10 Things for Curating Reproducible and FAIR Research

About this document: This document was developed by the RDA CURE-FAIR Working Group. The final version of this document will be made available under a Creative Commons Attribution License.

Version	Date	Authors
v.1	October 7, 2021	CURE-FAIR co-chairs, Mandy Gooch
CURE-FAIR WG Open comment period: October 7-19, 2021 https://docs.google.com/document/d/1CoiFboiIKj7FFgdKqhzKk6fwK6orHcPybM1UnWYMmt0/edit (or https://bit.ly/3Ft00Id)		
v.2	October 28, 2021	CURE-FAIR co-chairs, Mandy Gooch
Plenary sprint period: October 28-Nov 3, 2021 & Post-plenary extended period until Nov 25, 2021 https://docs.google.com/document/d/14A1sqdxhGAsFIPoyHUglrS89rcLY5aP_xizO2ZckcKA/edit		
v.3	November 28, 2021	CURE-FAIR co-chairs, Mandy Gooch
Revisions based on community comment https://docs.google.com/document/d/152YkCYcNIg_0kPAi7IDTnPDTVRvIJ8D0saqdRzX9Tx4/edit		
v.4	March 22, 2022	Lead Authors: Florio Arguillas, Thu-Mai Christian, Mandy Gooch, Tom Honeyman, Limor Peer (CURE-FAIR WG co-chairs) Contributors: Erin Clary, Christopher Erdmann, Ana Van Gulick, Daniel S. Katz, Katherine E. Koziar, Wanda Marsolek, Peter McQuilton, Qian Zhang, and members of the CURE-FAIR WG
Added content, incorporating community comment https://docs.google.com/document/d/1xNQtSlsCw_eyCim-iKd13x_g-hy9FYdgPKPrCVu1Tc0/edit#		

Table of contents

Introduction

Computational reproducibility is defined as the ability to obtain consistent computational results using the same input data, computational steps, methods, code, and analysis conditions as those used in the original study. Computational reproducibility is a means of making scientific claims more transparent. It is imperative for verifying and building upon reported findings, preserving a complete scientific record, and enhancing pedagogical strategies for research methods training. For various reasons, the computational reproducibility standard has not been adopted as an integral part of normative scientific practice.

Computational reproducibility involves the assembly of a "reproducibility file bundle" or "research compendium" that includes all of the research artifacts (e.g., data, code, documentation) necessary for the computation. Prior to publishing or archiving the research compendium, the objects contained within require curation to ensure they meet quality standards for computational reproducibility.

> ***The object of curation in the context of reproducibility is the scientific claim: We are curating the digital artifacts that underlie the claim.***

Curation for FAIR and reproducible research is the process of reviewing and enhancing a research compendium for long term reuse.

> ***We curate to: a) assess whether computational reproducibility can be achieved using the digital artifacts contained within the research compendium, and b) ensure that the quality of the digital artifacts aligns with the FAIR principles and community standards for long-term archival preservation.***

Both activities are essential aspects of curation in the context of reproducible research. We call these practices Curation for FAIR and Reproducible Research, or CURE-FAIR.

Curation is often carried out near the end of the research lifecycle by data and archive professionals who may also be subject matter experts. However, there are many data and code management actions that other key stakeholders can take earlier in the lifecycle to facilitate the production of FAIR and computationally reproducible research compendia.

Curators can be viewed as the first reusers of the research compendium. Prior to publication or archiving, curators can flag issues with a research compendium that preclude computational reproducibility, then take actions to remedy problems or recommend an appropriate course of action. For example, curators can make sure that software configuration and dependency information is well-documented so that an independent researcher can recreate the computational environment with the proper technical specifications.

> ***This document includes standards-based guidelines for CURE-FAIR best practices in archiving and publishing computationally reproducible studies that rely on quantitative data, primarily in the social sciences. Our hope is that these "10 Things for Curating Reproducible and FAIR Research" will serve as a starting point for the development of curatorial guidelines to extend beyond the specific concerns of the social sciences community and other domains and disciplines that use similar***

> ***methods, and to the particular curatorial concerns and requirements of an archives or publisher.***

Computational reproducibility requires a village. This document is primarily for data curators and information professionals who are charged with verifying that a computation can be executed and that it can reproduce prespecified results. Secondarily, it will be of interest to researchers, publishers, editors, reviewers, and others who have a stake in creating, using, sharing, publishing, or preserving reproducible research.

Additional resources can be found in the CURE-FAIR Zotero Library on Curating for Reproducibility.

10 Things for Curating Reproducible and FAIR Research

We have followed the approach taken by several groups to create a list of “Things” that can serve as stand-alone guides for self-directed inquiry (e.g., Library Carpentry and collaborators’ Top 10 FAIR Data & Software Things; Australian Research Data Commons’ 23 (research data) things; Griffith University Library and eResearch Services’ Reproducible-Research-Things; Briney et al. Foundational practices in RDM).

Our focus is specifically on the key issues of curating reproducible and FAIR research. We build on prior work and only include standards and practices that enhance the publication and archival preservation of FAIR and computationally reproducible research compendia.

In practice, the “Things” are interrelated to some extent, but we try to tease out the fundamental principles. Each “Thing” begins with a section describing the issue followed by three actionable sections: Get started, Learn more, and Go deeper.

The “10 Things for curating reproducible and FAIR research” address the following questions:

Does the research compendium contain everything needed to reproduce a predefined outcome in an organized and parsimonious way?

Thing 1: Completeness: The research compendium contains all of the objects needed to reproduce a predefined outcome.

Thing 2: Organization: It is easy to understand and keep track of the various objects in the research compendium and their relationship over time.

Thing 3: Economy: Fewer extraneous objects in the compendium mean fewer things that can break and require less maintenance over time.

Is descriptive information about the research compendium and its components available and easy to understand?

Thing 4: Transparency: The research compendium provides full disclosure of the research process that produced the scientific claim.

Thing 5: Documentation: Information describing compendium objects is provided in enough detail to enable independent understanding and use of the compendium.

Is information about how the research compendium and its components can be used available and easy to understand?

Thing 6: Access: It is clear who can use what, how, and under what conditions, with open access preferred.

Thing 7: Provenance: The origin of the components of the research compendium and how each has changed over time is evident.

Is information about the research compendium and its components embedded in code?

Thing 8: Metadata: Information about the research compendium and its components is embedded in a standardized, machine-readable code.

Thing 9: Automation: As much as possible, the computational workflow is script- or workflow-based so that the workflow can be re-executed using minimal actions.

Is there a plan for reviewing the research compendium for FAIR and computational reproducibility standards over time?

Thing 10: Review: A series of managed activities needed to ensure continued access to and functionality of the research compendium and its components for as long as necessary.

Thing 1: Completeness

Related FAIR Principles: R1, R1.2, R1.3

Related FAIR4RS Principles: R1, R2, I2

A research compendium contains all of the objects required to reproduce a predefined output.

When curation includes reproducing scientific claims or findings, you will need to check that all of the requisite objects and information are included in the research compendium. This often includes data, code, outputs, a codebook or data dictionary, and supporting documentation.

Completeness ensures re-users of the research compendium are well-informed and have the context necessary to reproduce the results without needing to contact the original producer or track down missing components.

Get started

When preparing a research compendium for reproducibility and curation, the goal should be to include all pertinent information, documentation, and files necessary to successfully run and recreate research output(s). Consider the following questions about each compendium component:

Data

- If specific input data are necessary to reproduce a prespecified outcome, are the data included as a separate digital object?
- If not, has detailed information been provided on where and how to obtain the specific input data, or how to generate them?
- Are the data sensitive or restricted by the data producer? Is there enough information within the compendium to allow re-users to find and request access to these data?
- If input data are available from an external source, is there some promise of permanence? If there is a chance the data will change over time, will re-users be able to access or recreate the version of data used in this study?

Code

- Are computing instructions included as digital objects in the research compendium?
- How are these computing instructions represented (e.g., personal scripts, bespoke manual steps, archived source code repository, a fully automated workflow available via open software libraries)?
- What environment, including hardware and software, does the computation rely on, and are all the packages specified?

Output

- Does the compendium include an output as a separate digital object?
- What is the nature of the output in relation to the other objects in the compendium (i.e., is it a log from a statistical program, an executable paper, new data, new file, table, figure)?

Documentation

- Does the compendium include a codebook or data dictionary that defines the contents of the data file?
- Are variable definitions clear? Are units of measure included where appropriate?
- Does the compendium include a README file that outlines the contents of the compendium and the relationships among them?

Container

- If the data or other objects are not in the container, are there precise instructions for accessing them during computation?
- Does the container require separate maintenance? Under what conditions will the container open?

Learn more

A research compendium should be able to stand on its own, with little interaction between a compendium producer and a re-user. Once deposited into a repository, the compendium should be accessible and reusable. More information on constructing a robust compendium can be found in Thing 5: Documentation and Thing 6: Accessibility.

Academics have been describing research compendiums and the necessary components that comprise a robust compendium for a few years now. Learn more about these recommendations below:

- **Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten simple rules for reproducible computational research. *PLoS Computational Biology*, *9*(10), e1003285. https://doi.org/10.1371/journal.pcbi.1003285**
  The authors present a simple set of rules that can be used to make research more reproducible by enacting changes within your own research habits.
- **Berenson, K. R. (2017). *Managing your research data and documentation*. American Psychological Association. https://www.apa.org/pubs/books/4313048**
  In this book, Berenson outlines and describes how to construct a research compendium with reproducibility in mind.

- **Krafczyk, M. S., Shi, A., Bhaskar, A., Marinov, D., & Stodden, V. (2021). Learning from reproducing computational results: introducing three principles and the Reproduction Package. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2197), 20200069. https://doi.org/10.1098/rsta.2020.0069**
This article takes a slightly different approach of investigating the reproducibility of seven different articles. After performing their checks, the authors developed three principles for making research more reproducible based on the roadblocks they encountered in their own attempts at computation reproducibility.

