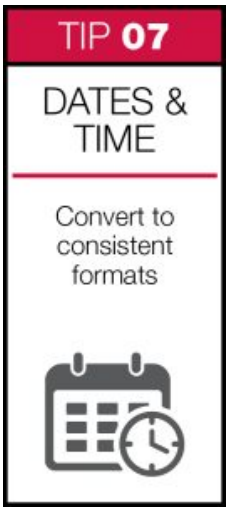
ID	BIRD	LOCATION	JUVENILE
1	robin	Quebec St	no
2	swallow	Cork Street	yes
3	crow	Moffatt St	no
4	pigeon	Victoria Street	no
5	crow	Steffler St	no
6	crow	Extra St	yes
7	robin	Doyle St	yes
8	robin	Oxford Street	no
9	crow	Dublin St	no
10	pigeon	First Street	no
11	pigeon	Sixth Street	yes
12	pigeon	North ST	no
13	swallow	Lake St	yes

Table 7. *Data in string format.*

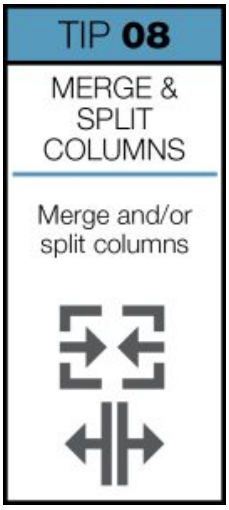
View Solutions for answers.

Tip 7: Dates and Time

There are many ways to format dates in a dataset. Sometimes dates are formatted as strings. If date data are needed for analysis, then at a minimum, change the field type from "string" to "date" so dates are recognized in the analysis tool of choice. With time values, you will need to select a convention and use it throughout the dataset. For example, you can choose to use either the 12- or 24-hour clock to define time in a dataset, but whichever you choose, you should be consistent throughout. You may also need to change the format to ensure that all dates and times are formatted in the same way.



Tip 8: Merge and Split Columns



After closer inspection of the newly acquired dataset, there may be a chance to either (1) merge two or more columns into one or (2) split one column into two or more. Retain the original columns used to merge or split the columns. Then, use the original columns to verify the transformation of the new column and delete the original once the transformation is confirmed to be correct. For example, you may want to split a column that contains a full name into a first and last name (Table 8). Or you may want to split a column with addresses into street, city, region, and postal code columns. Or the reverse might be true. You may want to merge a first and last name column into a full name column or combine address columns.

Tip 08 Exercise: Merge and Split Columns

In Table 8, split the NAME column into two, for first and last names.

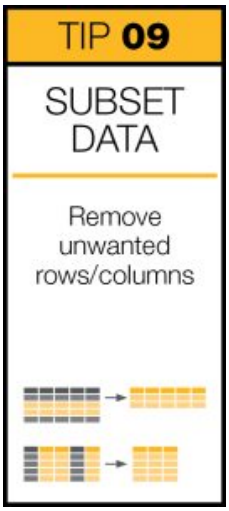
HINT: If using Excel, look for functions to "Combine text from two or more cells into one cell" and "Split text into different columns."

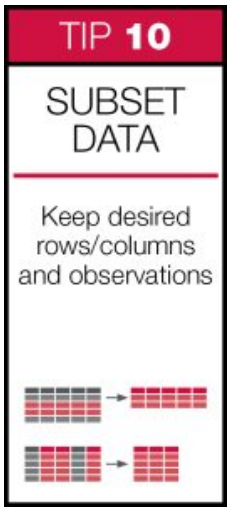
ID	AGE	NAME	EMAIL
1	17	James Smith	jsmith@gmail.com
2	21	Michael Smith	msmith@gmail.com
3	18	Robert Smith	smithr@aol.com
4	19	Maria Garcia	mgarcia@hotmail.com
8	18	David Smith	davidsmith@gmail.com
12	21	Maria Rodriguez	mariar@gmail.com
13	18	Mary Smith	marysmith@gmail.com
14	19	Maria Hernandez	hernandez@outlook.com
15	18	Maria Martinez	mmartinez@mail.com
16	21	James Johnson	james@gmail.com
17	18	Lee Hartman	hartman@mail.com
19	19	Patricia Smith	smithp@mail.ca
20	18	Ben Smith	bensmith@mail.com

Table 8. *Data with columns that may be split.*

View Solutions for answers.

Tip 9: Subset Data





Sometimes data files contain information that is unnecessary for an analysis, so you might want to create a new file containing only variables and/or observations of interest, which will involve selectively removing unwanted columns and/or rows. In this example, the researcher removed the JUVENILE column (Table 9). Or you may need to investigate only certain observations in the file, so you can delete rows in the dataset. In this table, all swallow observations will be deleted. One advantage of this type of cleaning is that programs will run more quickly because the data file is smaller.

Tip 09 Exercise: Subset Data

Create a subset of data in Table 9 to include only observations of juveniles (JUVENILE = 1).

HINT: As always, it is important to keep a copy of your original data.

ID	BIRD	LOCATION	JUVENILE
1	robin	Quebec St	1
2	swallow	Cork St	0
3	crow	Moffatt St	1
4	pigeon	Victoria St	1
5	crow	Steffler St	1
6	crow	Extra St	0
7	robin	Doyle St	0
8	robin	Oxford St	1
9	crow	Dublin St	1
10	pigeon	First St	1
11	pigeon	Sixth St	0
12	pigeon	North St	1
13	swallow	Lake St	0

Table 9. Subset data.

View Solutions for answers.

Cleaning data is an important activity that focuses on removing inconsistencies and errors, which can impact the accuracy of models. The process of cleaning data also provides an opportunity to look closer at the data to determine whether transformations, recoding, or linking additional data is desired.

4. Enriching Data

Sometimes a dataset may not have all the information needed to answer the research question. This means you need to find other datasets and merge them into the current one. This can be as easy as adding geographical data, such as a postal code or longitude and latitude coordinates; or demographic data, such as income, marital status, education, age, or number of children. Enriching data improves the potential for finding fuller answers to the research question(s) at hand.

It's also important to verify data quality and consistency within a dataset. Let's take a closer look at validating data, so that the models provide accurate results.

5. Validating Data

Data validation is vital to ensure data are clean, correct, and useful. Remember the adage by Fuechsel — "garbage in, garbage out." If the incorrect data are fed into a statistical analysis, then the resulting answers will be incorrect too. A computer program doesn't have common sense and will process the data it is given, good or bad, and while data validation does take time, it helps maximize the potential for data to respond to the research question(s) at hand. Some common data validation checks include the following:

1. Checking column data types and underlying data to make sure they're what they are supposed to be. For example, a date variable may need to be converted from a string to a date format. If in doubt, convert the value to a string and it can be changed later if need be.
2. Examining the scope and accuracy of data by reviewing key aggregate functions, like sum, count, min, max, mean, or other related operations. This is particularly important in the context of actual data analysis. Statistics Canada, for example, will code missing values for age using a number well beyond the scope of a human life in years (e.g. using a number like 999). If these values are inadvertently included in your analysis (due to 'missing values' not being explicitly declared) any results involving age will be in error. Calculating and reviewing mean, minimum, maximum, etc. will help identify and avoid such errors.

3. Ensuring variables have been standardized. For example, when recording latitude and longitude coordinates for locations in North America, check that the latitude coordinates are positive and the longitude coordinates are negative to avoid mistakenly referring to places on the other side of the planet.

It's important to validate data to ensure quality and consistency. Once all research questions have been answered, it's good practice to share the clean data with other researchers where confidentiality and other restrictions allow. Let's take a closer look at publishing data, so that it can be shared with other researchers.

6. Publishing Data

Having made the effort to clean and validate your data and to investigate whatever research questions you set out to answer, it is a key RDM best practice to ensure your data are available for appropriate use by others. This goal is embodied by the **FAIR principles** covered elsewhere in this textbook, which aim to make data Findable, Accessible, **Interoperable**, and Reusable. Publishing data helps achieve this goal.

While the best format for collecting, managing, and analyzing data may involve proprietary software, data should be converted to nonproprietary formats for publication. Generally, this will involve plain text. For simple spreadsheets, converting data to **CSV** (comma separated values) may be best, while more complex data structures may be best suited to XML. This will guard against proprietary formats that quickly become obsolete and will ensure data are more universally available to other researchers going forward. This is discussed more in chapter 9, "A Glimpse Into the Fascinating World of File Formats and Metadata."

If human subject data or other private information is involved, you may need to consider anonymizing or de-identifying the data (which is covered in chapter 13, "Sensitive Data"). Keep in mind that removing explicit reference to individuals may not be enough to ensure they cannot be identified. If it's impossible to guard against unwanted disclosure of private information, you may need to publish a subset of the data that is safe for public exposure.

For other researchers to make use of the data, include documentation and **metadata**, including documentation at the levels of the project, data files, and data elements. A data dictionary outlines the names, definitions, and attributes of the elements in a dataset and is discussed more in chapter 10. You should also document any scripts or methods that have been developed for analyzing the data.

Data Cleaning Software

OpenRefine (https://openrefine.org/) is a powerful data manipulation tool that cleans, reshapes, and batch edits messy and unstructured data. It works best with data in simple **tabular formats**, like spreadsheets (CSV), or **tab-separated values files (TSV)**, to name a few. OpenRefine is as easy to use as an Excel spreadsheet and has powerful database functions, like Microsoft Access. It is a desktop application that uses a browser as a graphical interface. All data processing is done locally on your computer. When using OpenRefine to clean and transform data, users can facet, cluster, edit cells, reconcile, and use extended web services to convert a dataset to a more structured format. There's no cost to use this **open source** software and the source code is freely available, along with modifications by others. There are other tools available for data cleaning, but these are often costly, and OpenRefine is extensively used in the RDM field. If you choose

to use other data cleaning software, always check to see if your data remain on your computer or are sent elsewhere for processing.

Exercise: Clean and Prepare Data for Analysis using OpenRefine

Go to the "Cleaning Data with OpenRefine" tutorial and download the Powerhouse museum dataset, consisting of detailed metadata on the collection objects, including title, description, several categories the item belongs to, provenance information, and a persistent link to the object on the museum website. You will step through several data cleaning tasks.

Conclusion

We have covered the six core data cleaning and preparation activities of discovering, structuring, cleaning, enriching, validating, and publishing. By applying these important RDM practices, your data will be complete, documented, and accessible to you and future researchers. You will satisfy grant, journal, and/or funder requirements, raise your profile as a researcher, and meet the growing data-sharing expectations of the research community. RDM practices like data cleaning are crucial to ensure accurate and high-quality research.

Key Takeaways

- Data cleaning is an important task that improves the accuracy and quality of data ahead of data analysis.
- Six core data cleaning tasks are discovering, structuring, cleaning, enriching, validating, and publishing.
- OpenRefine is a powerful data manipulation tool that cleans, reshapes, and batch edits

messy and unstructured data.

Reference List

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods*, *2*(2), 131.

Lidwell, W., Holden, K., & Butler, J. (2010). *Universal principles of design, revised and updated: 125 ways to enhance usability, influence perception, increase appeal, make better design decisions, and teach through design*. Rockport Publishers.

