Note: This document is now closed for comments. Please see v.3 here.

10 Curation for Reproducible and FAIR Things

DRAFT

CURE-FAIR WG RDA VP18 sprint period: October 28-Nov 3, 2021

&

Post-plenary extended period until Nov 25, 2021

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?usp=sharing (or https://bit.ly/3Ft00ld)		
v.2	October 28, 2021	CURE-FAIR co-chairs, Mandy Gooch
CURE-FAIR WG Plenary sprint period: October 28-Nov 3, 2021 https://docs.google.com/document/d/14A1sqdxhGAsFlPoyHUglrS89rcLY5aP_xizO2ZckcKA/edit?usp=sharing		

How to contribute to this document:

1. **Before the Plenary Session (Breakout 3 | 22:30 - 00:00 UTC):**
   Sign up for a task
   You will find discrete tasks highlighted in orange throughout the document below. If you would like to contribute to this document during the live plenary session, please sign up for a task by adding your name to it in the highlighted space provided. For example: for the task, [INSERT EXAMPLE], you will add [INSERT EXAMPLE, John Doe] to sign up. Multiple people may sign up for a task, and you may sign up for as many tasks as you like. During the live plenary session, we will have an open workshop period to allow attendees to work on the tasks available.
2. **At the Plenary Session (Breakout 3 | 22:30 - 00:00 UTC):**
   Contribute to the document
   We have highlighted each discrete task that is available for completion during this plenary session. To contribute, simply scroll to the highlighted task you wish to take on (or the task(s) you signed up for in advance) and add your contribution as a suggestion. You are more than welcome to contribute to multiple tasks and add comments with questions or feedback throughout the document. Please park your cursor in the designated areas throughout the document to make this process as manageable as possible.

If you can contribute to the document, please make sure to include your name in the Contributor List below for proper acknowledgment.

Name	**Institution**
Tom Honeyman	Australian Research Data Commons
Limor Peer	Institution for Social and Policy Studies, Yale
Mandy Gooch	The Odum Institute for Research in Social Sciences, University of North Carolina at Chapel Hill
Daniel S. Katz	University of Illinois at Urbana-Champaign
Peter McQuilton	Independent Consultant
Thu-Mai Christian	The Odum Institute for Research in Social Sciences, University of North Carolina at Chapel Hill
Florio Arguillas	Cornell Center for Social Sciences, Cornell University
Ana Van Gulick	Figshare
Wanda Marsolek	University of Minnesota

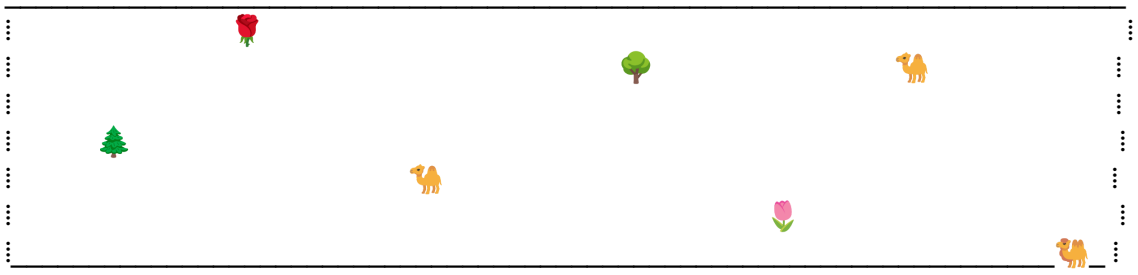


Table of contents

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. As a means of making scientific claims transparent, computational reproducibility 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 to 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.

> ***We curate to: a) assess whether computational reproducibility can be achieved using the objects contained within the research compendium, and b) ensure that the quality***

> ***of the objects meets community standards for for long-term discovery, access, reuse, and archival preservation.***

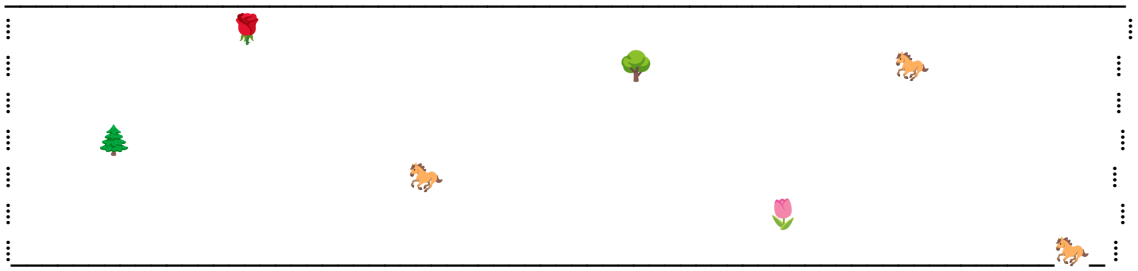
Both activities are essential aspects of curation in the context of reproducible research. We call these practices CURE-FAIR or Curation for FAIR and Reproducible Research.