Go deeper

For examples of good compendiums, please see the following citations from various data repositories:

- **Ansolabehere, Stephen; Kuriwaki, Shiro, 2021, "Replication Data for: Congressional Representation: Accountability from the Constituent's Perspective", https://doi.org/10.7910/DVN/QOVWMM Harvard Dataverse, V1, UNF:6:2J2iCVDbVbiImIApYE8mgw== [fileUNF]**

- **Berry, William; Fording, Richard; Hanson, Russell; Crofoot, Justin, 2021, "Replication Data for: The Validity of the Enns and Koch, and Berry et al. Measures of State Policy Mood: Continuing the Debate", https://doi.org/10.15139/S3/CQTBQU, UNC Dataverse, V1, UNF:6:iKVoC8ysGlWUdf5Gv/Cd9g== [fileUNF]**

- **Decarolis, Francesco, and Rovigatti, Gabriele. Data and Code for: From Mad Men to Maths Men: Concentration and Buyer Power in Online Advertising. Nashville, TN: American Economic Association [publisher], 2021. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2021-09-29. https://doi.org/10.3886/E130502V1**

Thing 2: Organization

Related FAIR Principles: F2, I3, R1.3

Related FAIR4RS Principles: F2, I2, R2, R3

As research projects progress, they often evolve and things get more complicated and opaque over time; memory fails; project team members move on to new ventures. It is difficult to predict what may happen with a project down the line, so it is important to properly organize a compendium throughout the project itself.

Organization is necessary to understand and to keep track of the various artifacts in the research compendium and the relationships among them over time.

Get started

The organization of a research compendium depends on the types of files, their relationship to one another, as well as their interactions within the computational workflow. There is not a single, prescribed way for organizing research artifacts; however, these are some questions to consider as a project is being organized within a compendium:

- How are the objects organized within the research compendium?
- Can you identify what each object is?
- Can you determine the function of each object in relation to each other?
- Are names of files indicative of their content?
- How is the compendium packaged?
- Is the folder structure logical, and does it enhance understanding of the computational workflow?
- Does the current structure permit streamlined reproducibility without errors?
- What if the objects cannot be organized in folders? What changes would you need to make to your scripts to incorporate a step to download (and perhaps unzip) the files prior to analysis and other steps?

Learn more

For more information on recommended organizational best practices and tips, please see the following resources:

- ***Tier protocol 4.0*. TIER Protocol 4.0 | Project TIER | Teaching Integrity in Empirical Research. (n.d.). Retrieved October 22, 2021, from https://www.projecttier.org/tier-protocol/protocol-4-0/.**

The TIER Protocol provides a guided hierarchy with descriptions and information on structuring a research compendium for computational reproducibility and curation.

- **Open Science Framework (OSF). Best practices [File naming, Organizing files, Version control].**
  https://help.osf.io/hc/en-us/sections/360003624133-File-Management-and-Licensing
  The Open Science Framework offers many helpful tips and guides on best practices for sharing data and research workflows.

- ***File management* | *Research Data Management Service Group*. (n.d.). Retrieved December 14, 2021, from https://data.research.cornell.edu/content/file-management**
  The Cornell Research Data Management Service Group has a robust resource for file management recommendations and best practices.

- **Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK. (2017). Good enough practices in scientific computing. PLoS Comput Biol 13(6): e1005510. https://doi.org/10.1371/journal.pcbi.1005510**
  Tips on project organization, including a project layout are described within this article.

Enabling reuse may mean separately handling the "objects" of the reproducible file bundle. This may be true when curating a bundle where the data is available elsewhere. If the data are archived separately from the code, then an additional step will likely need to be incorporated into the processing scripts to fetch the data. While adding extra labor, there are several benefits to this approach. This means that subsequent work can call upon the same data, rather than redepositing the same data with different code.

Explore how the links to data are captured in the landing page of the following datasets:

- **Irino, Tomohisa; Tada, Ryuji (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. *PANGAEA*, https://doi.org/10.1594/PANGAEA.726855, *Supplement to:* Irino, T; Tada, R (2000): Quantification of aeolian dust (Kosa) contribution to the Japan Sea sediments and its variation during the last 200 ky. *Geochemical Journal*, 34(1), 59-93, https://doi.org/10.2343/geochemj.34.59**

In some disciplines, journals require archiving the full compendium in specific archives, which may have certain requirements for organizing the research artifacts being deposited.

- **Willis, C., & Stodden, V. (2020). Trust but Verify: How to Leverage Policies, Workflows, and Infrastructure to Ensure Computational Reproducibility in Publication. *Harvard Data Science Review*, *2*(4). https://doi.org/10.1162/99608f92.25982dcf**

- **Replication Data and Publishing in Public Opinion Quarterly: https://academic.oup.com/poq/article/83/2/309/5557800#163322985**

- **Vilhuber, L. (2020). Reproducibility and Replicability in Economics. *Harvard Data Science Review*, *2*(4). https://doi.org/10.1162/99608f92.4f6b9e67**

Go deeper

A variety of tips, guidance, and online modules have been developed for further training on organization geared towards curation and reproducibility. Here are a few more resources to explore to go deeper:

- **University of Illinois at Urbana-Champaign Research Data Services. Data Nudges:**
  https://emails.illinois.edu/newsletter/158822.html
  This Data Nudge recommends tips for file naming and outlines known issues with using special characters in naming conventions.
  https://emails.illinois.edu/newsletter/160474.html
  Data Nudge #2,2017 highlights useful elements when crafting a file naming convention.

- **File Structure and Naming Prompt Sheet:**
  https://app.box.com/s/xiipln16t84y1ldwp5c7iefd395sypt4
  The File Structure and Naming Prompt Sheet is an adaptation of an MIT Libraries worksheet to help researchers think through their naming conventions before adopting them for a project.

- **Research Data MANTRA. Organising Data. (n.d.). Retrieved December 15, 2021, from https://mantra.ed.ac.uk/organisingdata/**
  Part of the Research Data MANTRA online course, this section provides lessons on naming and re-naming conventions, file and code versioning, as well as cloud collaboration tools. The interactive course gives valuable examples and activities to further educate researchers and curators on data organization best practices.

- **RO-Crate (Research Object Crate). https://w3id.org/ro/crate**
  The intent of RO-Crate is to assist researchers in packaging their data, research objects, and workflows for curation.

Thing 3: Economy

Related FAIR Principles: R1

Related FAIR4RS Principles: R1

When curating a reproducible file bundle, consider any extraneous parts that can be cut to make the overall bundle simpler to streamline computational reproduction.

Economizing everything means fewer research objects can break while also requiring less care and maintenance over time.

Get started

Simplifying and commenting out code are a couple of methods for tackling economization; however, the approach may vary depending on the type of software or methods being used for analysis. During file review and/or curation, here are some questions to consider:

- Can the scripts be simplified by removing redundancies or using loops and functions, for example?
- Are code blocks ordered logically according to the presentation of the results in the publication?
- Is there a master script that groups together all the other scripts? Are there additional scripts outside the master script and if so, are they necessary?
- Are the dependencies to the scripts or code all necessary?
- Are there comments in the code to help understand the computational workflow?
- Are there notebooks?

Learn more

More information on economization in the context of a research compendium can be found in Thing 5: Documentation and Thing 9: Automation. The following resources contain relevant information as well:

- ***Literate Programming*. (n.d.). Retrieved December 15, 2021, from http://www.literateprogramming.com/**
  A collection of best practices and guidance for programming, documentation, and code commenting.

- **Gillespie, C., & Lovelace, R. (n.d.). *Efficient R programming*. Retrieved December 15, 2021, from https://csgillespie.github.io/efficientR/**
  This book covers not only programmer efficiency, but also computational efficiency to write more effective and streamlined code using R.

- **Martin, R. C. (Ed.). (2009). *Clean code: a handbook of agile software craftsmanship*. Prentice Hall. https://enos.itcollege.ee/~jpoial/oop/naited/Clean%20Code.pdf**

A guide on writing clean and concise code covering topics such as good vs. bad commenting, slow code, and formatting.

- **Gentzkow, Matthew and Jesse M. Shapiro. (2014). Code and Data for the Social Sciences: A Practitioner's Guide. University of Chicago. https://faculty.chicagobooth.edu/matthew.gentzkow/research/CodeAndData.pdf, last updated January 2014.**
  Chapter 6, in particular, discusses the three rules of abstraction when writing code to eliminate redundancy and improve readability of the final product.

Go deeper

As mentioned previously, the approach to economizing code is dependent upon many factors. Here are resources specific to a few disciplines that highlight some best practices:

- **Benureau, Fabien C.Y., and Rougier, Nicolas P. (2018). Re-run, Repeat, Reproduce, Reuse, Replicate: Transforming Code into Scientific Contributions. *Frontiers in Neuroinformatics*, 11. https://doi.org/10.3389/fninf.2017.00069.**

- **Palomino, Jenny, Wasser, Leah, and Joseph, Max. (2021). *Earth Lab. Earth Data Analytics*. Intro to Earth Data Science. Section 7 - Write Efficient and Clean Code Using Open Source Python. https://www.earthdatascience.org/courses/intro-to-earth-data-science/write-efficient-python-code/**

- **Battig, W. F. (1962). Parsimony in Psychology. Psychological Reports, 11(2), 555–572. https://doi.org/10.2466/pr0.1962.11.2.555**

There are also tools available to assist with cleaning code. For example:

- **A Tool for Writing Cleaner, More Transparent Code. (n.d.). Retrieved December 15, 2021, from https://docs.ropensci.org/Rclean/**

Thing 4: Transparency

Related FAIR Principles: I3, R1, R1.2, R1.3

Related FAIR4RS Principles: I2, R1, R1.2, R2, R3

Transparency is a precondition of computational reproducibility. Ideally, the research compendium provides full disclosure of the research process that produced the scientific claim.

The inability to reproduce results is often caused by missing or insufficient information about the knowledge production process: where and how the data were obtained or generated, the analytical steps taken to produce results, the technical specifications of the requisite computing environment, the scripts used to process or manipulate the data, and any other information (often taking the form of tacit knowledge) needed to understand and re-execute the analysis. Despite lengthy methods sections in scholarly journal articles, many of these critical “backstage” details are either not included or not within easy reach. This problem of such opaque, or “black box,” research can undermine the integrity of the scientific record when no individual besides the original investigator is able to undertake the analysis process to verify the reported findings.

Transparency requires that details about how the data were captured and transformed, and how they were analyzed to produce published results are included in the research compendium.

Get started

Think about a re-user of the research compendium and the specific information they would need in order to re-execute the analysis to generate expected results. Consider the questions below for each key aspect of research transparency.

Access transparency
Access transparency requires that re-users of the research compendium are provided specific information about the location of the materials, protocols for accessing them, and their origin (i.e., provenance).

- Are explicit instructions provided on where and how to request and access materials not included in the compendium because they cannot be made publicly available due to licensing or sensitive data restrictions or for any other reason?

- For research that involves secondary analysis of extant data or uses existing code, are original sources cited with a persistent identifier or stable URL that resolves to the specific data files used in the analysis?

Production transparency
Production transparency is evident when details about the study design and research methods (e.g., sampling frameworks, response rates, validity testing) are clearly articulated.

- Are key aspects of the research design and analytic methods disclosed to allow for understanding and evaluation of the research process?