About the author

Lucia Costanzo

Lucia Costanzo is the Research Data Management (RDM) Librarian at the University of Guelph. She recently completed a secondment at the Digital Research Alliance of Canada (the Alliance) as the Research Intelligence and Assessment Coordinator. As part of this role, Lucia coordinated the activities of the Research Intelligence Expert Group, which included informing and advising the Alliance RDM Team and Alliance management on emerging developments and directions, both nationally and internationally, in RDM and broader Digital Research Infrastructure ecosystems. Before the secondment, Lucia actively supported, enabled, and contributed to the learning and research process on campus for over twenty years at the University of Guelph. Email: lcostanz@uoguelph.ca | ORCID: 0000-0003-4785-660X

8.

FURTHER ADVENTURES IN DATA CLEANING: WORKING WITH DATA IN EXCEL AND R

Dr. Rong Luo and Berenica Vejvoda

Learning Outcomes

By the end of this chapter you should be able to:

1. Explain general procedures for preparing for data cleaning.
2. Perform common data cleaning tasks using Excel.
3. Import data and perform basic data cleaning tasks using the R programming language.

Introduction

Data cleaning is an essential part of the research process. In the previous chapter, you were introduced to some common, basic data cleaning tasks. In this chapter, we will delve more deeply into data exploration, manipulation, and cleaning using some flexible general-purpose research tools. The tools that we highlight are, in some cases, the same tools researchers use for analyzing their data, and it's helpful for curators and data managers to be familiar with them.

General Procedures to Prepare for Data Cleaning

Without preparation before the data cleaning process, you may run into critical issues such as data loss. In this section, we will discuss general steps that you should take before the data cleaning process.

Making a Backup

Research data management (RDM) practices recommend creating secure backups of your data to ensure that if it is incorrectly altered during the cleaning process, the original data can always be restored. This backup copy of the original data should not be modified in any situation. You should also keep a record/log of the changes you make. You would be surprised by how many researchers create errors in their original data while they are trying to "improve" it. If a data analyst needs to access the original data, you should either send or share a copy of the data or allow read-only access to the original data.

Understanding the Data

The first step in cleaning data is understanding the data that is being cleaned. To understand the data, begin by doing some basic data exploration (or **Exploratory Data Analysis**) and get a sense of what problems, if any, exist within the data. Check the data values against their definitions in the **metadata** file or documentation for issues such as out-of-range or impossible values (e.g., negative age or age over 200). Ensure you have and understand workable data column names. Check **delimiters** that separate values in text files and ensure your data values don't embed the delimiter itself. If you didn't number your observations, you should add a unique record number to individual observations within the dataset so that you can easily find problem records by referencing that number.

Planning the Cleaning Process

Data cleaning must be done systematically to ensure all data is cleaned using the same procedures. This ensures data integrity and allows data to be easily processed during analysis. To create a plan for cleaning a specific field in a dataset, ask yourself the following three questions:

- What is the data you are cleaning?
- How will you identify an issue within the dataset that should be cleaned?
- How should it be cleaned?

Choosing the Right Tools

One of the most important stages of data cleaning is choosing the right tool for a specific purpose. The previous chapter highlighted **OpenRefine**, a handy, special-purpose data-cleaning tool. Here, we discuss Excel and R, two powerful, general-purpose software tools, and highlight a few of the data cleaning features of each.

Data Cleaning Tools

The data cleaning tool you choose will depend on factors including your computing environment, your level of programming expertise, and your data readiness requirements. You have a wide variety of software and methods to choose from for cleaning and transforming data. We'll review Excel/Google Sheets and R programming language.

Microsoft Excel/Google Sheets

Excel and Google Sheets are great tools for data cleaning and contain a variety of built-in automated data cleaning functions and features. Excel is widely available for both Windows and MacOS as a desktop program, and Google Sheets is available online. They are similar and easy to learn, use, and understand. They can both import and export the commonly used **CSV data file** format and other common spreadsheet formats. When exporting, be sure to check the exported data column names for usability, as some statistical packages will have issues if column names contain embedded spaces or special characters. Common data cleaning techniques used in Excel and Google Sheets for editing and manipulation are summarized in the Function table below.

Table 1. ***Excel functions.***

Function	**Description**
= CONCATENATE	Combines multiple columns
= TRIM	Removes all spaces from a text string except for single spaces between words
= LEFT	Returns the first character or characters in a text string, based on the number of characters you specify
= RIGHT	Returns the last character or characters in a text string, based on the number of characters you specify
= MID	Returns a specific number of characters from a text string, starting at the position you specify, based on the number of characters you specify

Function	Description
= CONCATENATE	Combines multiple columns
= LOWER	Converts a text string to all lower case letters
= UPPER	Converts a text string to all upper case letters
= PROPER	Converts a text string to proper case so that the first letter in each word is upper case and all other letters are lower case
= VALUE	Converts text string to numeric
= TEXT	Converts numbers to text
= SUBSTITUTE	Replaces specific text in a text string
= REPLACE	Replaces part of a text string, based on the position and number of characters specified, with a different text string
= CLEAN	Removes all non-printable characters from a text string
= DATE	Returns the number that represents the date in Microsoft Excel date-time code
= ROUND	Rounds a selected cell to a specified number of digits
= FIND	Returns the starting position of one text string within another text string. FIND is case-sensitive
= SEARCH	Returns the number of the starting position of a specific character or text string within another text string, reading left to right (not case-sensitive)

To understand some of these functions, we will consider a number of common errors seen in imported data, including line breaks in the wrong place, extra spaces or no spaces in and between words, improperly capitalized or all upper case/lower case text, ill-formatted data values, and non-printing characters.

	A	B	C
1		**CONCATENATE & TRIM**	
2			
3	**Data Imported**	**Formula Used**	**Results**
4		=CONCATENATE(A5, A6, A7)	University ofWindsor
5	University	=TRIM(CONCATENATE(A5, A6, A7))	University ofWindsor
6	of	=CONCATENATE(TRIM(A5), TRIM(A6), TRIM(A7))	UniversityofWindsor
7	Windsor	=CONCATENATE(TRIM(A5)," ",TRIM(A6)," ",TRIM(A7))	University of Windsor

Figure 1. *The CONCATENATE and TRIM functions with original and cleaned content side by side.*

Figure 1 illustrates combinations of CONCATENATE and TRIM nested in various ways to find the best output configuration for how you want the text to appear.[1] It is an example of how you can generate a single line of text from the contents of three rows by nesting two Excel functions. CONCATENATE will merge the three cells into one, but it does nothing about the extra spaces you see in the text. TRIM will remove all spaces except for single spaces between words, but it won't add needed spaces, so we need to add quotation marks for Excel to add the needed blank spaces in between words.

	A	B	C
8	**LEFT, RIGHT, MID**		
9			
10	**Data Imported**	**Formula Used**	**Results**
11	BUS256XD	=MID(A11,4,3)	256
12	DRT578XC	=MID(A12,4,3)	578
13	SACR1373	=RIGHT(A13,4)	1373
14	KINE5301	=LEFT(A14,4)	KINE

Figure 2. *The LEFT, RIGHT, and MID functions with original and cleaned content side by side.*

The LEFT, RIGHT, and MID functions in Figure 2 demonstrate how to process data from certain directions depending on where the text or number you wish to extract are in the string.

Rows 11 and 12 show how to use the MID function to extract numbers from the middle of a text string. The MID function takes three **arguments**: a reference to the string you're working with, the location of the first character you want to extract, and the number of characters you want to extract. So MID(A11,4,3) first looks up the contents of cell A11 and finds the string "BUS256XD," and then returns three characters starting with the fourth character: 256. The data in C11 and C12 are the results of using the MID function in rows 11 and 12.

The LEFT and RIGHT functions only take two arguments: the string and the starting point. These functions then return the rest of the string, going either left or right. C13 and C14 show portions of course numbers that have been extracted from A13 and A14 using the RIGHT and LEFT functions.

1. The original spreadsheet files for each figure in this chapter are available in an accessible format in Borealis.

	A	B	C
15	**FIND & SEARCH**		
16			
17	**Data Imported**	**Formula Used**	**Results**
18	INtrOducTion	=FIND("o",A18)	11
19	to	=FIND("o",A19)	2
20	COMputers	=FIND("o",A20)	#VALUE!
21	INtrOducTion	=SEARCH("o",A21)	5
22	to	=SEARCH("o",A22)	2
23	COMputers	=SEARCH("o",A23)	2
24	COMputers	=SEARCH("a",A24)	#VALUE!

Figure 3. *The FIND and SEARCH functions with original and cleaned content side by side.*

Figure 3 illustrates the difference between FIND and SEARCH. The FIND function in Excel is used to return the position of a specific character or substring within a text string and is case sensitive. The SEARCH function in Excel also returns the location of a character or substring in a text string. Unlike FIND, the SEARCH function is not case sensitive. Both FIND and SEARCH return the #VALUE! error if the specific character or substring does not exist within the text.

	A	B	C
25	**UPPER, LOWER, PROPER**		
26			
27	**Data Imported**	**Formula Used**	**Results**
28	INtrOducTion	=UPPER(A28)	INTRODUCTION
29	INtrOducTion	=LOWER(A29)	introduction
30	join smith	=PROPER(A30)	Join Smith

Figure 4. *The UPPER, LOWER, and PROPER functions with original and cleaned content side by side.*

Figure 4 shows how the UPPER, LOWER, and PROPER functions are used to produce the contents for data. The UPPER function changes all text to upper case. The LOWER function changes all text to lower case. The PROPER function changes the first letter in each word to upper case and all other letters to lower case, which is useful for fixing names.

	A	B	C
31	VALUE & TEXT		
32			
33	**Data Imported**	**Formula Used**	**Results**
34	12345	=VALUE(A34)	12345
35	ABCD	=VALUE(A35)	#VALUE!
36	12345	=TEXT(A36,"00000")	12345
37	12345	=TEXT(A37,"0000000")	0012345

Figure 5. *The VALUE and TEXT functions with original and cleaned content side by side.*

Excel aligns strings in a column based on how they are stored: text (including numbers that have been stored as text) are left aligned, numbers are right aligned. In Figure 5, the VALUE function converts text that appears in a recognized format (such as a numbers, dates, or time formats) into a numeric value. If text is not in one of these formats, VALUE returns the #VALUE! error. The TEXT function lets you change the way a number appears by applying format codes, which is useful in situations where you want to display numbers in a more readable format. But keep in mind that Excel now "thinks" of the number as text, so running calculations on it may not work or may lead to unexpected results. It's best to keep your original value in one cell, then use the TEXT function to create a formatted copy of the number in another cell.