Curation is often carried out near the end of the research lifecycle by data and archive professionals who are often subject matter experts. However, there are many data management actions that other key stakeholders can take earlier in the lifecycle that can 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, curators can flag issues with a research compendium that precludes computational reproducibility and 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 computation.

> ***This document includes standards-based guidelines for CURE-FAIR best practices in publishing and archiving computationally reproducible studies in the social sciences for research that relies on quantitative data to produce results. Our hope is that these "10 Curation for Reproducible and FAIR Things" will serve as a starting point for the development of curatorial guidelines tailored to the specific concerns of the social sciences community, other domains, and disciplines, and to the particular curatorial concerns and requirements of an archive 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 can reproduce prespecified results. Secondarily, it is for researchers, publishers, editors, reviewers, and others who have a stake in creating, using, sharing, publishing, or preserving reproducible research.



Ten Curation for Reproducible and FAIR Things

We follow 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 FAIR Curation for Reproducibility. 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 and any applicable standards, followed by three actionable sections: Get started, Learn more, and Challenge me.

The “10 Curation for Reproducible and FAIR Things” addresses the following questions:

Does a 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 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 full context necessary to understand the research process is available.
>
> Thing 5: Documentation: The process and reasoning required to computationally reproduce a scientific claim are readily available and understandable.

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

> Thing 6: Rights: It is clear who can use what, how, and under what conditions. Open is preferred.
>
> Thing 7: Provenance: The origin of the components of the research compendium and how each has changed over time is evident.

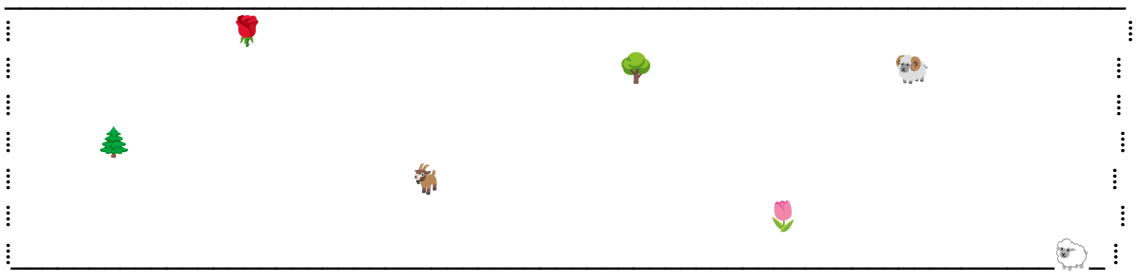
Is information about the research compendium and its components embedded in code?

Thing 8: Machine first: Information about the research compendium and its components is embedded in a standardized schematic code.

Thing 9: Automation: Information about the computational steps throughout the research process is script- or workflow-based.

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

Thing 10: Periodic 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

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.

[ADD RELEVANT STANDARD(S)]

[MAP TO RELEVANT FAIR PRINCIPLE]

Get started

Review a bundle of files you're going to submit or are currently curating, or have a look at this example [INSERT EXAMPLE] and consider the following:

- Data:
    - If specific input data are necessary to reproduce a prespecified outcome, are the data included as a separate digital object?
    - If not, do you have detailed information on where and how to obtain the specific input data or how to generate them?
- 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)?
- Codebook
    - Does the compendium include a codebook or data dictionary that defines the contents of the data file?
- README file

    - Does the compendium include a README file that outlines?

- Container

    - [QUESTIONS TO CONSIDER HERE]

Learn more

- 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:iKVoC8ysGIWUdf5Gv/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

- [LINK AND DEFINITION FOR SENSITIVE AND RESTRICTIVE DATA]

- "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."

    - Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 https://doi.org/10.25490/a97f-egyk

- [LINK TO EXAMPLE OF RESTRICTED DATA; REPOSITORIES THAT TAG DATA E.G., DATAVERSE OR ALLOW ACCESS TO RESTRICTED ENCLAVES E.G., ICPSR]

- [RECOMMENDATION: BASE CODE]

- [LINK TO EXAMPLE OF ANNOTATED OUTPUT TO ENHANCE REPRODUCIBILITY]

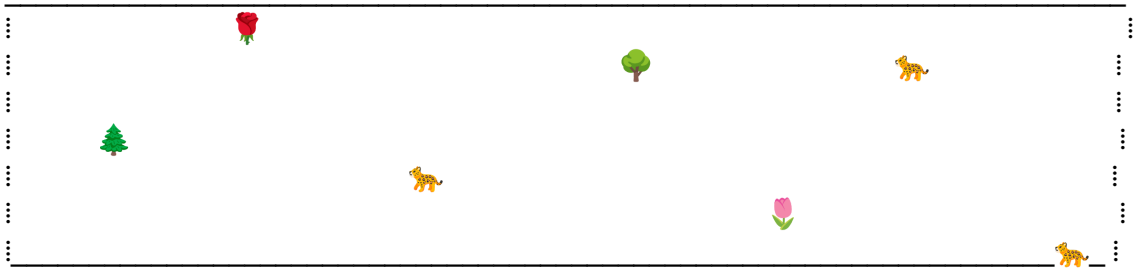
- Thing 5: Documentation

- Thing 6: Rights/Openness

Challenge Me

Exercises for incomplete compendia where the curator has to identify what is missing

- [IDEAS FOR WHAT COMPONENTS SHOULD BE INCLUDED IN THIS EXERCISE]



Thing 2: Organization

Organization is necessary to understand and to keep track of the various objects and the relationships among them over time.

Projects evolve and things get more complicated and opaque over time; memory fails;

[ADD RELEVANT STANDARD(S)]

[MAP TO RELEVANT FAIR PRINCIPLE]

Get started

Review a bundle of files you're going to submit or are currently curating, or have a look at this example [INSERT EXAMPLE] and consider the following:

- 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?

Learn More

Learn about various ways to organize a research compendium

- [TIER PROTOCOL (NOTE THAT CCSS ADDS INTERMEDIATE FOLDER, PREREQUISITES FOLDER, AND MORE)]
    - *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/.
- [USE HUMAN-UNDERSTANDABLE FILE NAMING CONVENTIONS (PROVIDE CITATION HERE)] CRATE/]

Challenge Me

What if the objects cannot be organized in folders?

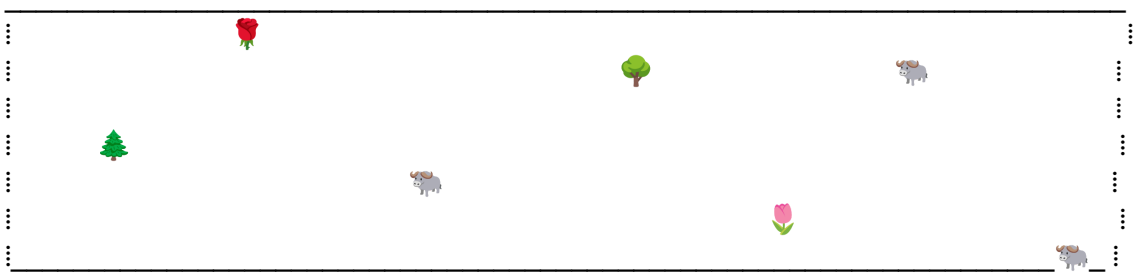
- Enabling reuse may mean separately handling the "objects" of your reproducible file bundle. Perhaps you are curating a bundle where the data is available elsewhere? If the data are archived separately from the code, then an additional step will likely need to be incorporated

into the processing scripts to fetch the data. While adding extra labour, 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:

- https://doi.pangaea.de/10.1594/PANGAEA.726855
- [ADD MORE EXAMPLES DEMONSTRATING THAT THE DOWNLOAD LINK IS ENCODED DIFFERENTLY DEPENDING ON THE ARCHIVE]

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?



Thing 3: Economize everything

As you curate your reproducible file bundle, consider if there are any extraneous parts you can cut out to make the overall bundle simpler to streamline computational reproduction. Fewer parts mean fewer things that can break, and less work caring for them over time.

[ADD RELEVANT STANDARD(S)]

[MAP TO RELEVANT FAIR PRINCIPLE]

Get started

Review a bundle of files you're going to submit or are currently curating, or have a look at this example [INSERT EXAMPLE] and consider the following:

- Can the scripts be simplified by removing redundancies or using loops and functions, for example?
- 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?