- Is information on how data were selected, captured, transformed, and analyzed provided in adequate detail to understand the choices made and steps taken by the researcher?

Analytic transparency
Analytic transparency refers to the disclosure of the analytical steps taken to produce reported results, which for computational analyses should be captured in code files.

- Are the technical specifications of the computing environment, including operating system, software, and packages requirements (and their versions), provided?
- Is the process used to transform raw or original data into the analysis dataset made evident in the script files?
- Do script files contain non-executable comments that explain the purpose of codeblocks to enhance human readability of the analytical workflow?

Learn more

Many scholars have been advancing the cause of transparency by highlighting its positive impact on scientific rigor and integrity while also describing the adverse effects of research that fails to disclose the processes that underpinned the results. Read the articles below to further understand the impetus for increasing transparency as a means to promote research reproducibility.

- **Elman, C., Kapiszewski, D., & Lupia, A. (2018). Transparent social inquiry: Implications for social science. *Annual Review of Political Science, 21*, 29–47. https://doi.org/10.1146/annurev-polisci-091515-025429**
  The authors of this article explain the value and practice of research transparency in terms of transparency's three primary elements--data access, production transparency, and analytic transparency.
- **Stark, P. B. (2018). Before reproducibility must come preproducibility. *Nature, 557*(7707), 613–613. https://doi.org/10.1038/d41586-018-05256-0**
  Philip Stark introduces the term "preproducibility" to describe the desired condition in which adequate information is provided to enable individuals to repeat an experiment or analysis.

Transparent research requires actionable strategies for exposing the knowledge-making processes that are often omitted from the published research article. Learn about some of these proposed strategies from the resources below.

- **Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L., Nosek, B. A., Petersen, M., Sedlmayr, R., Simmons, J. P., Simonsohn, U., & Van der Laan, M. (2014). Promoting transparency in social science research. *Science, 343*(6166), 30–31. https://doi.org/10.1126/science.1245317**
  This article makes the case for transparency in social science research and describes three specific practices--disclosure, registration and analysis plans, and open data and materials--that can increase transparency in scientific reporting.
- **Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich,**

E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. Science, 348(6242), 1422–1425. https://doi.org/10.1126/science.aab2374
This article introduces the Transparency and Openness Promotion (TOP) guidelines, which looks to scholarly journals to promote the principles of transparency, openness, and reproducibility by issuing policies that require authors to engage in transparent research practices.

- **EQUATOR Network. (n.d.).** ***Enhancing the QUAlity and Transparency Of health Research*****.** **https://www.equator-network.org/**
  The EQUATOR Network offers a searchable online database of almost 500 reporting guidelines for various types of research studies along with guidance on how to choose from among the guidelines.

- See also **Thing 5: Documentation** to learn how transparency is made evident by producing tangible documentation that records essential details of the research workflow and **Thing 7: Provenance** to understand the importance of tracking the origin, chain of custody, and ownership of research materials.

Go deeper

These resources go into more depth about research transparency by presenting disciplinary case studies, tools, and initiatives that promote transparency and specific situations in which transparency was crucial for addressing scientific challenges.

- **Jacobs, A. M., Büthe, T., Arjona, A., Arriola, L. R., Bellin, E., Bennett, A., Björkman, L., Bleich, E., Elkins, Z., Fairfield, T., Gaikwad, N., Greitens, S. C., Hawkesworth, M., Herrera, V., Herrera, Y. M., Johnson, K. S., Karakoç, E., Koivu, K., Kreuzer, M., … Yashar, D. J. (2021). The Qualitative Transparency Deliberations: Insights and implications.** ***Perspectives on Politics, 19*****(1), 171–208. https://doi.org/10.1017/S1537592720001164**
  While the focus has been on quantitative research, transparency standards do indeed apply for qualitative research. This article discusses how transparency is operationalized in qualitative research.

- **Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S., & Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency.** ***PLOS Biology, 14*****(5), e1002456. https://doi.org/10.1371/journal.pbio.1002456**
  This article explains how applying badges to transparent research can incentivize scientists to make their research more transparent.

- **The Center for Effective Global Action. (2020, June 23). Why COVID-19 makes research transparency more important than ever.** ***Medium*****. https://medium.com/center-for-effective-global-action/why-covid-19-makes-research-transparency-more-important-than-ever-84071f3586be**
  This article uses the COVID-19 pandemic as a case study for the importance of research transparency.

- **Carsey, T. M. (2014). Making DA-RT a Reality. *PS: Political Science & Politics, 47*(01), 72–77. https://doi.org/10.1017/S1049096513001753**
  Read about how a group of political scientists advocated for research transparency in their discipline by developing the Data Access and Research Transparency (DA-RT) statement, which has been incorporated into the ethics guidelines of the American Political Science Association, and adopted by a large number of political science journals.

Thing 5: Documentation

Related FAIR Principles: F2, R1, R1.2, R1.3

Related FAIR4RS Principles: F2, R1, R1.2, R3

An important strategy for demonstrating research transparency is to create comprehensive documentation that captures important details of the research process so that others are able to follow it to reproduce published research results. Failing to record essential information about the research compendium—descriptions of individual compendium files, technical requirements for using compendium files, the origin of materials used in secondary analyses, dataset variable and value code definitions, analytical steps for generating expected results—relegates the research process to a “black box,” making it difficult or even impossible to reproduce results.

Robust documentation is the outcome of an intentional practice that keeps in mind the information needs of the compendium re-user so that the research compendium and the materials contained within it are *independently understandable for informed reuse*.

Get started

Documentation can take on various forms. Common examples are README files, data availability statements, data dictionaries or codebooks, and code headers and non-executable comments. However the information about compendium components and the analysis workflow is presented, reproducibility relies on the presence of robust documentation that assumes no prior knowledge of the research process, thereby making the tacit more explicit.

README file

When reviewing the contents of a research compendium, it should be clear the contents of each file, the purpose it serves, and citations to the original sources of extant materials. This information is often communicated in a README file that provides the following:

- The purpose of the compendium
- Citation to the associated research publication
- List of files with descriptions of each file
- Citations to original sources of materials used in secondary analyses
- Operating system and software requirements including version information
- Specific step-by-step instructions on how to install, execute, and/or implement components of the compendium to reproduce expected results
- Licensing information that establishes acceptable uses of the materials

Data availability statement

Especially for data and other research artifacts that cannot be made publicly available, it is important to provide information on where and how to access the files in a data availability statement. This document should contain the following:

- Formal data citation that includes a persistent identifier (e.g., DOI) or other stable URL
- Instructions on where and how to access the materials
- Any specific protocols required to access the materials (e.g., IRB approval, data use agreement)
- Duration of any embargo placed on the materials

Codebook

A codebook, also referred to as a **data dictionary**, gives meaning to each variable and value that appears in a dataset. For every dataset in the compendium, a codebook should define each variable by providing:

- Variable names and human-readable variable labels
- Value codes and human-readable value labels for categorical and dichotomous variables
- Variable formats (e.g., string, numeric, date)
- Missing value codes
- Functions or expressions used to generate constructed or derived variables

Code headers and comments

Besides the scripts that execute the analysis workflow, code files should also include header information that identifies the associated study and technical specifications for executing the code, as well as non-executable comments that explain the analysis process. Specifically, code files should contain the following:

- Citation to the associated research publication
- Primary author name and contact information
- Operating system, software, and package dependencies including version information (this may be generated using a software command such as `print sessionInfo` in R or included in a requirements.txt file)
- Character encoding
- Date of most recent code update
- Date of most recent code execution
- Code license
- Non-executable, human-readable comments explaining the function of scripts and how they map to results presented in the publication

Learn more

The resources below offer additional guidance on documentation practices for reproducibility, including how-tos and exemplars.

- **DataONE. (n.d.). *Best practice: Describe*. Data Management Skillbuilding Hub. https://dataoneorg.github.io/Education/bp_step/describe/**
  DataONE's Data Management Skillbuilding Hub offers useful Best Practices Primers that focus on the research data lifecycle. This *Best Practice: Describe* primer outlines best practices to implement when creating comprehensive documentation for the research compendium.

- **Research Data Management Working Group. (n.d.). *README files*. Harvard Longwood Research Data Management. https://datamanagement.hms.harvard.edu/collect/readme-files**
  The Longwood Medical Area Research Data Management Working Group offers useful resources on various research data management topics. The README Files web page describes the purpose and contents of a README file with a template and example included.

- **Social Science Data Editors. (2021, December 15). *Guidance on data citations*. Social Science Data Editors. https://social-science-data-editors.github.io/guidance/addtl-data-citation-guidance.html**
  This guide covers all aspects of creating formal citations, which should be included in the compendium when using secondary materials to produce research results.

- **Cornell University Research Data Management Service Group. (n.d.). Guide to writing "README" style metadata. https://data.research.cornell.edu/content/readme**
  The resource offers comprehensive guidance on writing a README file, including a list of recommended content, formatting standards, and a README document template.

- **Hoces de la Guardia, F., & Sturdy, J. (2019). Code readability & dynamic documentation. In *Best practices for transparent, reproducible, and ethical research*. Inter-American Development Bank, Office of Strategic Development Effectiveness. http://dx.doi.org/10.18235/0001564**
  The "Code Readability & Dynamic Documentation" section of this book describes best practices for documenting code and specific coding suggestions for enhancing understandability.

Go deeper

There are several tools available to assist with creating documentation that is complete, accurate, and understandable. The resources below include software packages, tutorials, and other tools to go one step further toward research transparency and reproducibility.

- **Arslan, R. C. (2019). How to automatically document data with the *codebook* package to facilitate data reuse. *Advances in Methods and Practices in Psychological Science,***

***2*(2), 169–187. https://doi.org/10.1177/2515245919838783**
This article introduces the codebook R package that makes it easier to produce a codebook containing the essential information about the dataset.

- **Rule, A., Birmingham, A., Zuniga, C., Altintas, I., Huang, S.-C., Knight, R., Moshiri, N., Nguyen, M. H., Rosenthal, S. B., Pérez, F., & Rose, P. W. (2019). Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks. *PLOS Computational Biology, 15*(7), e1007007. https://doi.org/10.1371/journal.pcbi.1007007**
  While computational notebooks have been praised for supporting research reproducibility, there are specific practices that should be implemented, which these authors outline as rules for using Jupyter notebooks.

- **Data Documentation Initiative (DDI) Alliance. (n.d.). *Create a codebook*. DDI. https://ddialliance.org/training/getting-started-new-content/create-a-codebook**
  This webpage offers a list of tools for generating a codebook from standardized DDI metadata captured in various statistical software packages.