	A	B	C
38	SUBSTITUE & REPLACE		
39			
40	**Data Imported**	**Formula Used**	**Results**
41	Time	=SUBSTITUTE(A41,"t","l")	Time
42	tuttle	=SUBSTITUTE(A42,"t","b")	bubble
43	tuttle	=SUBSTITUTE(A43,"t","b",1)	buttle
44	The dog	=SUBSTITUTE(A44,"the","a")	The dog
45	tuttle	=REPLACE(A45,3,2,"*")	tu*le

Figure 6. *The SUBSTITUTE and REPLACE functions with original and cleaned content side by side.*

Figure 6 illustrates how the SUBSTITUTE function replaces one or more text strings with another text string. This function is useful when you want to substitute old text in a string with a new string. However, it is not case sensitive. For example, in cell A41, the function will not substitute "t" for the "T" in "Time". There is a difference between the SUBSTITUTE and the REPLACE functions. You can use SUBSTITUTE when you want to replace specific characters wherever they occur in a text string, and you can use REPLACE when you want to replace any text that occurs in a specific location in a text string.

	A	B	C
46	**CLEAN**		
47			
48	**Data Imported**	**Formula Used**	**Results**
49	□Research data Management▯	=CLEAN(A49)	Research data Management
50	▯line break	=CLEAN(A50)	▯line break
51	▯line break	=SUBSTITUTE(A51, CHAR(127),"")	line break
52			

Figure 7. *The CLEAN function with original and cleaned content side by side.*

The CLEAN function shown in Figure 7 removes non-printable characters, such as carriage returns (↩) or other control characters, represented by the first 32 codes in 7-bit ASCII code from the given text. Data imported from various sources may include non-printable characters and the CLEAN function can help remove them from a supplied text string. In Excel, a non-printing character may show up as a box symbol (□). Note that the CLEAN function lacks the ability to remove all non-printing characters (e.g., "delete character"). You can specify an ASCII character using the Excel function CHAR and the number of the ASCII code. For example CHAR(127) is the delete code. To remove a non-printing character, you can simply substitute the offending non-printable character with nothing enclosed in quotation marks ("").

The exercise below shows the CHAR(19) non-printing character in row 10, which looks like "!!".

Exercise 1

Generate the cleaning results in column B from data imported in column A by using Excel/Google Sheets functions.

	A	B
1	**Data Imported**	**Results**
2	The	The school of social work
3	school of	
4	social work	
5		
6	3 or more	3
7		
8	to the	To The
9		
10	‼Monthly report‼	Monthly report
11		
12	Time	time
13		
14	SACR-126	126
15		
16	789	00789

View Solutions for answers.

R Programming Language

While spreadsheet software like Excel and Google Sheets provide common functions that can assist with data cleaning, it can be very difficult to use them to work with larger datasets. Additionally, if Excel or Google Sheets do not have a specific built-in function, you will require a considerable amount of programming to build that function. The R program can help. R is one of the most well-known, freely available statistical software packages that can be used in the data cleaning process. R is a fully functional programming language with features for working with statistics and data, but you don't need to know how to program to use some basic functions.

The two most important components of the R language are objects, which store data, and functions, which manipulate data. R also uses a host of operators like +, -, *, /, and <- to do basic tasks. To work with R you type commands at a prompt, represented by ">".

To create an **R object**, choose a name and then use the less-than symbol followed by a minus sign to save data into it. This combination looks like an arrow, <-. For example, you can save data "1" into an object "a". Wherever R encounters the object "a," it will replace it with the data "1" saved inside, like so:

```
> a <- 1
```

R comes with many functions that you can use to do sophisticated tasks. For example, you can round a number with the round function. Using a function is pretty simple. Just write the name of the function and then the data you want the function to operate on in parentheses:

```
> round (3.1415)
[1] 3
```

R packages are collections of functions written by R's developers. You may need to install other packages (dependencies, packages that other packages are dependent on) up front to get it to work. It is easier to just set up dependencies = TRUE when you install R packages.

> **install.packages("package name", dependencies = TRUE)**

Let's begin with downloading and installing the required software. R is available for Windows, MacOS, and Linux.

1. Install R and RStudio

R can be downloaded here: https://cran.rstudio.com/.

- For Windows users, https://cran.rstudio.com/bin/windows/base/.
- For Mac users, https://cran.rstudio.com/bin/macosx/.
- For Linux users, https://cran.r-project.org/bin/linux/.

Base R is simply a **command-line tool**: you type in commands at a prompt and see the results displayed on the screen. RStudio, on the other hand, is an **integrated development environment (IDE)**, a set of tools including a script editor, a command prompt, and a results window, as well as some menu commands for commonly used R functions. When people talk about working in R, they usually mean using R through RStudio. Please note that to use RStudio you will need to install R first.

Download and install RStudio Desktop, which is also free and available for Windows, Mac, and various versions of Linux here: https://posit.co/download/rstudio-desktop/.

2. Become familiar with RStudio

Before importing any data, you want to become familiar with RStudio.

The R studio has four quadrants (see Figure 8):

Table 2. ***Purposes of the quadrants in R Studio.***

Section	**Purpose**
Top Left	This section shows you the script(s) you are currently editing. An R script is a set of R commands and comments. They are commonly used to keep of track of the commands that need to be run and provide explanatory notes on the purpose of the commands through comments.
Top right	The "Environment" tab lists all the variables and functions defined and used in a session.
	The "History" tab lists all the commands typed in the R Console (bottom left of RStudio).
	The "Connections" tab can help you connect to an external database to access data that is not on your local computer.
Bottom left	The "Console" tab displays a command prompt to allow you to use R interactively, just like you would without RStudio.
	The "Terminal" tab opens a system shell to perform advanced functions, such as accessing a remote system.
Bottom right	The "Files" tab lets you keep track of, open, and save files associated with your R project.
	The "Plots" tab shows graphs being plotted.
	The "Packages" tab allows you to load and install packages to add additional R functions.
	The "Help" tab provides useful information about some functions.
	The "Viewer" tab can be used to view and interact with local web content.

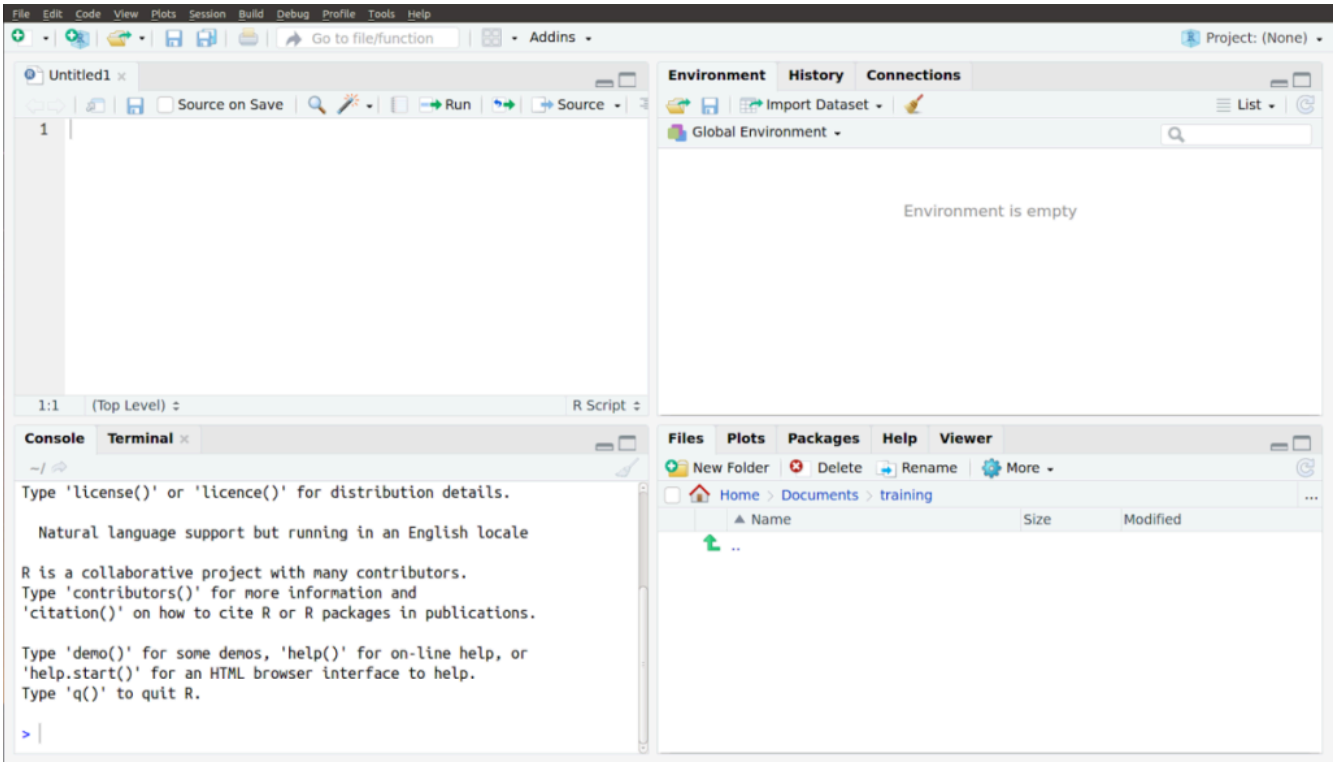


Figure 8. *R Studio quadrants.*

You can type R commands in the console at a prompt, just as you would if you were working without RStudio. You can then view the results in the "History" and "Environment" tabs. Your work is not automatically saved when R is closed. You can copy and paste your console to a text file to save it.

For example, you can open RStudio and type the following command in the console (the text after the prompt ">"):

```
> print("Hello")
```

R will return the following output:

```
[1] "Hello"
```

In addition to working interactively by typing commands at the prompt, you can also create R script files using the RStudio editor shown in the top right quadrant. **Script files** are text files containing a sequence of R commands that can be run one after another. You can select your commands in the script files and run them one at a time or all together. Writing and saving your commands for data cleaning in a script file allows you to better track your work, and you can more easily rerun code later and across new datasets. Tracking your work in this way is a good RDM practice.

Open a new script by selecting the top left icon:

Before importing your dataset, you should set your working directory to the dataset's location. From RStudio, use the menu to change your working directory to the directory where you saved the sample data file. In the "Session" menu, choose > Set Working Directory > Choose Directory.

Alternatively, in the console window (or script editor), you can use the R function setwd(), which stands for "set working directory". Forward slashes (/), rather than backslashes, are in the path. So, if you saved the data to "C:\data", you would enter the command:

```
> setwd("C:/data")
```

3. Importing data

You can import data in different formats using R. CSV files are commonly used for numeric data. While they look like standard Excel files, they are simply text files with columns separated by commas. You can export CSV data from Excel using "save as," and it is commonly used as a preservation format for data management, as it can be read by many programs.

For the next set of examples, we are going to be working with a sample dataset, sample.csv, that is available in Borealis. Please download and save this dataset to a new folder on your computer. The SPSS and Excel files needed for these examples are also available in Borealis.

To load the CSV file, first create a new script file in the script editor. Type the following command in the script to use R's built-in read.csv command, and then run the script.

```
> mydata_csv<-read.csv("sample.csv")
```

The default delimiter of the read.csv() function is a comma, but if you need to read a file that uses other delimiters, you can do so by supplying the "sep" argument to the function (e.g., adding sep = ';' allows a semi-colon separated file).

```
> mydata_csv<-read.csv("sample.csv", sep=';')
```

Please note that "mydata_csv" in the above command refers to the object (data frame) that will be created when the read.csv function imports the file "sample.csv". Think of the "mydata_csv" data frame as the container R uses to hold the data from the CSV file.