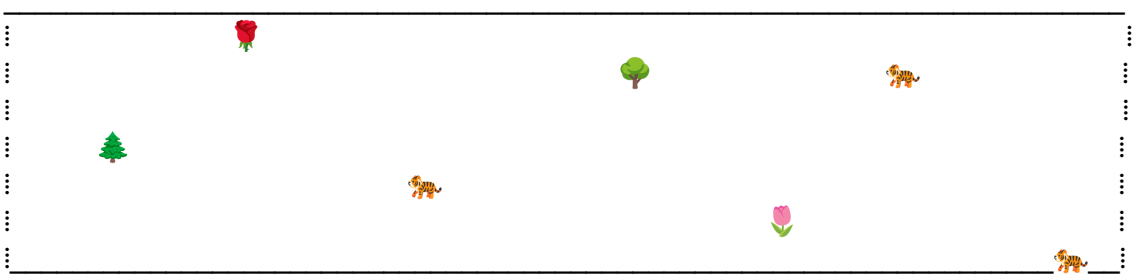
Learn More

You may like to review the following resources in considering the above questions

- [LINKS TO EXAMPLES, ARTICLES, ETC. FOR RECOMMENDATIONS ON CODING PRACTICES FOR REPRODUCIBILITY]
- [RECOMMENDED FILE FORMATS FOR CURATING CODE FILES?]
- Literate programming
  http://www.literateprogramming.com/
- Efficient R programming
  https://bookdown.org/csgillespie/efficientR/

Challenge Me

If your analysis code is written in R, try applying the RClean tool to your code to see if it can help you.



Thing 4: Transparency

Transparency is a precondition of computational reproducibility. Inability to reproduce results is often caused by lack of transparency. Problems of opaque (or “black box”) research is identified as a source of irreproducible research and leads to public doubts about the legitimacy and credibility of empirical scholarship.

There is a lot of tacit knowledge that researchers and research support functions hold that needs to be surfaced and made explicit.

Ideally these materials and information are as open as possible, following the principles of open science.

Importance of transparency for scientific integrity and public trust…

Initiatives such as the Data Access and Research Transparency (DA-RT) initiative that includes three main elements: data access, analytical transparency, and production transparency;

- Data transparency: Transparency about where the data can be accessed, how it was created, and any transformation it underwent (link to licenses; metadata)
- Analytics transparency: Transparency about analytical decisions, how the research measures, interprets, and analyzes evidence. Can include detailed documents specifying statistical models, dependent variables, covariates, interaction terms, and multiple testing corrections. (Reporting standards such as CONSORT for clinical trials require researchers to report all measures, manipulations, and data exclusions, as well as how they arrived at final sample sizes.)
- Production transparency: Transparency about a broader set of research design and method choices made in the research, for example, randomization procedures, probability assignments. Pre-registration and pre-analysis plans create a public record of studies that might otherwise go unpublished.

Transparency about the curation process itself is also important

    - HOW TO TRACK IT, REPORT IT [EXPAND]

[ADD RELEVANT STANDARD(S)]

[MAP TO RELEVANT FAIR PRINCIPLE]

Get started

Where is this information made available?

Are tools that help make this information transparent used? (e.g., pre-registration, pre-analysis plan)

Best practices for data:

- USE HUMAN-UNDERSTANDABLE FILE NAMING CONVENTIONS [ADD CONTENT]

Best practices for code:

- Call up data from correct location (i.e., relative path)
- Comment the code
    - Literate programing
- Specify computing environment (e.g., InfoSession)

Learn More

About DA-RT: A leading group of political scientists created the Data Access and Research Transparency (DART) 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.

Lupia, A., & Elman, C. (2014). Openness in political science: Data access and research transparency. PS: Political Science & Politics, 47, 19–42 and Appendices A and B. doi:10.1017/s1049096513001716

Transparency and qualitative research
https://www.princeton.edu/~amoravcs/library/TransparencyinQualitativeResearch.pdf

Tools: Active citation (AC) and Annotation for Transparency Initiative (ATI)
https://qdr.syr.edu/drupal_data/public/QDR_A%20Guide%20to%20ATI.pdf

On three mechanisms for increasing transparency in scientific reporting:
https://doi.org/10.1126/science.1245317

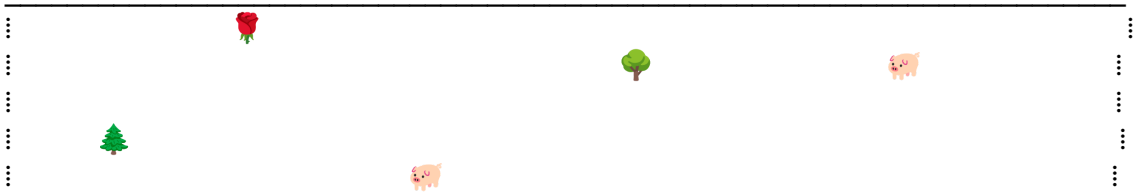
Why COVID-19 makes research transparency more important than ever
https://medium.com/center-for-effective-global-action/why-covid-19-makes-research-transparency-more-important-than-ever-84071f3586be