Thing 6: Access

Related FAIR Principles: F1, F2, F4, A1, A1.1, A1.2, I1, R1.1

Related FAIR4RS Principles: F1, F2, F4, A1, A1.1, A1.2, R1.1

Science is a community endeavor that sees maximum benefits when the materials produced in the course of research activities are made maximally available to community members to enable them to build upon, reuse, and verify scientific knowledge.

Fortunately, trustworthy repositories exist to preserve research artifacts for long-term discovery, access, and use. Trustworthy repositories are equipped with the policies, infrastructure, expertise, and workflows that ensure research compendia submitted to the repository are made publicly available alongside human- and machine-readable information (i.e, metadata) to facilitate access and use.

Ideally, the research compendium and the objects contained within it would be made as open as possible by placing the compendium in the public domain, which would permit anyone to freely use, modify, and redistribute the materials for any purpose. However, ethical and/or legal constraints may make it necessary to define and enforce acceptable uses of the research compendium or any of its component parts. This is best accomplished by assigning a formal license to the materials.

In addition to physical and legal access to the compendium, access is concerned with access to the technology and tools necessary to use the compendium. Requiring the use of costly, proprietary, obscure, or obsolete hardware or software to render compendium files, for example, impedes efforts to reproduce research results.

Because the research compendium serves as the evidence-base for published findings, enabling others to access it with minimal physical, legal, and technical barriers is essential for supporting and promoting computational reproducibility.

Get started

Physical access

Access to the compendium and its components is best preserved when placed in a trustworthy repository. When considering physical access to a compendium, address the following questions:

- Is the research compendium (and all of its component parts) made available in a trustworthy repository that provides long-term archival access to materials without undue burden?
- Is the research compendium assigned a digital object identifier (DOI) or other unique, persistent identifier that allows web access to the compendium, even if its online location should change?
- Is standardized metadata included with compendium materials to facilitate discovery, access, and use of the compendium?

Legal access

Whether a research compendium is placed in the public domain or has use restrictions applied, acceptable and/or conditional uses of the compendium and its individual contents should be declared using precise language in the form of a waiver or machine-readable license that addresses the following:

- Are there any compelling reasons that the compendium cannot be placed in the public domain to maximize access?
- If access to the materials must be restricted, who is permitted to request access?
- For what purposes may the materials be used?
- What types of uses of the materials are prohibited (e.g., commercial uses, modification, redistribution)?
- What are the specific protocols (if any) that one must follow to access and use the materials (e.g., IRB approval, use of a secure computing workstation)?
- What obligations must be fulfilled as conditions for accessing and using the materials (e.g., data citation, funder statement)?

Technical access
As much as possible, the technology (i.e., hardware and software) required to render and use compendium files to reproduce the associated published results should be reasonably accessible by scholars for whom the research is relevant. Determine if this is the case by addressing the following:

- Is required software open-source or in common use by the research community?
- Is comprehensive, up-to-date documentation available to facilitate use of the technology?
- Does the technology have longevity, i.e, is it unlikely that the technology will become obsolete in the near future, making it unusable?
- Are there alternatives that would be suitable for rendering and using compendium files should the technology become difficult or impossible to access? In other words, are file formats hardware-, operating system-, and software-agnostic?

Learn more

The resources below offer a closer look at mechanisms for supporting physical, legal, and technical access to research compendium files as a means of promoting research reproducibility.

Trustworthy repositories
A trustworthy repository has provisions in place to ensure the long-term discoverability, accessibility, and usability of research artifacts. Read more about the criteria repositories must meet to be considered trustworthy, and how to select a repository best suited for the types of materials contained in the compendium:

- **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., Navale,**

V., Petters, J., Sierman, B., Sokolova, D. V., Stockhause, M., & Westbrook, J. (2020). The TRUST Principles for digital repositories. *Scientific Data, 7*(1), 144. https://doi.org/10.1038/s41597-020-0486-7
These authors, who are representatives of the digital repository community, present the TRUST (Transparency, Responsibility, User focus, Sustainability, and Technology) Principles that offer a framework for repository best practices.

- **CoreTrustSeal. (n.d.). *CoreTrustSeal*. CoreTrustSeal. https://www.coretrustseal.org/**
  The CoreTrustSeal website provides a list of repositories that have undergone a thorough audit and review of their practices, services, and infrastructure and earned the CoreTrustSeal indicating trustworthiness.

- **re3data. (n.d.). *re3data.org: Registry of research data repositories*. http://www.re3data.org/**
  Search for a repository using the re3data registry of research data repositories.

Data citation

According to the Joint Declaration of Data Citation Principles developed by the Data Citation Synthesis Group (2014), “Sound, reproducible scholarship rests upon a foundation of robust, accessible data. For this to be so in practice as well as theory, data must be accorded due importance in the practice of scholarship and in the enduring scholarly record.” Learn more about data citation below:

- **Data Citation Synthesis Group. (2014). *Joint Declaration of Data Citation Principles*. Force11. https://doi.org/10.25490/A97F-EGYK**
  The Data Citation Principles presented here emphasize the importance of data access to the scientific enterprise as it outlines the purpose, function, and value of data citations.

- **Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., Ault, L., Bell, P., Bowie, R., Leadbetter, A., Lowry, R., Moncoiffé, G., Harrison, K., Smith-Haddon, B., Weatherby, A., & Wright, D. (2012). Making data a first class scientific output: Data citation and publication by NERC’s environmental data centres. *International Journal of Digital Curation, 7*(1), 107–113. https://doi.org/10.2218/ijdc.v7i1.218**
  This article describes mechanisms for data citation and publication to promote research transparency in the context of NERC’s environmental data centers, though is applicable to many disciplinary contexts more generally.

- **National Research Council. (2012). *For attribution: Developing data attribution and citation practices and standards: Summary of an international workshop*. National Academies Press. https://doi.org/10.17226/13564**
  This National Academies publication summarizes a workshop during which participants addressed various questions on the current state of data citation practices; the importance of data citation; the technical, economic, legal, and cultural issues of data citation; and implementation of data citation practices and standards.

Licensing

Whether a research artifact is placed in the public domain or not, licenses let others know how they can use, modify, or redistribute the materials. The resources below offer more insight into the why and how of licensing in the sciences:

- **Stodden, V. (2009). Enabling reproducible research: Open licensing for scientific innovation. *International Journal of Communications Law and Policy*, 13, 22–47. Retrieved from https://ssrn.com/abstract=1362040**
  Victoria Stodden explains the necessity of a standard for research artifact licensing as a means of promoting sharing and attribution to support research reproducibility.

- **Open Knowledge Foundation. (n.d.). *Guide to open data licensing*. Open Definition. https://opendefinition.org/guide/data/**
  This guide breaks down the practical aspects of licensing data and legal intellectual property rights for data in various country jurisdictions.

- **Morin, A., Urban, J., & Sliz, P. (2012). A quick guide to software licensing for the scientist-programmer. *PLoS Computational Biology, 8*(7), e1002598. https://doi.org/10.1371/journal.pcbi.1002598**
  This article covers information on software licensing and considerations when choosing a software license.

- **Hrynaszkiewicz, I., Busch, S., & Cockerill, M. J. (2013). Licensing the future: Report on BioMed Central's public consultation on open data in peer-reviewed journals. *BMC Research Notes, 6*(318). https://doi.org/10.1186/1756-0500-6-318**
  This article provides responses to common questions about open licenses for data based on BioMed Central's public consultation on Open Data.

It is important to understand the differences among various standard licenses to be able to determine which is most appropriate for the specific type of research artifact to be made publicly available. Below are resources and tools to help with license selection.

- **Open Source Initiative. (n.d.). *Licenses & standards*. Open Source Initiative. https://opensource.org/licenses**
  The Open Source Initiative provides centralized access to information on various open source licenses along with an FAQ on open source licensing.

- **Kamocki, P., Straňák, P., & Sedlák, M. (2015). *Public license selector*. Institute of Formal and Applied Linguistics. http://ufal.github.io/public-license-selector/**
  This web-based tool helps to choose a license for data or software from among various licensing including Apache, CDDL, BSD, Creative Commons, GNU, and MIT licenses.

- **GitHub. (n.d.). *Choose an open source license*. https://choosealicense.com/**
  This tool, developed by GitHub with contributions from the developer community, walks through the selection of an open source license.

- **Ball, A. (2014, July 17). *How to license research data*. Digital Curation Centre. https://www.dcc.ac.uk/guidance/how-guides/license-research-data**
  This DCC How-to Guide explains the why and how of licensing for research data including specific information about various types of licenses and when their use is most appropriate.

Technical durability

Lack of access to the technology (i.e., hardware, software, computing systems) required to re-execute the research workflow precludes any attempt to confirm the reproducibility of results. The resources below provide information to help maximize the usability of technology over time.

- **Library of Congress. (2021, November 12). *Sustainability of digital formats: Planning for Library of Congress collections*. Library of Congress. https://www.loc.gov/preservation/digital/formats/index.html**
  This Library of Congress web site provides comprehensive information about the factors that affect the sustainability of various types of digital content.

- **Library of Congress. (2021). *Library of Congress recommended formats statement 2021-2022*. Library of Congress. https://www.loc.gov/preservation/resources/rfs/**
  The Library of Congress maintains a recommended formats statement that outlines the necessary characteristics of file formats that ensure long-term preservation and usability along with a discrete list of recommended formats for various file types.

- See also **Thing 10: Review** to learn more about ways to ensure sustained access to the technology required for research reproducibility.

Go deeper

Open Science

Public access to research artifacts is a cornerstone of Open Science. Learn more about the Open Science movement to situate knowledge and knowledge-making processes in the public domain.

- **Open Knowledge Foundation. (n.d.). The open definition. Open Definition. http://opendefinition.org/**
  The Open Knowledge Foundation provides a precise definition of "openness" as it refers to research artifacts.

- **Murray-Rust, P., Neylon, C., Pollock, R., & Wilbanks, J. (2010). *Panton Principles: Principles for open data in science*. http://pantonprinciples.org/**
  The Panton Principles are a set of recommendations for the adoption of specific practices to make data maximally open.

- **Molloy, J. C. (2011). The Open Knowledge Foundation: Open data means better science. *PLoS Biology, 9*(12), e1001195. https://doi.org/10.1371/journal.pbio.1001195**
  In this article, the author defines open data and describes the impetus behind the open data movement in science.

- **Hey, T., & Payne, M. C. (2015). Open science decoded. *Nature Physics, 11*(5), 367–369. https://doi.org/10.1038/nphys3313**
  This article focuses on reproducibility as the reason that open source code is as important as open data.