R commands follow a certain pattern. Let's go through this one from right to left. In above command, mydata_csv<-read.csv("sample.csv", sep=';'), read.csv is a function to read in a CSV file and has two parameters. The first, sample.csv, tells read.csv what file to read in, while the second, sep=';', tells it that the data points in the file are separated by semi-colons. After read.csv parses the file, the assignment operator, <-, assigns the data to mydata_csv, which is an object (data frame) created to hold the data. You can now use and manipulate the data in the data frame.

In R, <- is the most common assignment operator. You can also use the equal sign, =. For more information, use the help command, which is just a question mark followed by the name of the command:

```
> ?read.csv
```

To read in an Excel file, first download and install the readxl package. In the R console, use the following command:

```
> install.packages("readxl")
```

Please note that sometimes packages are dependent on other associated packages to function properly. By using existing packages, programmers can save time when creating new functionalities by using existing functions that were already implemented. However, it may be difficult to figure out if a package requires another package to function. As such, it is good practice to install packages by including the dependencies statement (TRUE tells R the dependencies should be included). By setting the dependencies parameter to TRUE, we tell R to also download and install all the required packages needed by the package that we are trying to install.

```
> install.packages("readxl", dependencies = TRUE)
```

After the package downloads and installs, use the library() function to load the readxl package.

```
> library(readxl)
```

Note that with the library function, unlike the install.package function, you do not put the name of the package in quotes.

Now you can load the Excel file with the read_excel() function:

```
> mydata_excel <- read_excel("sample.xlsx")
```

For more information, use the **?read_excel** help command.

Statistical Package for the Social Sciences (SPSS) SAV files can be read into R using the haven package, which adds additional functions to allow importing data from other statistical tools.

Install haven by using the following command:

```
> install.packages("haven", dependencies = TRUE)
```

After the package downloads and installs, use the library() function to load the package:

```
> library(haven)
```

Now you can load the SPSS file with the read_sav() function:

```
> mydata_spss <- read_sav("sample.sav")
```

You can import SAS and Stata files too. For more information, use the ?haven or ?read_sav help commands, or visit the https://haven.tidyverse.org/.

Data can also be loaded directly from the internet using the same functions as listed above (except for Excel files). Just use a web address instead of the file path.

```
> mydata_web <- read.csv(url("http://some.where.net/data/
sample.csv"))
```

Now that the data has been loaded into R, you can start to perform operations and analyses to investigate any potential issues.

4. Inspecting data

R is a much more flexible tool for working with data than Excel. We will cover the most basic R functions for examining a dataset.

Assume that the following text file with eight rows and five columns is stored as sample.csv.

```
1,4.1,3.5,setosa,A
2,14.9,3,set0sa,B
3,5,3.6,setosa,C
4,NA,3.9,setosa,A
5,5.8,2.7,virginica,A
6,7.1,3,virginica,B
7,6.3,NA,virginica,C
```

```
8,8,7,virginica,C
```

Now consider the following command to import data. Since the dataset does not contain a header (that is, the first row does not list the column names), you should specify header=FALSE. If you want to manually set the column names, you specify the col.names argument. In the command below, we asked read.csv to set the column names to ID, Length, Width, Species, and Site. Using colClasses, we can specify what data type (number, characters, etc.) we expect the data contained in the columns to be. In this case, we have specified to read.csv that it should treat the first and last two columns as factor (categorical data) and the middle two columns as numeric (numbers).

Enter the following command into your script editor and run it:

```
> mydata_csv <- read.csv("sample.csv", header = F, col.names =
c("ID","Length", "Width", "Species", "Site"),
colClasses=c("factor","numeric","numeric","factor", "factor"))
```

The data is now loaded into mydata_csv. To view the data we loaded, run the following line:

```
> mydata_csv
```

R will return the data from the file it's read in.

```
 ID Length Width   Species Site

1  1    4.1   3.5    setosa    A
2  2   14.9   3.0    set0sa    B
3  3    5.0   3.6    setosa    C
4  4     NA   3.9    setosa    A
5  5    5.8   2.7 virginica    A
6  6    7.1   3.0 virginica    B
7  7    6.3    NA virginica    C
8  8    8.0   7.0 virginica    C
```

The output above shows five columns of data. The first column specifies the row number and is automatically created by R when the data is loaded. The first row displays the column names that we have specified.

One of the first commands to run after loading the dataset is the dim command, which prints out the dimensions of the loaded data by row and column. This command allows you to verify that all entries have been correctly read by R. In this case, the sample dataset should have eight entries with five columns. Let's run dim to see if all the data are loaded.

```
> dim(mydata_csv)
```

After running the command above, R will output the following:

```
[1] 8 5
```

The output tells us that there are eight rows and five columns in the loaded data. This matches our expectation, so all the data have been loaded.

You can also run the summary command, which gives some basic information about each column in the dataset. The summary command returns the maximum and minimum values, the lower and upper **quartiles** (the lower quartile is the value below which 25% of the data in a dataset fall and the upper quartile is the value above which 75% of the data in a dataset fall), and the median for numeric columns and the frequency for factor columns (the number of times each value appears in a column).

```
> summary(mydata_csv)
```

```
       ID         Length          Width           Species  Site
 1      :1   Min.   : 4.100   Min.   :2.700   set0sa   :1   A:3
 2      :1   1st Qu.: 5.400   1st Qu.:3.000   setosa   :3   B:2
 3      :1   Median : 6.300   Median :3.500   virginica:4   C:3
 4      :1   Mean   : 7.314   Mean   :3.814
 5      :1   3rd Qu.: 7.550   3rd Qu.:3.750
 6      :1   Max.   :14.900   Max.   :7.000
 (Other):2   NA's   :1        NA's   :1
```

From the output, we can see that there are five columns. Since we asked R to read the "Length" and "Width" columns as numeric, it calculated and displayed summary information about the numbers in those columns, such as the minimum, maximum, mean, and quartiles. The information about each column is displayed in rows under the column's name. For example, in the "Length" column, you can see that the minimum value is 4.1, and the maximum is 14.9. The 1st Qu. shows the lower quartile, which is 5.4, and the 3rd Qu. shows the upper quartile, which is 7.55.

The NA's row tells us if there are any missing values. In R, missing values are represented by the symbol NA (not available). From the summary output, there are two missing data: one in the "Length" column and one in the "Width" column.

In the "Species" column, which was read in as factors (or categories), each of the rows displays the frequency with which a value appears in the column. From the output, we can see that there are three instances of setosa, four of virginica, and one of set0sa.

Here it is possible to identify recording errors. Instead of *setosa*, there is one flower mistakenly entered as *set0sa*. This kind of typo is very common when recording data, but it's very difficult to find since zero and the letter "o" appear very similar in most fonts.

R uses a basic function, is.na, to test and list whether data values are missing. This function returns a value of true and false for each value in a dataset. If the value is missing, the is.na function returns a value of "TRUE," otherwise, it will return a value of "FALSE."

```
> is.na(mydata_csv$Length)
[1] FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE
```

Note that the dollar sign ($) is used to specify columns. In this case we are investigating if the column "Length" in the CSV dataset contains missing values. As you can see from the output, there is one missing value, so the function returns "TRUE" for that entry in the column.

Outliers are data points that dramatically differ from others in the dataset and can cause problems with certain types of data models and analysis. For instance, an outlier can affect the mean by being unusually small or unusually large. While outliers can affect the results of an analysis, you should be cautious about removing them. Only remove an outlier if you can prove that it is erroneous (i.e., if it is obviously due to incorrect data entry). One easy way to spot outliers is to visualize the data items' distribution. For example, type the following command, which is asking R to generate a **boxplot**.

```
boxplot(mydata_csv$Length)
```

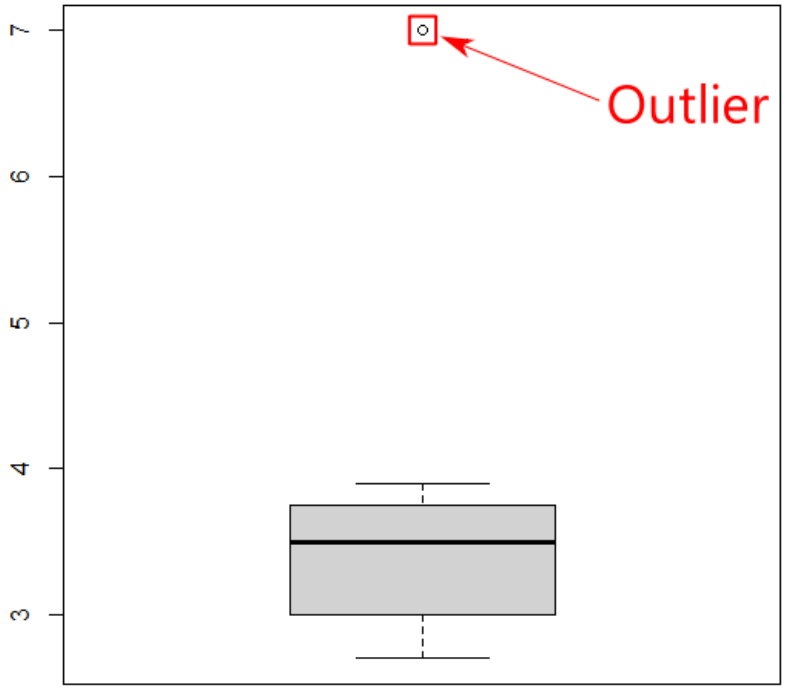


Figure 9. *Outlier in relation to a boxplot.*

Boxplots are useful for detecting potential outliers (see Figure 9). A boxplot helps visualize a quantitative column by displaying five common location summaries: minimum, median, first and third quartiles (Q1 and Q3), and maximum. It also displays any observation that is classified as a suspected outlier using the interquartile range (IQR) criterion, where IQR is the difference between the third and first quartile (see Figure 10). An outlier is defined as a data point that is located outside the whiskers of the boxplot. In the above boxplot output, the circle at the top represents a data point that is very far away from the rest of the data, which are mostly contained in the "box" of the plot.