Transparency in Science: Nature Geoscience https://www.nature.com/collections/kvfmrywcdg

On types of knowledge (tacit and explicit): Polanyi, M. (1958), Personal Knowledge: Towards a Post-Critical Philosophy, University of Chicago Press, Chicago.
https://press.uchicago.edu/ucp/books/book/chicago/P/bo19722848.html

Challenge Me

See Center for Open Science badges scheme: what badges would you give to [X] STUDY? [NEED TO IDENTIFY A STUDY OR STUDIES FOR THIS EXERCISE]




Thing 5: Documentation

Developing robust documentation will increase the reproducibility of your research by providing users all of the context necessary to not only recreate your compute environment, but also understand your entire research compendium. What we mean by documentation is [EXPAND HERE]. The effort here is to make **tacit** knowledge more explicit. The researcher communicates as if they assume nothing (for example, a future user will not use the original computing environment). What are all the ways that research and curation decisions are made explicit?

Create documentation with the following questions in mind:

- What is in this research compendium? Who created it? How do I cite it?
- Am I permitted to use it and in what capacity?
- Can I recreate the compute environment using all of the information provided in order to run the analyses?
- Are these the correct versions of the packages and files necessary for the analyses?
- Am I able to understand what I am looking at when opening the data file(s)?
- Does the documentation enable the compendium to be “independently understandable for informed reuse?”

[ADD RELEVANT STANDARD(S)]

[MAP TO RELEVANT FAIR PRINCIPLE]

Get Started

Review your research compendium, or this [EXAMPLE] compendium and consider the following:

COMPENDIUM

- README
    - [SPECIFY WHAT A README INCLUDES]

- DATA
    - Do we know where is it from and/or how to get it?
    - Do we know how it was constructed, transformed?
    - Do we know the version?
    - Codebook or data dictionary

- CODE
    - Do we know the software version?
    - Do we know the dependencies?

- Are there comments in the code?
- How to call up data from correct location (relative path)

- COMPUTATION / ANALYSIS
  - Is there a pre-registration, pre-analysis plan?

CURATION
- How to track it, report it
- Is there a **README** (or similar document) that details all dependencies, any folder hierarchies, and/or instructions for re-running the analyses?
- Is a data availability statement present within the documentation? Are there additional licenses or Terms of Use?
- Is there a **codebook** available that provides enough context for users to understand the data files provided?
- Are original data sources fully cited?

Learn More

If you'd like to read up on good documentation practices for curation and reproducibility, please explore the following articles and resources:

- Guidance on how to deposit data at the AEA Data and Code Repository https://aeadataeditor.github.io/aea-de-guidance/data-deposit-aea.html
- DataONE, Best Practice: Describe https://dataoneorg.github.io/Education/bp_step/describe/

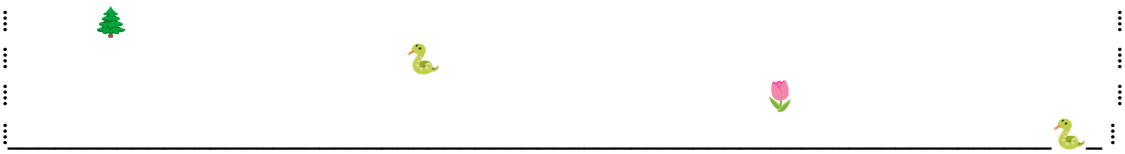
How to comment code, more on literate programing [EXPAND HERE]

Importance of transparency for scientific integrity and public trust, initiative such as DA-RT:
- See Thing 4: Transparency

Challenge Me

If you have not written a README or Codebook before, use one of your own research projects to draft both. You can use the following templates as guides, or if the analysis was conducted in Stata or R, try the following tips to generate a basic codebook to build from!

- [INSERT R/STATA TIPS HERE FOR GENERATING CODEBOOKS]
- [INCLUDE TEMPLATES FOR CODEBOOKS AND README FILES]



Thing 6: Openness

Science is a community endeavor that sees maximum benefits when the materials produced in the course of research activities are made fully available to community members to enable them to build upon, reuse, and verify scientific knowledge. Because research compendia serve as the evidence base for published findings, allowing others to access, download, and use them with minimal legal and technical barriers is essential for supporting and promoting computational reproducibility.

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, legal, and technical constraints may make it necessary to define and enforce acceptable uses of the research compendium or any of its component parts.

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 license.

Access to the compendium and its components:

- Data
- Code / software [DISCUSS THE USE OF OPEN SOURCE SOFTWARE OR SOFTWARE-AGNOSTIC FORMATS]
-

On the legal front:

Licenses -- consider pulling out (cf FAIR principles)?