Sensitive Data

While Open Science is the goal, there are instances in which access and use of materials must be restricted due to their sensitive nature. Datasets that contain personal health information or other identifiable data collected from human participants, data collected with Indigenous partners, traditional knowledge, and precise location information of vulnerable species or protected sites often must be safeguarded against unauthorized access or disclosure. Learn more about measures for assessing and mitigating risks to human participant data from the resources below:

- **Sensitive Data Expert Group. (2020). *Sensitive data toolkit for researchers part 2: Human participant research data risk matrix*. Zenodo. https://doi.org/10.5281/zenodo.4088954**
  This tool, developed by the Sensitive Data Expert Group of the Portage Network, helps to assess the level of risk to data containing personally identifiable information based on data content, context, and other factors impacting risk.

- **Sensitive Data Expert Group. (2020). *Sensitive data toolkit for researchers part 3: Research data management language for informed consent*. Zenodo. https://doi.org/10.5281/zenodo.4107178**
  Planning for the public release of human participant data begins with the informed consent process. This resource explains how consent language can impact if and how data can be shared.

- **Darragh, J., Hofelich, A. M., Hunt, S., Woodbrook, R., Fearon, D., Moore, J., & Hadley, H. (2020). *Human subjects data essentials data curation primer*. Version 2.0. https://github.com/DataCurationNetwork/data-primers**
  The Data Curation Network offers a primer on human subjects data that includes curation approaches for minimizing disclosure risk while providing access to the data or public-use versions of the data.

- **Steffensmeier, D., & Schwartz, J. (2021) *21st century corporate financial fraud, United States, 2005-2010*. Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR37328.v1**
  This ICPSR repository record is an example of how sensitive data can be made discoverable while limiting access to only those individuals who have followed specific protocols for access approval.

- **Centre for Applied Data Ethics. (2021). *Ethical considerations in the use of geospatial data for research and statistics. Confidentiality and disclosure risk.* UK Statistics Authority. https://uksa.statisticsauthority.gov.uk/publication/ethical-considerations-in-the-use-of-geospatial-data-for-research-and-statistics/pages/3/**
  Geospatial data can present particular risks when the data pinpoint an individual address or when overlapping data narrow a location small enough to identify a certain population. This resource provides important information on protecting geospatial data.

Thing 7: Provenance

Related FAIR Guiding Principle: R1.2

Related FAIR4RS Guiding Principle: R1.2

Provenance is a fundamental archiving principle that refers to information about the origins, chain of custody, and ownership of an object, which helps to establish the object's authenticity and authority. Provenance describes the lineage of data and other digital objects.

In the context of research compendium components, knowledge about their sources, creation, and transformation over time is critical to assessing the quality, reliability, and trustworthiness of the compendium as well as the research outputs it serves to substantiate.

The FAIR principles call for provenance to be captured as standardized machine-readable metadata. Provenance records a huge range of detail including the date the research object was created, the name of the creator, the tools and processes used to create the object, and any changes made to the object since its creation. An example of a formal standard for capturing provenance in a machine readable format is the WC3 PROV data model (see "go deeper").

Applies to **data, code and computation**

- For any transformation (including variable transformation), especially those done manually, e.g., when the full dataset cannot be made available due to the cost of collecting the data, privacy aspects of personal data, or intellectual property considerations, provenance should indicate how the subset was derived.

- If any external data sources are included or referenced, the creator, publisher, version and date accessed information should be documented. Where available, this includes a resolvable, globally unique permanent identifier such as a DOI.

- A history of the modifications made to the code, ideally including attributed sources when code snippets are incorporated. Reference to dependencies that includes the actual version used, not a reference to the latest version.

- Provenance of the research workflow that produced the outputs including scripts or representations in a formal workflow language such as the Common Workflow Language or similar, and a history of changes to these.

- Workflow systems such as **Kepler**, **Galaxy**, **Taverna**, NextFlow or SnakeMake can make it easier to capture provenance but require you to bring experiments or analysis into these systems prior to curation. "Code aware" data management/versioning systems such as DataLad (or the many tools and platforms used in MLOps) are another option. Care needs to be taken in making a choice of system, that provenance records in a useful form can be extracted for the target archive. See also Thing 9: Automation

Get started

Formal, machine-readable capture of provenance can be quite daunting for a newcomer, and for manually curated materials, can be quite time consuming. To get started, focus less on formal capture, and more on what the relevant categories of provenance detail are in your context. Think about protocols or processes to ensure that this detail continues to be captured thereafter. This includes:

- What was the overall intent or purpose driving the compendium (or where this is captured)
- What inputs were relevant when gathering data or writing code (funding, contributions, resources applied or consumed)
- How were changes in files captured over time
- Who contributed what to the outputs
- Where did the relevant data come from, and (how) can others do the same data
- What software tools and hardware were used in generating, preparing and analyzing the data
- What scripts and workflows were generated

Get started by sketching out these categories and more. You may not be able to access all of this information. In certain areas of research it may be necessary for you to capture precise detail about, for instance, where the data was gathered, or the conditions or constraints under which the project took place. You might like to explore the provenance records of a similar archive to the one you are targeting.

You may find that there are specific provenance requirements in order to be accepted to a certain archive. It is best to understand these requirements before data collection and analysis, as it can be impossible to uncover this detail once lost.

Learn more

A simple overview and suggestions for further reading is provided by the ARDC:

- **Australian Research Data Commons : Data Provenance**
  https://ardc.edu.au/resources/working-with-data/data-provenance/

Formal capture of data provenance is a major area of research, with many published papers:

- **Science Direct: Data Provenance**
  https://www.sciencedirect.com/topics/computer-science/data-provenance

See, for instance:

- **Leipzig J., Nüst D., Hoyt C.T., Ram K., Greenberg J. (2021). The role of metadata in reproducible computational research. Patterns 2(9): 100322.** https://doi.org/10.1016/j.patter.2021.100322
  Enabling reproducible research requires careful consideration of metadata. This resource unpacks metadata standards and steps for enabling reproducible *in-silico* experiments

Go deeper

The WC3 PROV data model (PROV-DM) offers a conceptual overview of provenance information for research objects, or entities. In this model, all steps in the provenance of an object are broken down into relations between agents, entities and activities. Formal capture of this level of detail is best suited to systems that automate many or most aspects of transformations. If this describes you, then it may be worth going deeper with formal PROV representation of metadata.

Otherwise, if you have experience in capturing basic provenance detail, you may like to further consider:

- What provenance information is necessary vs nice to have?
- Is traceability more important than reproducibility?
- Should provenance be gathered only if the users ask for it?
- Is there a provenance system that orchestrates the data flow and keeps a log to describe it?
- Do you agree that provenance of sensitive data processing is itself sensitive?

There are many resources available for formal capture of provenance.

Read an article:

- **Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, Crosas M, et al. (2014) Ten Simple Rules for the Care and Feeding of Scientific Data. *PLoS Comput Biol* 10(4): e1003542. https://doi.org/10.1371/journal.pcbi.1003542**
  The higher the quality of provenance information, the higher the chance of enabling data reuse. If you want your work to be reproducible, “then provenance information is a must (e.g., working pipeline analysis code, a platform to run it on, and verifiable versions of the data).”

Consider provenance in automated data collection: Increasingly, tools for gathering data are assisting in elements such as question order variation in computer assisted survey information collection or spatio-temporal metadata capture associated with electronic field data collection. In addition to capturing the steps for an analysis, if you use tools such as this, it may be worth checking to see if you can extract provenance detail capturing those automated steps as well.

Thing 8: Metadata

Related FAIR Principles: ALL

Related FAIR4RS Principles: ALL

Metadata are information that describe an object and may convey details such as who created the object, when and how the object was created, its purpose, its license or other terms of use, and how it is related to other objects.

Ideally structured and machine-readable, metadata about the research compendium and its component parts can be preserved and made accessible alongside the compendium to facilitate discovery and use.

Ask: is information about the research process (data, analysis, and production), the research compendium, and each of its components FAIR: findable, accessible, interoperable, and reusable?

The FAIR guiding principles make specific recommendations on types of metadata and ways of encoding it that are especially important for making data and code available. For Findability, identifiers are important as well as rich metadata (see below). For Accessibility making clear how to access the data or software (see Thing 6), and ensuring that metadata remains available are important. For interoperability, the use of standards in encoding metadata are important (see below). Finally, Provenance (see Thing 7) and licenses (see Thing 6) are specifically mentioned in the context of enabling Reusability.

Get started

Machine readable metadata refers to information structured such that it can be automatically read and processed by a computer. There are various metadata schemas available to accommodate the standards, requirements, and conventions of various domains, a number of which have tools or templates that can be used to generate structured metadata or verify the integrity of a structured metadata record.

Standards

A metadata standard is a set of required and optional metadata elements used to describe objects in a consistent manner. Use of a metadata standard can increase the probability of interoperability and contribute to a shared understanding of the meaning of the data.

To find a standard suitable for a particular discipline or data type, start with a metadata directory. Where a disciplinary standard doesn't exist, it is better to adopt terms from an existing standard(s), or use a general-purpose metadata schema such as DCMI Metadata Terms or DataCite Metadata Schema. You may also consider joining the Research Data Alliance to start work on defining a standard.

Sample Metadata directories:

- The Metadata Standards Catalog, created by the Metadata Standards Directory Working Group at the Research Data Alliance (RDA) is a directory of standards in use by various disciplines. Other relevant RDA work has been published by the Metadata Interest Group and Research Metadata Schemas Working Group, including Guidelines for publishing structured metadata on the Web.

- INCF standards and best practices portfolio for neuroscience includes many examples of discipline and method specific standards https://www.incf.org/resources/sbps
- FAIRSharing.org is a curated resource with many data and metadata standards, which can be narrowed by domain. https://fairsharing.org/standards/

Sample Metadata standards:

- Neuroscience: Brain Imaging Data Structure (BIDS) is a standard for organizing and describing neuroimaging and behavioral data.
    - **Gorgolewski, K., Auer, T., Calhoun, V. et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci Data* 3, 160044 (2016). https://doi.org/10.1038/sdata.2016.44**
      The Brain Imaging Data Structure (BIDS) is a standard for organizing and describing MRI datasets. The BIDS standard uses file formats compatible with existing software, unifies the majority of practices already common in the field, and captures the metadata necessary for most common data processing operations.
- Health care data exchange: Fast Healthcare Interoperability Resources (FHIR) is a standard for communicating research findings; the intended scope of FHIR is broad, covering human and veterinary, clinical care, public health, clinical trials, administration and financial aspects. The standard is intended for global use and in a wide variety of architectures and scenarios.
- Social and behavioral science: Data Documentation Initiative (DDI) is a standard used to describe survey data and observational data in social, behavioral, economic and health sciences – see below.