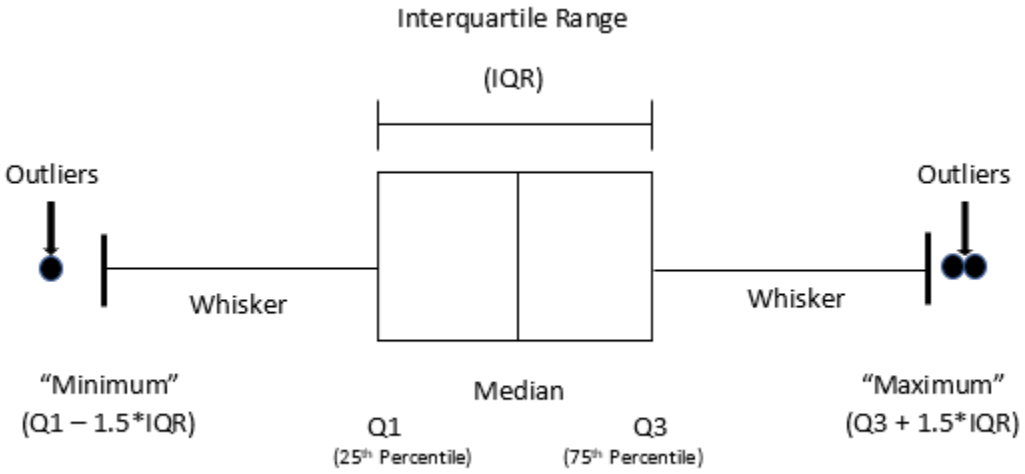


Figure 10. *Interpreting a boxplot.*

Another basic way to detect outliers is to draw a histogram of the data. A **histogram** shows the distribution of the different values in the data. From the histogram below, it appears there is one observation that is higher than all other observations (see the bar on the right side of the plot), which is consistent with the boxplot. The following command will generate a histogram:

```
> hist(mydata_csv$Length)
```

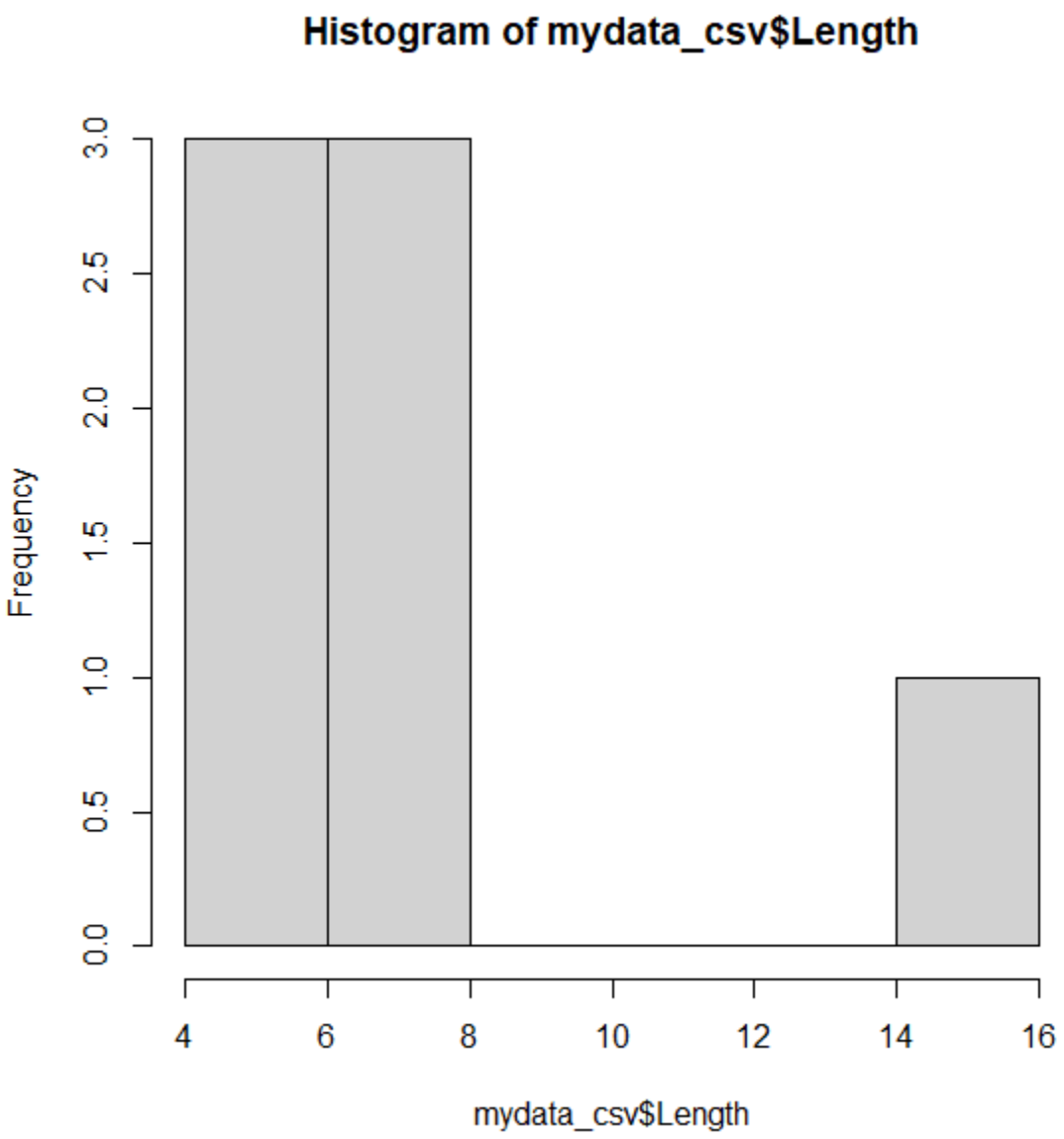


Figure 11. *Histogram of length.*

```
> summary(mydata_csv$Length)

Min.   1st Qu.  Median  Mean    3rd Qu. Max.    NA's
4.100  5.400    6.300   7.314   7.550   14.900  1
```

From the summary output, one value, 14.90, for length seems unusually large, although not impossible according to common sense. This requires further investigation. Such outliers can significantly affect data analysis, so it's important to understand their validity. Removing outliers must be done carefully since outliers can represent real, meaningful observations rather than recording errors.

Now that we have done some preliminary inspection on the raw data, suppose the raw data have several issues that need to be fixed. These issues include:

- Redundant "Site" column
- Typo in the "Species" column
- Missing values in the "Length" and "Width" columns
- Outlier in the "Length" column

With these issues in mind, we can move on to the next stage and start cleaning the data.

5. Cleaning the data

First, let's start by dropping an extra column. Using the data above, suppose we want to drop the "Site" column. As seen in the output of the summary command, "Site" is the fifth column in the dataset. To drop this, we can run the following command:

```
> mydata_csv <- mydata_csv[-5]
```

The above command uses the square brackets to specify the columns of the original data. By using a negative number, we tell R to retrieve all columns except for the specified column. In this case, the "Site" column is the fifth column. Since we want to remove the fifth column but keep all other columns, we can use -5 in the square brackets to tell R to fetch all columns except for the fifth column. Then, by reassigning the new data that we fetched to mydata_csv, we've effectively deleted the fifth column.

To verify that the column has been successfully removed, we can use the dim command, as seen before.

```
> dim(mydata_csv)
[1] 8 4
```

From the output, we can see that the data has four (versus five) columns now.

Next, it's time to clean the typos. In this case, since you know that the typo is *set0sa* and it should actually be *setosa*, you can replace all matching cells using the following command:

```
> mydata_csv[mydata_csv=="set0sa"] = "setosa"
> summary(mydata_csv)
       ID         Length           Width            Species
 1      :1   Min.   : 4.100   Min.   :2.700   set0sa   :0
 2      :1   1st Qu.: 5.400   1st Qu.:3.000   setosa   :4
```

```
3       :1   Median : 6.300   Median :3.500   virginica:4
4       :1   Mean   : 7.314   Mean   :3.814
5       :1   3rd Qu.: 7.550   3rd Qu.:3.750
6       :1   Max.   :14.900   Max.   :7.000
(Other):2   NA's   :1         NA's   :1
```

Note that the equality operator, "==", selects all instances of "set0sa" (in this case, one instance) and "=" assigns it to be "setosa". Using two equals signs to test for equality and a single equals sign to set something equal to something else is a common programming convention.

You can see that there are now zero entries in the "Species" column with the name *set0sa* from the summary() output. The data have been cleaned for this typo.

There are several ways to deal with missing data. One option is to exclude missing values from analysis. Prior to removing NA from the "Length" column, the mean() function returns NA as follows:

```
> mean(mydata_csv$Length)
[1] NA
```

This is done because it is impossible to use NA in a numeric analysis. Using na.rm to remove the missing value NA returns a mean of 7.314286:

```
> mean(mydata_csv$Length, na.rm = T)
[1] 7.314286
```

Exercise 2

Check if there are any outliers in the "Width" column of sample.csv by using a boxplot, then calculate the mean of length by removing the outliers.

View Solutions for answers.

Conclusion

Data cleaning procedures are important foundations for successful data analysis and should be performed before analyzing data. In this chapter, we have only scratched the surface of data cleaning issues and fixes that researchers need in order to create clean data using Excel/Google Sheets and R language. Extensive libraries of data manipulation functions exist, and they offer functionalities that might help you in your data cleaning process. Additional R documentation can be found at the following sites: https://cran.r-project.org/manuals.html and https://cran.r-project.org/web/packages/available_packages_by_name.html.

Key Takeaways

- General procedures in preparing for data cleaning are making a backup, understanding the data, planning the cleaning process, and choosing appropriate tools.
- Excel functions can be used to perform many basic data cleaning tasks.
- The R programming language is a useful and free software package that can be used for more advanced cleaning procedures.

Acknowledgment

The authors thank Kristi Thompson and other editors whose constructive comments have improved this chapter.

About the authors

Dr. Rong Luo

Dr. Rong Luo is a learning specialist in the Academic Data Center at the University of Windsor's Leddy Library. Rong's research interests have been focused on statistical modelling, missing data imputation, social survey data analysis, and information literacy skill assessment. She uses both quantitative and qualitative designs for her research projects.

Berenica Vejvoda

Berenica Vejvoda is the Research Data Librarian at Leddy Library where she is responsible for the coordination and management of Research Data Services. She is also responsible for strategic direction and implementation of research data management services for the University of Windsor as part of a campus-wide initiative. Berenica also serves as the Academic Director for the University of Windsor's Branch Statistics Canada Research Data Centre. Berenica's research interests focus on social determinants of health for marginalized populations with an intersectional lens as well as data inclusivity principles as applied to research data management.

9.

A GLIMPSE INTO THE FASCINATING WORLD OF FILE FORMATS AND METADATA

Émilie Fortin

Learning Outcomes

By the end of this chapter you should be able to:

1. Understand what a sustainable file format is.
2. Properly choose a file format that meets your needs.
3. Understand the usefulness of metadata.
4. Identify the different types of metadata.

Introduction

The research data lifecycle always includes a preservation stage, sometimes referred to as archiving or retention. This stage is linked to data reuse because no one can reuse damaged or inaccessible data. The chapter, "Digital Preservation of Research Data," addresses the issue of digital preservation; this chapter focuses on two elements that enable data retrieval and reuse: file formats and metadata.

File Formats

Pre-assessment

Answer the following questions as honestly as possible (Yes, No):

- Are you having trouble opening files that you created more than ten years ago?
- Do you think that ten years from now you will have difficulty opening files you created this year?
- Do you think a PDF file is a perfect preservation format?
- Do you wake up at night wondering if your great-grandchildren will still have digital photos of you?
- Do you love interactive apps and want all your projects to be as connected as possible?

If you answered yes to more than two questions, this section should help you.

What is a File Format?

Digital file formats are designed according to predefined rules that outline their structure and organization. These principles are usually listed in a specification document that provides details on the subdivisions, encoding, and internal relationships that allow a format to be constructed and validated. A format specification indicates the boundaries between **bit sequences**. These bit sequences can represent, for example, a character, an operation to be performed (machine instruction), or a colour selection.