- good = license name [ADD EXAMPLE]
- better = license link [ADD EXAMPLE]
- best = machine readable license and link [ADD EXAMPLE]

[EXPAND ON FAIR PRINCIPLES HERE]

On the tech front:

Protocols etc. [EXPAND HERE]

Long-term preservability/usability? (e.g., addressing anticipated software obsolescence by using formats more likely to be renderable into the future)

[ADD RELEVANT STANDARD(S)]

[MAP TO RELEVANT FAIR PRINCIPLE]

Get started

[EXPAND HERE]

Learn More

[EXPAND HERE]

- Open Knowledge Foundation: Open Definition
  http://opendefinition.org/
- Creative Commons licenses
  https://creativecommons.org/
- GNU licenses
  https://www.gnu.org/licenses/licenses.en.html
- MIT license
  https://opensource.org/licenses/MIT
- BSD licenses
  https://en.wikipedia.org/wiki/BSD_licenses
- Software licensing: Open Source Initiative
  https://opensource.org/licenses
- DCC: How to License Research Data
  https://www.dcc.ac.uk/guidance/how-guides/license-research-data
- Panton Principles for Open Data in Science
  http://pantonprinciples.org/

Challenge Me

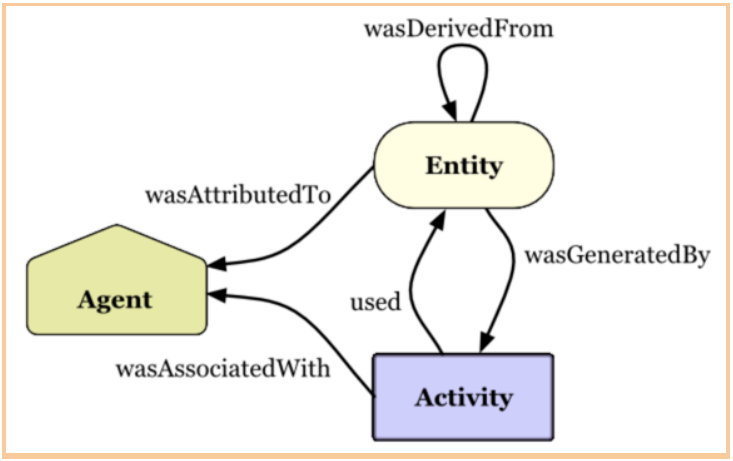
[EXPAND HERE]

Access the following data sources, can you identify the licenses applied to them? What are you permitted to do with the data? If there are no licenses provided, do the data producers provide terms of use or access request instructions? [ADD EXAMPLES, MORE QUESTIONS]

Thing 7: Provenance

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. 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.

Ideally, provenance is captured in standardized machine-readable metadata that records 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. The WC3 PROV data model (PROV-DM) offers a conceptual overview of provenance information for research objects, or entities.



WC3 PROV Data Model (PROV-DM)

[EXPAND TO EXPLAIN: TRACKING, LOGS, CHAIN OF CUSTODY, TRANSFORMATION FOR THE PURPOSE OF AUDITABILITY, VERSIONING]

[EXPAND TO REFER TO AUTHORSHIP, CREDIT HERE]

Applies to **data**

- Any transformation that is 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, we want to know how the subset was derived.
- Or variable transformation
- Need citation

Applies to **code**

- Need citation
- Version control/Versions of files, software (cf FAIR4RS principles)

- Poor coding practices and bugs that break the code

Applies to the **computation**

- Workflow tools?
- Can be cited

[ADD RELEVANT STANDARD(S)]

[MAP TO RELEVANT FAIR PRINCIPLE]

Get started

Where the author got the data

- Where can others get the data
- Persistence of data

Tools that automatically capture provenance information

Learn More

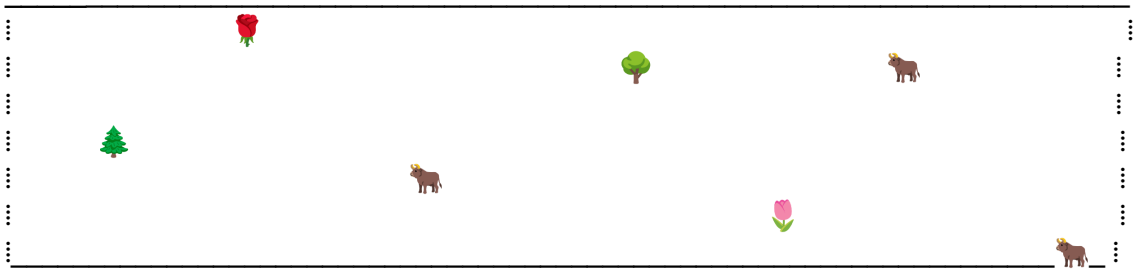
[EXPAND TO ADD CONTEXT FOR LINKS PROVIDED BELOW]