Learn more

Templates

There are various tools and resources available that can be used to generate structured metadata, or to validate structured metadata records. The resources below can help you get started.

- The Metadata Standards Catalog maintains a list of tools for generating, editing, or validating metadata records
- The Consortium of European Social Science Data Archives (CESSDA) “Documentation and metadata” training module includes links to a number of templates that can be used to describe resources in a standardized manner.
- CEDAR maintains a number of Metadata Tools that can be used to annotate and describe data; although originally designed for biomedical data, CEDAR’s Template Tools could be used to adapt templates for other disciplines
- GO FAIR Metadata for Machines Workshops

- The CodeMeta Generator can be used to structure metadata that describes a software object in JSON-LD notation. (See also the User Guide)

Crosswalks
A “crosswalk” is necessary when metadata created in one schema needs to be transformed to another schema, for example to facilitate sharing metadata between systems or to increase discoverability. For example, the crosswalk from MARC to Dublin Core by the Library of Congress.

For a basic overview, see the ARDC 23 (Research Data) Things lesson on metadata crosswalks.

The following resources provide more information:

- **Ojsteršek. (2021). *Crosswalk of Most Used Metadata Schemes and Guidelines for Metadata Interoperability* (1st ed.) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4420116.**
  This resource provides crosswalks among the most commonly used metadata schemes and guidelines to describe digital objects in Open Science. This resource is also referenced by FAIRsharing.org.

Go deeper

An in-depth look at the Data Documentation Initiative:

The Data Documentation Initiative (DDI) has been working to establish “metadata standards and semantic products for describing social science data, data covering human activity, and other data based on observational methods.” DDI Lifecycle provides free standards that “document and manage different stages of the research data lifecycle, including conceptualization, collection, process, distribution, discovery, and archiving” of social, behavioral, economic, and health sciences. The standard includes metadata at the study- and file-level, for example:

- Study-level
  - A researcher ID (e.g. ORCID) is specified for each author specified
  - Research organizations specified for author affiliations, ROR identifier added if possible
  - Research funding is specified, CrossRef Funder Registry Funder ID specified if possible. Grant ID is specified in standard format of funder (e.g. NIH funding IDs should include Activity code, Institute/Center, and Serial Number - see NIH guide)
  - License is standard, linked, and machine-readable license describing allowable reuse (see License section in Thing 6 for more details)
  - Study registrations are linked
    - Clinical Trials registrations include ClinicalTrials.gov Identifier
    - Study pre-registrations are linked with DOIs or links (e.g. OSF registries)

    - Associated materials are linked using persistent identifiers or links
        - Associated publications or preprints are noted with full citations and DOIs
        - Protocols deposited in repositories include DOIs (e.g. Protocols.io)
        - Related datasets, code, other digital materials deposited in other repositories are linked with DOIs
        - Relationship type between these items and the compendium is noted in a standard, machine-readable way if possible (see e.g., https://www.crossref.org/blog/doing-more-with-relationships-via-event-data/)
    - Title of compendium is meaningful, provides context to the work, and is not exactly the same as the title of an associated publication
    - A publication date is listed for the compendium, revision dates are also listed
- File-level
    - When possible, separate machine-readable and structured files are used to convey compendia metadata, such as
    - Citation using CITATION.cff or BibTeX files (CITATION.bib example)
    - Environment variables captured in environment.yml
    - License metadata can be done at least different 3 ways:
        - Captured as a standard text file (LICENSE.txt example)
        - Recorded in metadata via a vocabulary (e.g., SPDX terms)
        - Recorded in schema.org digest (as json) or markup in html
    - Some mechanisms for capturing software/code metadata include:
        - codemeta.json file
        - .zenodo.json (deposit metadata) file
        - CITATION.cff file
- Variable-level
    - Variable-level DDI metadata define the variables and specify units of measurement and permitted values (including missing value codes); it is applied to quantitative data.

  - This information is usually embedded within data files, e.g. as a header, or in column labels or generated from statistical package files.

The resources below offer a deeper look at metadata – what it is, how to use it, and some of the challenges.

- **Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 160018. https://doi.org/10.1038/sdata.2016.18**
  Machine-actionable metadata is information that enables the agent to have the capacity, "when faced with a digital object never encountered before, to: a) identify the type of object (with respect to both structure and intent), b) determine if it is useful within the context of the agent's current task by interrogating metadata and/or data elements, c) determine if it is usable, with respect to license, consent, or other accessibility or use constraints, and d) take appropriate action, in much the same manner that a human would."

- **CESSDA Training Team (2017 - 2020). *CESSDA Data Management Expert Guide*. Bergen, Norway: CESSDA ERIC. Retrieved from https://www.cessda.eu/DMGuide .**
  Start getting familiar with metadata with this guide from CESSDA.

- **Jedinger, Alexander. (2020). *What Is Metadata and How Do I Document My Data?*. Presented at the CESSDA Training Days 2019 (CTD2019), Cologne, Germany: Zenodo. http://doi.org/10.5281/zenodo.3923956**
  A video with Alexander Jedinger discussing metadata and data documentation.

- **Baca, Murtha, ed. (2016). *Introduction to Metadata* (3rd ed.). Los Angeles: Getty Publications http://www.getty.edu/publications/intrometadata.**
  An overview of metadata, including its types, roles, and characteristics; a discussion of metadata as it relates to web resources; and a description of methods, tools, standards, and protocols for publishing and disseminating digital collections.

- **Alper, Brian S., Joshua E Richardson, Harold P Lehmann, Vignesh Subbian (2020). It is Time for Computable Evidence Synthesis: The COVID-19 Knowledge Accelerator initiative, *Journal of the American Medical Informatics Association*, Volume 27, Issue 8, August 2020, Pages 1338–1339, https://doi.org/10.1093/jamia/ocaa114**
  Efforts to establish universal standards for biomedical data exchange, including proposed metadata crosswalk.

- **Manninen, Lauren (2018) Describing Data: A Review of Metadata for Datasets in the Digital Commons Institutional Repository Platform: Problems and Recommendations, *Journal of Library Metadata*, 18:1, 1-11, https://doi.org/10.1080/19386389.2018.1454379**
  A review of data hosted in Digital Commons repositories found that there is a need for a standardized language in describing research data within Digital Commons Institutional Repositories.

Thing 9: Automation

Related FAIR Principle: Reusable

Related FAIR4RS Principle: Reusability

Automation refers to a process, for example, computation, transformation, or curation, that makes computational reproduction efficient and minimizes, if not eliminates, human intervention in reproducing research. Automation goes beyond documentation to avoid the need to manually set up processes (e.g., installing software, setting up the compute environment).

There are many opportunities to automate aspects of the research process at various points during data handling and computation.

Get started

Automation means that information about the computational steps throughout the research process is script- or workflow-based, and manual processes are avoided to the fullest extent possible.

Including a good README file is a good first step: Instructions will always be needed.

> Social Science Data Editors: “We note that a good replication package uses a minimal number of automated scripts, with no manual interventions required unless absolutely unavoidable.”

Reasons for automation in research

The main steps in a project should be encoded in software and documented so they can be reliably reproduced.

- It’s better to use loops, functions or macros to automate any repetitive tasks for concise code (see Reproducible Research Things: “Can you automate any repetitive tasks?”).

- Avoids human error and undocumented idiosyncratic processes typical in manual workflows. Many disciplines use software-based or software-dependent means of data collection which help minimize human error. For example, in survey research, Computer-Assisted Telephone Interviewing (CATI) and Personal Interviewing (CAPI) are methods of coding information from telephone or personal interviews directly into a computer during the interview. CATI/CAPI software usually has built-in consistency checks, will not allow wild codes to be entered, and automatically prompts the interviewer for correct skip pattern questions (ICPSR).

- Increases the likelihood that code will run on multiple machines, that necessary packages will be loaded, that data are pulled from the appropriate sources in the correct way. Automation helps with portability, that is, moving the execution of the same code with the same data to either a similar or different computing environment (see TIER Protocol on the importance of portability).

Learn more

Relative paths (versus absolute paths)
High quality code can help with automation. A key example is how input data are called.

When calling files, the *absolute path* starts from the root. In a typical computer, it would be the C drive: However, the *relative path* starts from the current working directory. Relative path calls files relative to their position from the current working directory. Using relative paths makes your reproducibility package portable and can be installed in any folder location. When you double-click on the program file to open the statistical software like SAS, STATA, and R, the location of the program file becomes the current working directory and you can leverage this behavior to simplify the process of reproducing your study.

Workflows and pipelines
Script-based computation may cover data transformation, data analysis, figure and table generation, automated testing.

Consider dual workflows, “in which the data flow creates variables and datasets, while the analysis flow analyzes datasets, but does not construct variables or save datasets." (see Scott Long remarks)

In addition to the data creation and data analysis workflows, the review and verification workflow should be automated as much as possible. See example of a curation tool, YARD.

Computing environments
Rather than enabling scripts to run on different machines, constraining the execution environment to a specific operating system and specific dependencies may make it easier to ensure re-execution (see below for more). It may also be the truest representation of the compendium. An increasingly popular approach is to provide a build script for a container which fetches or contains all the relevant elements for reproduction. Doing so does not negate the need to properly document data and software, as the suitability of these approaches tend to change rapidly.

Holding the computation environment as fixed: Platforms that make this easy
One way to aid automation is to hold the execution environment as fixed. Fixing the execution environment (hardware, operating system, software and dependencies) increases the likelihood that code that ran in the same environment will run again without issue in the future.

A common approach to making it easier to rerun code is to do the computation on a cloud based service or platform. Examples of this approach include WHOLETALE, Code Ocean, or MyBinder. Many of these are services built on top of JupyterHub or RStudio. Integrations of similar functionality are also starting to become available within journals, notably the reproducible article from eLife, which brings the reproduction of results as close as possible to the published article.

More complex solutions are explored below.

Go deeper

Fifteen (15) Code Review Things
This is a list of 15 code review-related tasks that help improve the quality and efficiency, automation

of verification, reproducibility package preparation, and reuse.

1. File names of code are prefixed with the order of execution for easy identification of which code to run first and the next one.

2. Use a Master program file to call all programs used in the study (this is similar to the idea of "push button replication, see 3ie's push-button replication (PBR) project).

3. Note the software, packages, operating system used, and versions. And when the code was last executed.

4. Convert absolute file paths to relative file paths.

5. Check code for the presence of non-executable comments that document analysis processes.

6. Map code processes to paper sections using comments.

7. Remove or comment out unnecessary codes.

8. Label new variables and their values immediately after creation. These are needed should re-users need to rebuild their computing environment.