In summary, a file format is a specific and conventional series of 1s and 0s used to recognize a format.

From the moment you use a computer media, no matter what you use it for, keep in mind that you are using, creating, or modifying formats.

What is a Sustainable Format?

No format is truly sustainable. Those that are deemed acceptable for long-term preservation are formats that remain accessible over time despite technological developments. A good format today can become obsolete in two, five, or ten years.

Here are some criteria for judging the sustainability of a format:

- complexity
- backwards compatibility
- encoding
- dependency
- openness
- metadata
- property
- usage
- evolution
- protections

Complexity. A format must provide good capabilities, without being too complex, or it will be difficult to maintain over time due to its many features. The complexity of a format can be defined by its readability by humans, its level of compression, and the variety of its functionalities. The more effort needed to decipher a format, the more likely it will not be fully understood.

Backwards compatibility. Is a format known for its backwards compatibility? When a new software version is released, how feasible is it to open formats created with older versions of the software? Are the generations of the same format very different from each other?

> **Interesting fact:** Did you know that Adobe provides backwards compatibility of PDF formats up to version 1.3 (released in 1999) only?

Encoding. In the Western world, formats will likely rely on **ASCII** or **Unicode encoding**. If you use other symbols or non-Latin characters, encoding is important because you want the letters and symbols to display properly no matter who opens your files.

Dependency. This is a question of the format's dependence on its software, but also on a specific technology or hardware, on other files, or on its computer environment. Can the format be opened only by specific software? Is the format a container in which we find other formats (ZIP type compression format, video embedded in a text file, video file with a soundtrack, etc.)? Does the format need to connect to your environment to work (for example, an interactive book that is connected to your phone's camera)?

Resources external to your file can be lost over time, so the more dependencies a format has, the harder it will be to preserve in its current form.

Openness. An **open format** is preferable.

Examples of open formats: Office files with an X (e.g., XLSX, DOCX), PDF, TXT, JPG, PNG, CSV.

Interesting fact: Some **extensions** sometimes hide files in open formats. For example, a script file may have extensions like HTML, XML, SC, but they are actually plain text formats.

Interesting fact: Some open formats have become standards over time. For example, PDF and PDF/a are ISO standards.

Metadata. This refers to the file's internal metadata. Think about the file properties that you can access in a software application and through your operating system.

Identifying a format is a first step but documenting the content and the container as much as possible within the format is also very useful. The more a digital object is documented, the better it can be understood in the years to come. A file format that can embed metadata is advantageous, because if the file no longer opens, it is sometimes possible to retrieve valuable information thanks to its metadata (e.g., title, creator, software used to save the format). For more details on this, please see the "Metadata" section.

Property. A proprietary format belongs to a legal entity. It may or may not be open. Its evolution is controlled by its owner. These formats are generally attached to specific software. When the formats are **non-proprietary**, their evolution is controlled by a community of users and they are for the most part open.

- Examples of non-proprietary formats: MKV, TXT, XML, CSV, PNG
- Examples of proprietary and open formats: Office files with an X (e.g., DOCX, XLSX), PDF, RAR
- Examples of proprietary formats: AutoCAD, PSD, WMA

Usage. If only ten people use a format, even if it is open and non-proprietary, it will disappear. On the other hand, an extremely popular proprietary format is very unlikely to die out in the next few years.

If a closed, proprietary format is adopted as a standard by a library, archive, or research community, there is a good chance that the format will live on thanks to its popularity. However, its development needs to be closely monitored.

Evolution. The format should follow a continuous improvement cycle but avoid excess. Systems change, and software and formats must evolve; a static format is not necessarily better than a format that is in development. However, releasing a series of new versions of a format within a limited time frame can be unwise, as frequent changes threaten long-term accessibility.

Protections. There are several technical file protection measures. For example, encryption and the use of a password are good methods for protecting sensitive data, but they are not compatible with long-term preservation. Just imagine the impact that losing a password can have!

Similarly, certain measures to protect the intellectual property of a file, such as locks on e-books, may compromise access to content.

> **Interesting fact:** Some platforms allow for restricted access to files by applying permissions checks. This method is far preferable to locking the files themselves.

How to Choose a Format for a Research Project?

The criteria that define a sustainable format are important, but it is essential to choose them in ways that meet your project needs. It is not necessary to comply with all the criteria. Also, if your area of research requires you to use a format that does not meet any sustainable format criteria, you don't need to refrain from using it; just be aware that there will be an impact on data preservation.

Here are some questions you can ask yourself to help you choose the best format:

- Do you need to preserve your data long term? If you plan to delete all your data in five years and not share it, think only of your own immediate usage needs.
- If you use research instruments/equipment, do you have a choice of format? If so, try to opt for a sustainable format if doing so would have no impact on your research.
- Is the data's appearance or layout important, or just the data itself? If the data layout is not important, you can opt for a simpler format. For example, a textual document stored as a PDF helps preserve the look and feel of a document, but content reuse is complex. However, if the text document is converted to TXT format, the formatting is lost, but the content can easily be reused.
- Are the data independent or linked to other data? If your data are linked to equations or other files, you must preserve those links.
- Do you need to control for file size? If you are limited on space, you may not have a choice but to opt for compression. Try using **lossless compression**.
- In your discipline, is there a format that is used by most of your colleagues and that is considered essential?

In some cases, it is possible to keep data both in its original format and in a sustainable format, but this duplication must have a purpose. For example, your data may serve two very different communities that do not use the same level of technology. However, you should avoid the confusion that two versions of the same dataset could cause.

Another option could be to keep only the original format and to generate lighter copies of the files when necessary. This option is risky in the sense that it involves a dependency on software to read the original format.

You should also keep in mind that unreadable data in ten years will no longer be useful to anyone, including yourself.

Most national libraries publish a list of recommended formats. I've included a number of these lists in the Additional Resources section of this chapter; it may be useful to consult them. The lists include some of the formats that are generally accepted as sustainable in 2023.

Databases

A database involves values, but also a structure and relationships between values. The most commonly used databases at the time of writing are Microsoft Access, Oracle, MySQL, and PostgresSQL. When looking at long-term preservation of databases, one must assess future needs: is the database still in use? Will the preservation of values alone be sufficient? Must the structure of the database and relationships between data also be documented?

Databases are complex to preserve given their structure and the evolution of their content. It is important to define needs before choosing a preservation format.

Some recommended formats include:

- Formats with value separators (CSV, TSV, TXT): preserve data, but not relationships or formulas. Especially useful for simple and small databases.
- Database Preservation Format (SIARD 1.0 and 2.0): an open format established for preserving databases but only usable for certain types of databases.
- Lightweight Relational Database Format (SQLITE): a simple format used for relational databases.

Tabular Data

The main challenge with these formats is dealing with formulas, macros, and embedded content. It should also be remembered that exporting a tabulated file to cloud computing software, or vice versa, can cause losses or errors.

Note that SPSS's SAV format is sometimes recommended, although its documentation is unofficial and backwards compatibility is not guaranteed.

Some recommended formats include:

- Data with Delimiters (CSV, TXT, TSV): simple files, but there is a loss of formulas and cell relationships.
- Microsoft Excel (XLSX): documented and open format, but not recommended by some repositories, as it is a complex proprietary format. In some cases it remains unavoidable. If used, be sure to create a file with Office 2013 or later.
- OpenDocument (ODS, FODS): usually associated with LibreOffice, a software suite developed as an open equivalent of Microsoft software. Structure based on XML. Version 1.2 is certified as an ISO standard; version 1.3 has achieved standard status.

Text

A text document can be very simple, but it can also bring about some challenges. For example, using cloud-based word processing software makes collaboration much easier, but exporting these documents to save them locally can sometimes affect their formatting and hyperlink functionalities. Also, you should ask yourself which versions to keep; it is irrelevant to preserve all revisions and comments to a text. A solution would be to preserve some intermediate versions along with the final version.

If the text document contains embedded objects, such as an image or a table, the selected format may vary. The choice of fonts can also affect the preservation of a textual document.

The text might also refer to other documents to help contextualize or better explain the content. These relationships are important and must be maintained.

The most appropriate format is the one that will retain the most functionalities from the original document while allowing for long-term access.

Some recommended formats include:

- OpenDocument (ODT, OTT): usually associated with LibreOffice, a software suite developed as an open equivalent of Microsoft software. Structure based on XML. Version 1.2 is certified as an ISO standard; version 1.3 has achieved standard status.
- Plain text (TXT): no page layout, but easily accessible and does not depend on any program, which is why it is highly recommended for **README files**.
- PDF and PDF/a: common format, often used for long-term preservation. Ideally, make sure to only keep versions 1.3 and later.
- Electronic Publication (EPUB): an open format, widely used for digital publishing.

> **Interesting fact:** Commercial EPUB files may contain built-in protections to protect intellectual property by preventing copying and sharing. These digital locks are incompatible with long-term preservation.

Images

Most digital preservation experts agree on the most secure image formats to use. The formats mentioned below are raster files; that is, they consist of a series of dots called pixels.

The quality of a format can vary according to several factors such as resolution (the best known), but also colour space or colour depth. Often, the higher the quality of an image, the larger the file.

The RAW proprietary format is not recommended for long-term preservation. Conversely, an image created with a compressed format (e.g., GIF, JPG, BMP) could be preserved as is. Ultimately, technological, human and financial needs and resources need to be assessed before choosing an image format.

Some recommended formats include:

- Tagged Image File Format (TIFF): most used format for preserving images, but heavy.
- Joint Photographic Experts Group 2000 (JP2): lighter than TIFF, but less widely used.
- Joint Photographic Expert Group (JPG): widely used, but the image is compressed.
- Portable Network Graphics (PNG): uses lossless compression. Fairly commonly used, but not always supported by software.

Audio

An audio format is a container with one or more audio data streams.

Several characteristics need to be considered that will influence the rendering and authenticity of the sound: channels, compression, number of bits per sample, number of samples per second, etc. If the original file is already compressed (e.g., MP3, AAC), it may not make sense to migrate it to another format.

Note that MP3 is a compressed format not generally recommended for long-term preservation, but its widespread adoption makes it a fairly reliable format if the original file was created that way.

Some recommended formats include:

- Free Lossless Audio Codec (FLAC): file with lossless compression, lighter format than WAVE.
- PCM WAVE (WAV): quality format used by several national libraries during digitization.
- Broadcast WAVE (BWF): allows the addition of metadata in the files.
- Ogg Vorbis (OGG): open format with better compression than MP3, but less popular.

Video

Video formats are complex, ever-changing, and there is no consensus on any one format in the digital preservation community.

Video formats are generally containers with images or streams of video and sound data. Several characteristics (e.g., colour, compression, sound) can influence their long-term preservation. More than one format can be used for a project depending on the different project goals or outputs, which could range from video creation, to editing, to distribution.

The biggest challenge is balancing file weight and file quality.

Some recommended formats include:

- MP4 with H.264: compressed format mainly used for broadcasting; very widespread.
- QuickTime (MOV) or uncompressed Audio Video Interleaved (AVI) 4:2:2: very heavy formats, but good quality.
- Matroska with FFV1 codec (MKV): standardized format not overly compressed.
- Material Exchange Format with JPG 2000 (MXF): recommended by some national libraries, well documented, but little used by the public.
- Digital Picture Exchange (DPX): very heavy format used when digitizing film stock.