- WC3: An Overview of the PROV Family of Documents (PROV-Overview)
  https://www.w3.org/TR/prov-overview/
- Australian Research Data Commons : Data Provenance
  https://ardc.edu.au/resources/working-with-data/data-provenance/
- Australian National Data Service: Data Provenance
  https://www.ands.org.au/working-with-data/publishing-and-reusing-data/data-provenance
- Science Direct: Data Provenance
  https://www.sciencedirect.com/topics/computer-science/data-provenance
- T7 Workshop: Provenance for Transparent Research
  https://t7.tmcphillips.dev/abstracts/#manir

Challenge Me

- 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?



Thing 8: Metadata / Machine first

Structured and machine-readable metadata about the research compendium and its component parts is preserved and made accessible alongside the compendium. [EXPAND HERE]

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

For example, (e.g., DDI) (CF FAIR PRINCIPLES - EXPAND ON FAIR PRINCIPLES IN THE CONTEXT OF METADATA)

- Study-level [EXPAND HERE - WHAT DOES THIS MEAN, PROVIDE EXAMPLE]
- File-level [EXPAND HERE - WHAT DOES THIS MEAN, PROVIDE EXAMPLE]
- Variable-level (for quant data; other information for qual data) [EXPAND HERE- WHAT DOES THIS MEAN, PROVIDE EXAMPLE]

[ADD RELEVANT STANDARD(S)]

[MAP TO RELEVANT FAIR PRINCIPLE]comment

Get started

[EXPAND HERE]

Learn More

[EXPAND HERE]

- Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18

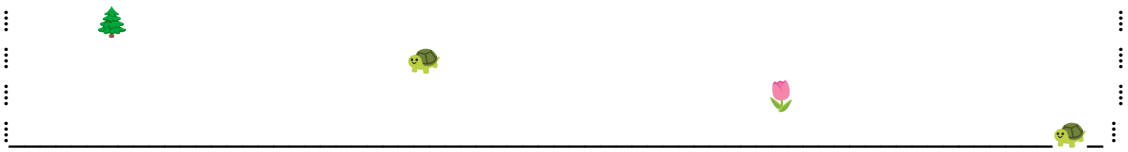
Challenge Me

[EXPAND HERE]

Notes: Standard for communicating research findings/science -- FSRI HSL7 FHIR format

EBM on FHIR

Electronic data exchange standards



Thing 9: Automation

Information about the computational steps throughout the research process is script- or workflow-based, avoiding manual processes to the fullest extent possible. [EXPAND - CONSIDER DIFFERENTIATING FROM ECONOMIZE ABOVE]

- Main steps in a project should be encoded in software and documented so they can be reliably reproduced
- Avoids human error and undocumented idiosyncratic processes typical in manual workflows.
- 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 (see also tacit knowledge!)
- Script-based computation (Dual workflows, see Scott Long TIER Project talk on "dual workflow in which the data flow creates variables and datasets, while the analysis flow analyzes datasets, but does not construct variables or save datasets.")
    - Data transformation
    - Data analysis
    - Figure and table generation
    - Automated testing
- The review and verification should be automated, if possible.
    - Using Jupyter notebook and Binder and GitHub actions can help [EXPAND HERE]
    - Other tools? [EXPAND HERE]

[ADD RELEVANT STANDARD(S)]

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.” -- https://social-science-data-editors.github.io/template_README/

[MAP TO RELEVANT FAIR PRINCIPLE]

Get started

[EXPAND HERE]

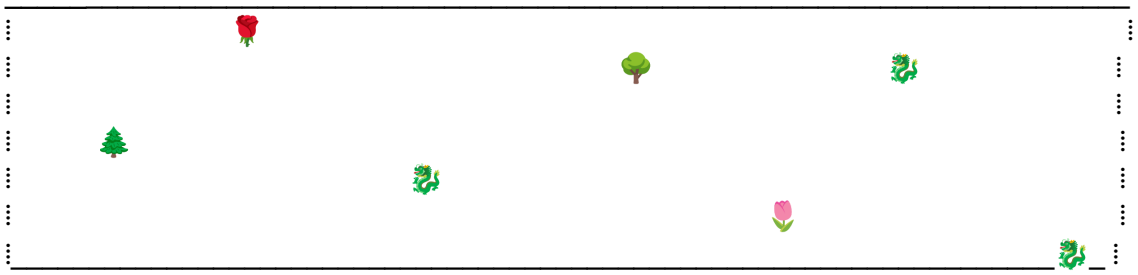
Find a published readme file and compare to this guideline: https://social-science-data-editors.github.io/template_README/

Learn More

[Expand here]

Challenge Me

[EXPAND HERE]



Thing 10: Periodic review

The idea here is 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 the essence of digital preservation. As a curator, you will need to perform periodic reviews, ideally in a scheduled and automated way, and remediate if need be.