9. Run the code from start to finish in one go, using a different computer (not the one used to run the analysis). Rebuild the computing environment if necessary.

10. Address errors in the code. Repeat until no more errors are encountered.

11. Note the date the code was last executed from start to finish in the code.

12. Compare code output to findings presented in the article. The output should conform to a predetermined set of possible outputs (e.g., know what output you should expect from re-executing the workflow).

13. Verify that code generates output in the same order as they appear on the manuscript (if possible).

14. Insert a coding header to help researchers cite the article and code, contact the author, know the computing environment and software version used, and when the code was last updated or executed.

15. Create a Manuscript-Code-Output Checklist (Optional). It is a spreadsheet that lists the tables, figures, and in-text numbers found in your manuscript, specifies the program name and code line number that created them and verified that they were replicated.

Automating aspects of code publishing

- Github Actions with projects developed on GitHub github can be used to automate validating or generating some types of metadata:
  - Use “howfairis” to automate assessing some FAIR aspects of the code components of you compendium
  - Use “cffconvert” to validate a citation file format (CFF) file or use “CodeMeta2CFF” to convert a CodeMeta json file to CFF (CFF support is also available within github)
  - Use “pipreqs” to generate a requirements.txt via a github action (if working in python)
- Consider running license compliance tests. These check that dependencies provide suitable licenses, and that the licenses between dependencies are compatible.
  - The OSS Review Toolkit (ORT) will work with projects written in several programming languages (see also the experimental github action)
  - The “R Dependency License Report” is a github action that can be used with projects written in the R language
- A small number of domain-agnostic research archives include integrations such that ingestion from a github repository (or some other git based code development platforms) can be automated. Project metadata must be provided within the repository using a valid format. For instance, see the Zenodo and Figshare integrations with GitHub.
  - You can publish versions of your code using these integrations

Holding the computation environment as fixed: Virtual Machines, Containers and other solutions

For desktop based workflows, the computation environment can be fixed in place using a number of solutions. This could be using a virtual machine, or especially common these days is to use containerisation. This allows for a ready made (or built) computation environment for code that assists with portability and re-execution. This can be done by carefully constructing images or build scripts to use vagrant, docker or singularity, usually starting with an image or container that include all or most of the software required. The image or container might be a suitable target for archiving (being mindful of licensing restrictions), or the scripts and/or config files that describe how to build the image or container, fetch the appropriate code and data, and then combine them is potentially an approach that makes longer term preservation easier. There are also other tools that make it easier to use these technologies including ReproZip.

There are also desktop based approaches that do fix the software and dependencies, but not the hardware or operating system. An example of this is conda, or a level of abstraction above with repro-packs.

This “desktop based” approach isn’t necessarily bound to desktop use either, as many of the tools described above can also be deployed in the cloud in a scalable fashion.

See also the discussion of workflow or pipeline systems in Thing 7.

Improving the technical infrastructure
In addition to scripts, the broader technical infrastructure can help with automation.

The technical infrastructure supporting the creation, publication, and distribution of computationally reproducible research artifacts has advanced greatly in recent years.

Automation requires that a well-developed technical infrastructure is trusted, and verified (See Willis & Stodden, 2020).

README practice
Try this: Find a published README file and compare to this guideline: https://social-science-data-editors.github.io/template_README/

- **Gutiérrez, Germán, and Sophie Piton. "Revisiting the Global Decline of the (Non-Housing) Labor Share."** ***American Economic Review: Insights*** **2, no. 3 (September 2020): 321–38. https://doi.org/10.1257/aeri.20190285**.
  The README here follows the Social Science Data Editors guidelines which includes Data Availability Statements describing how to acquire the data, the location of the data in the reproducibility package, software requirements, codes and their locations, and contact information. Data, Code, and README location: https://www.openicpsr.org/openicpsr/project/115761/version/V1/view

Thing 10: Review

Related FAIR Principles: Reusable

Related FAIR4RS Principle: Reusability

Computational reproducibility is a particular type of Reuse. It refers to a user's ability to execute computations in order to reproduce results reported in scholarly publications (NASEM, 2019). A primary reason for curation in this context is to assess whether computational reproducibility can be achieved using the digital artifacts contained within the research compendium.

Curators helping prepare and review materials for archiving and publishing are often the "first reusers" of these materials.

A review is a series of managed activities that helps ensure continued access to and functionality of the research compendium and its components for as long as necessary.

Get started

Sharing of data and code is increasingly encouraged or mandated by various stakeholders in the scientific landscape. The OECD in 2020 notes that data and code are often required to achieve reproducibility of scientific results, a key goal: "sharing access to the data underpinning scientific publications allows peers to test and reproduce scientific results. In practice, data alone are often insufficient to test reproducibility, and enhanced access to analysis software is also necessary."

The 2021 UNESCO Recommendation on Open Science states that, "scientific outputs related to publications (e.g. original scientific research results, research data, software, source code, source materials, workflows and protocols, digital representations of pictorial and graphical materials and scholarly multimedia material) that are openly licensed or dedicated to the public domain should be deposited in a suitable open repository, following appropriate technical standards that allow them to be properly linked to publications."

This document also recommends changes to "peer review and other evaluation and verification methods" in order to achieve improved "quality, reproducibility and impact of science, and thereby the reliability of the evidence needed for robust decision making and policy and increased trust in science."

Reviewing code

When curating for reproducible research, the Data Quality Review Framework can help organize review activities into four categories: Files, Documentation, Data, and Code. Many of the Things in this document address concepts and strategies that cover reviewing files, documentation, and data.

We focus here on "Code Review" as it is a newer practice, but one that is integral to the archival and publication of a computationally reproducible research compendium.

In general terms, code review includes an assessment of,

- The purpose of the code (e.g., recoding variables, manipulating or testing data, testing hypotheses, analysis), and
- Whether the code accomplishes this goal. In many cases, that means a verification that the code is executable and that reported findings are verified.

Code review can take place before the research compendium is published (“pre-publication review”) or after it is published (“post-publication review”).

The practice of code review can be implemented by professional staff in repositories and archives, academic institutions, scholarly associations and commercial publishers, and third-parties or by peers.

Learn more

Curators can identify problems with code files and remediate them before materials are archived and published. The resources below include examples of code-related irreproducibility, guidelines, tools, and services for managing the review and verification process, and information about badges.

Examples of code-related irreproducibility
Challenges to reproducibility that involve code may include,

- Un- or non-executable code. The code cannot be compiled, or running the code returns errors.
- Error in variable operation. The code executes but the operation diverges from its description in the manuscript. For example, a manuscript defines obese as BMI >28, but the code defines obese as BMI>= 28.
- Output discrepancy. The code executes but produces different results. For example, R code produces a result that is different from the result reported in the manuscript:
- Messy code. The code executes but requires high computing power.
- Hardcoded values. The code executes but values reported in the code are not generated by the code. The value may be generated by hand calculator or a different application (e.g., Excel) and results are entered into a program file. This can cause errors when data is further refined or if a mistake was made when calculating it.
- Outdated comments. The code executes but it has been evidently updated such that comments in the code no longer apply. For example, comments have typos or have not been updated.

How to review code and address code-related irreproducibility
Curators, data and information professionals, researchers and others all can contribute to preventing and addressing these problems to ensure that re-users of the research compendium are able to reproduce the results without needing to contact the original producer or track down missing components. Re-users might be the original authors at a later point in time, curators, peer reviewers, and others.

Guidelines. Several organizations engaged with code review offer useful guidelines for how to curate materials for reproducibility. Guidelines tend to be domain-specific because the requirements are specific. Some examples of domain-specific guidelines include,

- The CURE Consortium (Curating for Reproducibility) provides materials
- Fifteen Code Review Things (see Thing 9: Automation)
- Cornell Center for Social Sciences R2 provides Reproduction Materials Review Recommendations

Tools. There are also several tools for managing the review, verification, and curation process. For example, Whole Tale, CODECHECK, and CASCaD are options for general-purpose projects. See this conversation about solutions for reproducibility on the Taking the Pulse blog.

Some examples of domain-specific tools include,

- Yale Application for Data Review (YARD)
- Social Science Reproduction Platform (SSRP)

See the “Go deeper” section below for more resources about the practice of reproducibility checks.

Who reviews code

Researchers, curators, publisher and repository staff, and others may review the code to verify reproducibility. For example, a number of ACM (the Association of Computing Machinery) conferences and journals have instituted formal processes for artifact review.

Third-party services are individuals or organizations not affiliated directly with the primary stakeholders in the research lifecycle (e.g., researchers, repositories, institutions, or publishers). Third-party services can be contracted to assist with the verification process.

Some examples of domain-specific services include,

- Cornell Center for Social Science R2
- Odum Institute for Research is Social Science

Who reviews and how is a function of the goal of the review as well as the timing (i.e., pre- or post-publication).

Researchers can do a lot to facilitate computational reproducibility of their own research.

Anticipating and preventing code-related irreproducibility early in the research life-cycle is prudent. Researchers can avoid potential problems and help ensure a research compendium is ready for archiving and publication by following these five review steps before submitting a manuscript for publication:

1. Review your code. Run your code from start to finish in one go. If possible, use a different computer with just the base package of the statistical software used in the analysis installed. Correct all errors encountered. Note the following in the README:
    a. Operating System used and its version
    b. CPUs and RAMs
    c. Statistical software package used and their version
    d. Packages/libraries/Ados used and their version

    Why? So you can verify that the code is error-free; detect and note the packages, libraries, and ado files required to run your analysis; and so that future re-users of your files have sufficient information to rebuild the computing environment used in the analysis, if necessary.

2. Review your manuscript. Verify that methods and operationalization of variables stipulated in the manuscript were correctly implemented in the code. For example, if obesity was defined as bmi >28 in the manuscript, then it should be coded as bmi >28 in the code, and not bmi >=28. Ensure the following are cited in your manuscript:
    a. Dataset used. Cite yourself if you created the dataset.
    b. Statistical software used
    c. Packages, libraries, and ado files used

3. Review the output produced by your code. Compare the results generated by your code to the findings presented in the manuscript. They should be identical up to the last decimal place. Verify you are pulling numbers from the correct table columns or cells.

4. Review your data. Make sure all original and computed variables are labeled. And prepare a Data Availability Statement for each dataset used in the study, which details the steps to acquire the data and its license.

5. Review your README. Make sure it has instructions on how to reproduce the study, contact information of the primary author, data and code citations and licenses, data availability statements, and information about the computing environment and packages, libraries, and ados used in the study.

Note: A useful, domain-specific, resource is offered by the American Economic Association (AEA) Data Editor. These are guidelines for authors on how to prepare a "replication package" for an AEA journal.