Geospatial Data

Geospatial data are also covered in the chapter, "Geospatial Research Data in Canada: An Overview of Various Projects." These data usually consist of a series of files that complement each other. They can be intrinsically linked to the geographic information system that uses them. Metadata, coordinate referencing systems, and coordinate precision (i.e., how close an observed and recorded value is to the actual value) must be preserved with the data.

Listing recommended formats for the long-term preservation of geospatial data is almost impossible given their complexity (ie., several types of different structures, many proprietary formats). There is no consensus on this and keeping the original format may be the best solution.

Some recommended formats include:

- Geospatial Tagged Image File Format (GEOTIFF): an open format that allows geographic coordinates to be added to an image.
- Geographic Markup Language (GML): an open format based on a standard, but it is complex.
- Keyhole Markup Language (KML, KMZ): XML language that can be associated with several other files that must also be archived (avoid using hyperlinks). Open and widely used format.
- ESRI Shapefile (SHP SHX, DBF, PRJ, SBX, SBN): proprietary, but open and widely used format.

Digging Deeper: How to Identify a Format?

To identify a file format, it is usually sufficient to look at the final section of the file name, which is its extension. For example, the file "my-notes.xlsx" is an Excel file while "my-photo.jpg" is an image. This method has limitations since an extension can be modified, voluntarily or not, or it may be completely unknown. Some operating systems are even configured by default to hide the extension of files, which can complicate the task.

The best way to identify a format is by using its **signature**. A file signature is a series of bits that are strung together in a predictable fashion at the beginning, end, or at both ends of a file.

A tool like PRONOM, widely used in the digital preservation community, works by saving the start and end signatures of a file (known as *Beginning of File* (BOF) and *End of File* (EOF)). This allows a user to retrieve the unique identifier of a format. As an example, the signature x-fmt/398 identifies JPG version 2.0. Knowing a format will be helpful to those who want to view datasets and better understand how to open them.

Some file format identification tools include:

- PRONOM: http://www.nationalarchives.gov.uk/pronom/
- Siegfried: https://www.itforarchivists.com/siegfried
- FIDO: https://github.com/openpreserve/fido or https://fido-js.glitch.me/

Tools that allow viewing files in hexadecimal code:

- HexEd.it: https://hexed.it/
- Literate-binary: https://github.com/marhop/literate-binary

Metadata

Pre-assessment

Answer the following questions as honestly as possible (Yes, No):

- Do you understand what "data about data" means?
- Do you know that there is more than one type of metadata?
- Do you know that some metadata are automatically written into your files?
- Do you know that your brother-in-law could appear as the author of a file you created when using his computer?
- Do you realize the power of metadata?

If you answered no to more than two questions, this section should help you.

An Introduction to Metadata

Metadata are pieces of information used to describe the content or container of a resource. They can be structured or not.

To understand what metadata are, let's start with an example of raw data:

> CCTTTATCTAATCTTTGGAGCATGAGCTGGCATAGTTGGAACCGCCCTCAGCCTCCT
> CATCCGTGCAGAACTTGGACAACCTGGAACTCTTCTAGGAGACGACCAAATTACAA
> TGTAATCGTCACTGCCCACGCCTTCGTAATAATTTTCTTTATAGTAATACCAATCATG
> ATCGGTGGTTTCGGAAACTGACTAGTCCCACTCATAATCGGCGCCCCCGACATAGCA
> TTCCCCCGTATAAACAACATAAGCTTCTGACTACTTCCCCCATCATTTCTTTTACTTC
> TAGCATCCTCCACAGTAGAAGCTGGAGCAGGAACAGGGTGAACAGTATATCCCCCTC
> TCGCTGGTAACCTAGCCCATGCCGGTGCTTCAGTAGACCTAGCCATCTTCTCCCTCC
> ACTTAGCAGGTGTTTCCTCTATCCTAGGTGCTATTAACTTTATTACAACCGCCATCAA
> CATAAAACCCCCAACCCTCTCCCAATACCAAACCCCCCTATTCGTATGATCAGTCCT
> TATTACCGCCGTCCTTCTCCTACTCTCTCTCCCAGTCCTCGCTGCTGGCATTACTAT
> ACTACTAACAGACCGAAACCTAAACACTACGTTCTTTGACCCAGCTGGAGGAGGAG
> ACCCAGTCCTGTACCAACACCTCTTCTGATTCTTCGGCCATCCAGAAGTCTATATCC
> TCATTTTAC

Raw data from research, devoid of metadata, are interesting, but not meaningful to most people. It is easy to see that there is a large gap between the raw data extracted during a research project and their meaning and, thus, usability for humans.

If a geneticist wants to describe the raw data above, she could add the following description, which would be the first level of metadata:

> >Seq1 [organism=*Carpodacus mexicanus*] *C. mexicanus* clone 6b actin (act) mRNA, partial cds

A second level of metadata would be the description of the dataset that this sequence is a part of: genetic sequencing, in this case, of Carpodacus mexicanus, a species of bird.

> This is a nucleotide sequence of *Carpodacus mexicanus* (clone 6b). (A = Adenine, G = Guanine, C = Cytosine, T = Thymine: nucleic acid bases).

A third level of metadata would make it possible to better characterize the previous metadata by standardizing the nomenclature used, which will facilitate search and retrieval in other resources, such as article databases or institutional repositories:

- House finch – Genetics
- Nucleotide sequence

A fourth level would link this metadata to other relevant information, such as an image.

Carpodacus mexicanus QC by Simon Pierre Barrette is licensed CC BY-SA 3.0.

The main goal of metadata is to describe and enable retrieval. Any metadata present should facilitate the tasks performed when using general or academic search engines, which are:

- Finding: finding resources that match the search criteria.
- Identifying: to establish the context of the data and to confirm that the resource that is described corresponds to the resource that is sought or to be able to distinguish between two or more resources with similar characteristics.
- Selecting: selecting a resource that is relevant to the needs of the searcher.

Metadata necessary for preservation are those that ensure the authenticity and long-term accessibility of digital resources and that allow recovered files to be accessible, readable, and intelligible. Metadata need to be managed and discovered independent of the systems they were created with.

Metadata Normalization

Some metadata can be standardized, such as the names of those responsible for a research project, the methods of data collection and analysis, variable titles, subjects covered by the research, as well as temporal or geographical coverage. Other types of metadata will adopt less precise description rules. They aim to standardize the display of the resource being described. This includes, for example, the title attributed to a research project or an abstract describing a dataset.

The more metadata are standardized, the more they contribute to the **FAIR principles** (detailed further in chapter 2, "The FAIR Principles and Research Data Management") and the more they allow for the Findability, Accessibility, Interoperability, and Reuse of the resources they represent. When describing a resource, whether it is data or a dataset, it is necessary to select the most useful metadata to maximize time and effort.

Several methods can be used to standardize metadata. However, there is often terminological confusion, as certain terms are used to incorrectly describe varying concepts.

Metadata Schemas

To fully understand **metadata schemas**, imagine an online form with empty fields to fill in. The schema hides behind the form and gives meaning to the information added to each field.

Some schemas specify the syntax with which elements should be encoded, while others, such as **Dublin Core** and **Data Documentation Initiative (DDI)**, only provide fields for storing information without giving any indication as to how the content should be entered or its syntax.

Let's take the house finch as an example. A birdwatcher wants to enter a sighting of the bird in a repository that uses the Darwin Core metadata schema. He will need to fill in the following fields:

Fields to fill	**Darwin Core elements behind the scenes**
Time of sighting	eventDate
Observer	identifiedBy
Scientific name	scientificName
Kingdom	kingdom
Class	class
Order	order

Family	family
Genus	genus

There are many schemas, some are general while others are disciplinary. A standardized schema which is widely used can be **machine-readable**, which increases the visibility of data and the possibility of its reuse. However, these advantages are lost when creating an in-house metadata schema.

In summary, a metadata schema serves as a structure and a container for information about datasets and, to some extent, adds to its meaning.

Description Rules

Description rules make it possible to standardize, normalize, and structure information relating to datasets. These rules will prescribe the transcription of information, the use of capital letters, as well as element syntax or order. Rules are schema independent and can be used in any data repository.

To illustrate, let's use the example of the finch-enthusiast birdwatcher. He wants to know if this species has been sighted in his area on a specific date. He looks up three repositories that use the Darwin Core schema. Searching with the date October 10, 2021, he finds results in only one of the repositories. Why? Because repositories use different description rules for dates. One has no requirements, being the repository where the October 10, 2021 entry is found; the other asks for the ISO 8601 standard, which is YYYY-MM-DDTHH:MM:SSZ and where the date is indicated as 2021-10-10; and the last one requires the form DDMMYYYY and where the desired entry is represented by 10102021.

Clear description rules are also very useful for personal names, especially in the case of common names. It's important to avoid the use of initials, homonyms, or pseudonyms. Depositing data gives visibility to researchers, but to do this, it is necessary to be able to identify, without ambiguity, the person responsible for the data.

A name is sometimes not enough to distinguish between people, and this is why it is recommended to also use **Persistent Unique Identifiers (PID)**, like **ORCiD**.

Controlled Vocabularies

Controlled vocabularies standardize indexing and make it easier to find and locate information. It is a set of terms recognized, standardized, and validated by a group or community of practice used to index or analyze the content of a resource.

If several terms refer to the same concept, only one of them will be chosen and identified as the "preferred term;" all others, considered as possible synonyms, will be mentioned as "rejected terms."

Let's go back to the birdwatcher who, this time, is looking for information on the finch in an English-language data repository. The data in this repository is indexed with free vocabulary, but also with FAST (Faceted Application of Subject Terminology). To retrieve all the information on the species, the birdwatcher searches for the term "finch" and discovers that "house finch" is the term chosen by FAST. With this term, he can successfully search the repository and retrieve all available data.

Thesauri and subject heading directories are the most common and well-known examples of controlled vocabularies. There are encyclopaedic vocabularies, but also specialized vocabularies specific to certain disciplines, e.g., ERIC, a thesaurus that specializes in education, or WORMS, a catalogue of the names of marine organisms.

Several of these vocabularies are multilingual, or process linguistic equivalents, which is a valuable contribution for **interoperability**.

Digging deeper: Ontologies

An ontology is a theoretical representation of a domain of knowledge with concepts linked by semantic and logical relations. It includes vocabularies and definitions, and specifies how concepts are interrelated. An ontology makes it possible to establish a set of relations and to describe specific situations in a given domain. It also imposes a structure on the domain and limits the possible interpretations of terms. Put simply, an ontology makes it possible to offer a common language to blocks of information linked to each other. It is to metadata what grammar is to language.

One of the main advantages of using an ontology is the interoperability, reuse, and sharing of metadata. The main difference between an ontology and a controlled vocabulary is that the controlled vocabulary proposes semantic relations between the elements that compose it, while the ontology will propose functional relations making it possible to describe situations precisely.