[ADD RELEVANT STANDARD(S)]

See UNESCO Recommendation on Open Science (https://en.unesco.org/science-sustainable-future/open-science/recommendation) “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)

[MAP TO RELEVANT FAIR PRINCIPLE]

Get started

Review a bundle of files you're going to submit or are currently curating, or have a look at this example [INSERT EXAMPLE] and consider the following:

- What in the research compendium requires periodic review? Or, what might you worry about?
    - Files:
        - Periodic bit check to ensure file integrity[EXPAND HERE]
        - File format assessment to determine need for format migration to avoid issues due to software obsolescence
    - Data:
        - Access / license changes over time [EXPAND HERE]
    - Code:
        - Reasons it might break over time [EXPAND HERE]
        - Availability of package versions
    - Computation
        - Is there evidence that the researchers did a post-manuscript review of the reproducibility package? [EXPAND HERE]
        - Are file revisions necessary to successfully reproduce the results?

    - Do current computational workspaces continue to support reproducibility?

Learn More

You may like to review the following resources in considering the above questions

[Reproducibility check (add Florio's 15 code review things!)]

Trusted repositories and certification [EXPAND ON WHAT THESE THINGS ARE/MEAN]

- CoreTrustSeal provides a list of repositories that have undergone a thorough audit and review of the practices, services, and infrastructure. Repositories with a CoreTrustSeal have met the 16 requirements deemed necessary characteristics of trustworthy data repositories. (https://www.coretrustseal.org/)
- ISO 16363 is another audit and certification for trustworthy digital repositories. More details can be found here: https://www.iso.org/standard/56510.html
- Center for Research Libraries. (2007). Ten principles. http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/core-re
- Lin, D., Crabtree, J., Ingrid, D., Downs, R. R., Rorie, E., Giaretta, D., . . . Westbrook, J. (2020). The TRUST principles for digital repositories. *Scientific Data, 7*(1) doi:http://dx.doi.org/10.1038/s41597-020-0486-7

[EXPAND ON NDSA LEVELS OF PRESERVATION (SEE LEVEL 4)]

[EXPAND TO TALK ABOUT CONTINUOUS INTEGRATION]

Who is responsible for periodic review? How to plan for periodic review and remediation? Repositories that have demonstrated their trustworthiness through audit and certification have as their mission to preserve objects for long-term discovery, access, and use. One significant benefit of depositing research compendia into a trusted repository is that it is equipped with the necessary policies, technology, and expertise to perform many of the peer review actions described here.

Tools for tracking the review / verification process

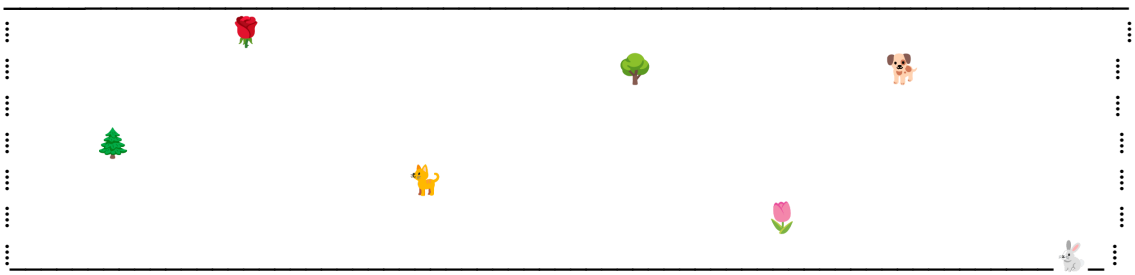
- YARD: https://datascience.codata.org/articles/10.5334/dsj-2020-028/
- SSRP: https://www.bitss.org/introducing-the-social-science-reproduction-platform-a-resource-for-teaching-and-improving-computational-reproducibility/

Challenge Me

Confirming computational reproducibility at a single point in time is not a guarantee for long term reproducibility. What would you want to know about a point-in-time confirmation? [EXPAND HERE WITH MORE QUESTIONS/EXAMPLES]

- When is the repro check done, by whom, was it certified?

- Why might we want to review again? How often?



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