Go deeper

Review is an integral aspect of the scientific process and an important aspect of reproducible research. Learn more about key topics related to review.

FAIR principles and code review

Gold standard data review has been a province of data archives for a long time (e.g., Inter-university Consortium for Political and Social Research (ICPSR), UK Data Archive). Many of the curation practices common in such archives conforms to the FAIR principles for data.

Recently, new platforms have been created to review and assess data and metadata specifically for "FAIRNESS" (e.g., FAIR Metadata Assessment tool is in development by DataONE; FAIR Evaluator developed by FAIRsharing.org ; F-UJI developed by PANGAEA under the umbrella of the FAIRsFAIR project; Australian Research Data Commons' FAIR data self assessment tool).

Efforts have been made to apply FAIR to research software.

Journal policies and computational reproducibility

Perhaps inspired by the Transparency and Openness Promotion (TOP) Guidelines, scholarly journals have been implementing policies requiring authors to provide data and code and submitting these materials to various levels of scrutiny and review.

The National Academies of Science, Engineering, and Medicine in the U.S. made the following recommendation in its report:

> RECOMMENDATION 6-4: Journals should consider ways to ensure computational reproducibility for publications that make claims based on computations, to the extent ethically and legally possible.

The NASEM report goes on to acknowledge that this can present some challenges:

> Although ensuring such reproducibility prior to publication presents technological and practical challenges for researchers and journals, new tools might make this goal more realistic. Journals should make every reasonable effort to use these tools, make clear and enforce their transparency requirements, and increase the reproducibility of their published articles.

Examples of journal policies that require data and code sharing and that include a code review:

General

- Science Data and Code Deposition Policy
- Nature Reporting Standards and Availability of Data, Materials, Code and Protocols
- PLOS Materials, Software, and Code Sharing

Political Science

- Journal of Politics Data Replication Policy
- American Journal of Political Science (AJPS) Verification Policy (read the Christian, et al., (2018) article on the collaboration between AJPS and the Odum Institute Data Archive to develop a workflow that bridges manuscript publication and data review processes.)

Economics & Management

- American Economic Association (AEA) Code and Data Availability Policy
- Royal Economic Society Replication Policy
- Management Science Policy for Data and Code Disclosure

Resources on the practice of reproducibility checks

- **Christian, T.-M. L., Lafferty-Hess, S., Jacoby, W. G., & Carsey, T. (2018). Operationalizing the Replication Standard. *International Journal of Digital Curation*, 13(1), 114–124. https://doi.org/10.2218/ijdc.v13i1.555**
  A case study of the collaboration between AJPS and the Odum Institute Data Archive to develop a workflow that bridges manuscript publication and data review processes.

- **Nüst D. and Eglen S.J. (2021). CODECHECK: An Open Science Initiative for the Independent Execution of Computations Underlying Research Articles During Peer Review to Improve Reproducibility [version 2; peer review: 2 approved]. *F1000Research*, 10:253. https://doi.org/10.12688/f1000research.51738.2**
  The CODECHECK system uses open infrastructure and tools and can be integrated into review and publication processes in multiple ways.

- **Peer, L., Green, A., and Stephenson, E. (2014). "Committing to Data Quality Review." *International Journal of Digital Curation* 9 (1): 263–91 https://doi.org/10.2218/ijdc.v9i1.317**
  "We can think of a continuum of data curation that progresses from a basic level where data are accepted 'as is' for the purpose of storage and discovery, to a higher level of curation which includes processing for preservation, improved usability, and compliance, to an even higher level of curation which also undertakes the verification of published results."

- **Peer, L. and Dull, J. (2020). YARD: A Tool for Curating Research Outputs. *Data Science Journal*, 19(1), p.28. http://doi.org/10.5334/dsj-2020-028**
  YARD (Yale Application for Research Data) is an adaptable curation workflow tool that enhances research outputs and associated digital artifacts designated for archival and reuse. YARD helps create high quality data packages that are FAIR, transparent, and reproducible.

- **Peng, R.D. (2011). Reproducible Research in Computational Science. *Science*, 334 (6060), 1226-1227. https://doi.org/10.1126/science.1213847**
  Argues that articles that have passed the reproducibility review "convey the idea that a knowledgeable individual has reviewed the code and data and was capable of producing the results claimed by the author. In cases in which questionable results are obtained, reproducibility is critical to tracking down the 'bugs' of computational science."

Reproducibility badges

Badges are considered "simple, effective signals to promote open practices and improve preservation of data and materials by using independent repositories.

- The Association for Computing Machinery (ACM) lists several badges for reproducibility reviews.

- The Center for Open Science (COS) badges acknowledge open science practices (sharing data, materials, or pre-registration).

For more information on badges, see:

- **Hardwicke, T.E., et al. (2021). Analytic Reproducibility in Articles Receiving Open Data Badges at the Journal Psychological Science: An Observational Study. *Royal Society Open Science*. 8: 201494. https://doi.org/10.1098/rsos.201494**
  Evaluated analytic reproducibility in 25 Psychological Science articles awarded open data badges between 2014 and 2015. 16 articles (64%) contained at least one major numerical discrepancy, requiring additional input from original authors.

- **Kidwell, M.C., et al. (2016). Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLoS Biology* 14(5): e1002456. https://doi.org/10.1371/journal.pbio.1002456**
  Found that when badges were earned, reportedly available data were more likely to be actually available, correct, usable, and complete than when badges were not earned. Badges are simple, effective signals to promote open practices and improve preservation of data and materials by using independent repositories.

Active maintenance

Confirming computational reproducibility at a single point in time is not a guarantee of long term reproducibility. While some aspects of curation are more or less “one and done” (e.g., assigning a persistent ID to a file, documenting a dataset), others call for repeated assessment. As a curator, you will need to perform periodic reviews, or ***maintenance***, ideally in a scheduled and automated way, and remediate if need be.

The idea that digital curation is an ongoing activity, a “series of managed activities necessary to ensure continued access to digital materials for as long as necessary” is central to digital preservation. It is reinforced in the foundational text that defines the FAIR principles: “Beyond proper collection, annotation, and archival, *data stewardship includes the notion of ‘long-term care’* of valuable digital assets, with the goal that they should be discovered and re-used for downstream investigations, either alone, or in combination with newly generated data” (Wilkinson et al., 2016; Italics added).

Maintenance highlights the time dimension of curation. Consider this: It is common practice for an archive concerned about the authenticity and integrity of its digital content to make use of checksums at the time of deposit, and in subsequent periodic reviews (see, for example, ICPSR Digital Preservation Policy Framework). Similarly, concern about computational reproducibility calls for not only curation of the research compendium (that includes all the other Things in this document) at one point in time, often indicated with a badge (see, for example, the ACM’s Artifact Review and Badging), but a periodic review.

When planning to publish or archive a research compendium you will want to think about what might require periodic review. What might you worry about breaking down over time? Specifically, consider the effect of the following scenarios on computational reproducibility:

- Digital files may get corrupted, inaccessible, or unavailable.
- Permissions to access data may change.
- Computer programs, scripts, or code may become unavailable or may not compile or execute.
- Computing environments and parameters may change.

The following resources may be useful in considering the above scenarios:

On the rationale for maintenance:

- **Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Scientific Data* 3, 160018. https://doi.org/10.1038/sdata.2016.18**
  The position that long-term care of digital materials is a continuous activity is articulated in the Wilkinson et al 2016 article that originally introduced the FAIR principles: “Beyond proper collection, annotation, and archival, *data stewardship includes the notion of ‘long-term care’* of valuable digital assets, with the goal that they should be discovered and re-used for downstream investigations, either alone, or in combination with newly generated data” (Italics added).

- **UNESCO (2020, March 2). *UNESCO Recommendation on Open Science*. https://en.unesco.org/science-sustainable-future/open-science/recommendation**
  The report states that, “Open research data… [should be made] available in a timely and user-friendly, human- and machine-readable and actionable format, in accordance with principles of good data governance and stewardship, notably the FAIR (Findable, Accessible, Interoperable, and Reusable) principles, *supported by regular curation and maintenance*” (Italics added).

- **Spinellis, D., Louridas, P., & Kechagia, M. (2021). Software Evolution: The Lifetime of Fine-Grained Elements. *PeerJ. Computer science*, 7, e372. https://doi.org/10.7717/peerj-cs.372**
  A challenge for ensuring long term computational reproducibility is rooted in the fact that it relies on code and software. When it comes to code, the time horizon for “regular curation and maintenance” is not clear. The lifespan of code is, by this estimation, less than two and a half years.

- **Peer, L., Orr, L., & Coppock, A. (2021). Active Maintenance: A Proposal for the Long-Term Computational Reproducibility of Scientific Results. *PS: Political Science & Politics*, 54(3), 462-466. https://doi.org/10.1017/S1049096521000366**
  Proposing that researchers dedicated to computational reproducibility should have a plan in place for “active maintenance” of their analysis code, and offering concrete suggestions for how data archives, journals, and research communities could encourage and reward the active maintenance of scientific code and data.

On strategies for maintenance:

- **Jeffrey M. Perkel (2020). Challenge to Scientists: Does Your Ten-Year-Old Code Still Run? *Nature* 584, 656-658. https://doi.org/10.1038/d41586-020-02462-7.**
  As a dynamic digital object, code “will eventually decay, and you will have to repair it, and you’ll have to replace it,” as Carole Goble points out.

- **Akhlaghi, M., Infante-Sainz, R., Roukema, B. F., Khellat, M., Valls-Gabaud, D., & Baena-Gallé, R. (2021). Toward Long-Term and Archivable Reproducibility. *Computing in Science* Engineering, 23(3), 82–91. https://doi.org/10.1109/MCSE.2021.3072860**
  Introducing “Maneage” (managing data lineage), which enables cheap archiving, provenance

extraction, and peer verification that has been tested in several research publications. The article includes 8 criteria for longevity (i.e., “the length of time that a project remains functional after its creation”). Includes review of the longevity of commonly used tools (e.g., Docker, Jupyter).

- **National Digital Stewardship Alliance (2019).** ***The Levels of Digital Preservation V2.0 Matrix*****,** **https://ndsa.org/publications/levels-of-digital-preservation/**
  A framework for thinking about the specific actions that curating for reproducible and FAIR research entails. The Framework specifies the actions necessary to ensure that digital content is known, protected, monitored, and sustained.

Version	Date	Authors
v.1	October 7, 2021	CURE-FAIR co-chairs, Mandy Gooch
v.2	October 28, 2021	CURE-FAIR co-chairs, Mandy Gooch
v.3	November 28, 2021	CURE-FAIR co-chairs, Mandy Gooch