For example, in a controlled vocabulary, "house finch" is the preferred term. It is related to "Carpodacus," which is the general term, as well as "Mexican finch" and "*Carpodacus mexicanus*," which are two rejected terms. In an ontology, "house finch" could be linked through the relationship "habitat" to the terms "suburb" and "semi-desert." The ontology could also point to the "feeding" relationship to make a link between the finch and other "granivores" and "insectivores."

Types of Metadata

There are various ways of categorizing metadata. In this chapter, the following groupings will be used: descriptive, structural, technical, access, and preservation metadata. The last three types of metadata in the list are less straightforward to understand. They are introduced below for those interested in gaining more advanced knowledge on the topic.

Beyond these categories, metadata can also be classified by their source (internal, external), their mode of creation (manual, automatic), their status (static, dynamic), their structure (structured or not), and other characteristics. For more information on this, please consult the resources at the end of this chapter.

Descriptive Metadata

As their name suggests, descriptive metadata are used to describe a resource's content and ensure that it can be found, whether by humans or by machines. The title of a work, the name of its creator, and the date of creation are examples of descriptive metadata found in data repositories, library catalogues, or databases.

In the case of research data, descriptive metadata generally refer to fields to be filled in data repositories. In addition to metadata, in cases where the data are not deposited in the repository, a text file, such as a README file, can be used to support descriptive metadata.

Project metadata describe the "who, what, where, when, and why" of a dataset, which provides context for understanding the purpose of data collection, methodology, and use.

Dataset metadata are more granular. They describe and contextualize the data in more detail, including, for example, variables, units of measurement, and observations. This information may also be present with the data themselves.

The rules to follow for descriptive metadata are not insignificant. The better a dataset is described, the more it will be identifiable and the easier it will be to attribute credit to the right people. In this sense, the use of unique identifiers such as DOIs and ORCiDs as well as controlled vocabularies such as FAST and its French-language equivalent, RVMFAST, makes it possible to disambiguate people and digital objects. Metadata standardization also supports interoperability between systems.

The best way to harness the power of descriptive metadata is to:

- use unique identifiers where possible.
- use existing metadata schemas well established in your research community.
- standardize metadata where possible (names, subjects, geospatial coordinates, dates, etc.), ideally with

controlled vocabularies.
- follow the advice suggested by repositories for completing their metadata fields, i.e., mandatory fields, recommended fields, and optional fields.

Each discipline uses their own metadata, schemas, ontologies, and controlled vocabularies. For some examples of these particularities, see the chapters "Managing Quantitative Social Science Data" and "Managing Qualitative Research Data."

> **Interesting fact:** Many files have descriptive metadata embedded in their format. Have you ever looked at the file properties attributed by a software application or your operating system? You might be surprised! Sometimes a software application automatically fills in the "author" information with the name of the owner of the software or inserts geographic coordinates into the file of a photo taken with a cellphone!

Structural Metadata

Structural metadata help establish links between and within files. It is as much about the physical structure of a file (the links between different pieces of content) as it is about the logical structure of a document (the links between files). For example, you might have an article in a PDF format and the associated graphics in a different file, in DOCX. You might also have information about where text and images are located on a page, and information about page order.

Some of these metadata are generated automatically, others must be entered manually. They can be useful if you have to switch from a complex format to a simple format and doing so would require breaking down your data. You may need to describe the links between your files to represent the original format. This information can be noted in a text file or by using code.

If your files are not independent or they refer to other files, think about the structural metadata. They will allow you to fully understand your data.

Digging Deeper: Other Types of Metadata

Descriptive and structural metadata are fairly easy to understand, even though their exact definitions may be debatable. However, definitions for technical, access, and preservation metadata are more ambiguous.

Sometimes these metadata are grouped together under the term "administrative metadata." The divisions below are used for explanatory purposes only.

Most of the metadata below are created automatically within files and it is not essential to know them. It is possible to modify some of this internal metadata and indeed, some software applications allow their extraction to keep them separate. However, good knowledge of formats and metadata is recommended before attempting to do this.

As mentioned previously, a format change can be positive for the long-term preservation of files. Such a conversion may impact the file's internal metadata. Extracting these metadata from the original format and keeping them alongside the digital object allows for the provenance and authenticity of the files to be documented.

Technical Metadata

Technical metadata are highly format-specific and mostly always embedded within files. They document the creation of the file (software used, version, operating system, date of creation and last modification, etc.) and the characteristics of the digital objects which vary according to the type of format.

Examples of technical metadata include:

- For text: encoding, structure in XML ...
- For images: resolution, colour profile, encoding depth ...
- For sound: bitrate, codec, sample rate ...
- For video: number of frames per second, colour profile, duration ...
- For web content: format declared in the header, server response collected ...

Extracting technical metadata helps prove that a format is what it claims to be. It provides information about an unknown or corrupted digital object.

Access and Use Metadata

Access and use metadata include information that allows the research community to download data and reuse it legally.

To avoid any rights violations, these metadata provide information on the provenance, the possibilities of access (open access, embargo, confidentiality form, etc.) and of use (free, with citation, read-only, etc.). It may

also include **digital signatures**. These metadata make it possible for repository administrators to carry out preservation actions in a legal manner.

Preservation Metadata

Preservation metadata are usually tied to specific metadata schemas like METS or PREMIS and represent the actions performed on files to preserve them.

They include everything related to the integrity and authenticity of a digital object (see the chapter, "Digital Preservation of Research Data," for more on this topic). Minimally, a **checksum** should be calculated. With preservation metadata, you can trace all changes made to a file such as format changes, checksum checks, and physical media moves, as well as those who made the changes.

Conclusion

The title of this chapter refers to a fascinating world for good reasons. We have only offered a preliminary survey into the world of file formats and metadata. Be assured, however, that it is not essential to master all the secrets of file formats, controlled vocabularies, or metadata schema to ensure accessible and reusable data in the long term.

Reflective Questions

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://ecampusontario.pressbooks.pub/canadardm/?p=149#h5p-6

Key Takeaways

- The choice of a format depends on several factors, but mainly on the needs and capacities of those who use them.
- The best research data cannot be found and understood, including by those who created them, without quality metadata. Quality is preferred over quantity.
- View formats and metadata as allies and not obstacles; you may find they are, at times difficult, but always reliable friends!

Additional Readings and Resources

Corti, L., Van den Eynden, E., Bishop, L., Woollard, M., Haaker, M., & Summers, S. (2019). *Managing and sharing research data: a guide to good practice* (2nd ed., vol. 1). Sage.

Formats

Canadian Resources

Bibliothèque et Archives nationales du Québec. (2020, March). *Guide concernant les formats recommandés par BAnQ.* https://numerique.banq.qc.ca/patrimoine/details/52327/4076856

Bieman, E., & Vinh-Doyle. W. (2019). *National heritage digitization strategy – Digital preservation file format recommendations*. Government of Canada, Canadian Heritage Information Network. https://www.canada.ca/en/heritage-information-network/services/digital-preservation/recommendations-file-format.html

Library and Archives Canada. (2022). *Guidelines on file formats for transferring information resources of enduring value*. https://library-archives.canada.ca/eng/services/government-canada/information-disposition/guidelines-information-management/Pages/guidelines-file-formats-enduring-value.aspx

Library and Archives Canada. (n.d.). *File format guidelines for preservation and long-term access version 1.0.* https://www.councilofnsarchives.ca/sites/default/files/

LAC%20File%20Format%20Guidelines%20for%20Preservation%20and%20Long-term%20v1_2010-12_0.pdf

Other Resources

Bibliothèque nationale de France. (n.d.). *Fiches formats*. https://github.com/hackathonBnF/FichesFormat/wiki

Caplan, P. (2008). What Is digital preservation? *Library Technology Reports*, *58*(2). https://journals.ala.org/index.php/ltr/article/view/4224/4809/

Caplan, P. (Ed.). (2010). Digital preservation [Special issue]. *Information Standards Quarterly, 22(*2). https://www.niso.org/sites/default/files/2019-07/ISQ%20Spring%202010.pdf

Centre de coordination pour l'archivage à long terme de document électroniques. (n.d.). *Catalogue des formats de fichiers pour l'archivage*. https://kost-ceco.ch/cms/kad_main_fr.html

Dappert, A. (2016). *Digital preservation metadata and improvements to PREMIS in version 3.0* [PowerPoint Presentation]. https://www.loc.gov/standards/premis/v3/tutorialslides.pdf

Digital Preservation Coalition. (2015). *Digital preservation handbook* (2nd Ed.). https://www.dpconline.org/handbook

Digital Preservation Coalition. (n.d.). *Technology watch publications*. https://www.dpconline.org/digipres/discover-good-practice/tech-watch-reports

Digital Preservation Coalition, & Artefactual System. (2021). *Preserving audio*. http://doi.org/10.7207/twgn21-11

Digital Preservation Coalition, & Artefactual System. (2021). *Preserving databases*. http://doi.org/10.7207/twgn21-06

Digital Preservation Coalition, & Artefactual System. (2021). *Preserving documents*. http://doi.org/10.7207/twgn21-07

Digital Preservation Coalition, & Artefactual System. (2021). *Preserving GIS*. http://doi.org/10.7207/twgn21-16

Digital Preservation Coalition, & Artefactual System. (2021). *Preserving moving images*. http://doi.org/10.7207/twgn21-12

Digital Preservation Coalition, & Artefactual System. (2021). *Preserving raster images*. http://doi.org/10.7207/twgn21-13

Digital Preservation Coalition, & Artefactual System. (2021) *Preserving spreadsheets*. http://doi.org/10.7207/twgn21-09

Federal Agencies Digital Guidelines Initiative. (n.d.). *Guidelines, file format comparison projects*. https://www.digitizationguidelines.gov/guidelines/File_format_compare.html

Federal Records Management. (n.d.). *Appendix A: Tables of file formats*. National Archives and Records Administration. https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html

Library of Congress. (n.d.). *Recommended formats statement*. https://www.loc.gov/preservation/resources/rfs/

Loftus, C. (2019, August 23). File format identification: A student project at the University of Sheffield Library. *Digital Preservation Coalition*. https://www.dpconline.org/blog/file-format-identification-sheffi-uni

McLellan, E. P. (2007) *General study 11 final report: Selecting digital file formats for long-term preservation*. InterPARES 2 project. http://www.interpares.org/display_file.cfm?doc=ip2_file_formats(complete).pdf

UK Data Service. (n.d.). *Recommended formats*. https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/recommended-formats/

Vitam. (2020). *Identification des formats de fichier*. https://www.programmevitam.fr/ressources/DocCourante/autres/fonctionnel/20200131_NP_Vitam_preservation-identification-format-v2.0.pdf

Games on Formats

Archives & Records Association. (2022). *File format or fake?* https://www.exploreyourarchive.org/archives/digital-preservation/

Fortin, É., & Ruest, J.-F. (2022). *Mille formats*. Bibliothèque de l'Université Laval. https://www5.bibl.ulaval.ca/formations/tutoriels-en-ligne/autres-tutoriels/mille-formats

Metadata

Baca, M. (Ed.). (2016) *Introduction to metadata* (3e éd.). Getty Publications. http://www.getty.edu/publications/intrometadata/