




Scientific Data Management

Notes for the 21st Century Scientist

Louis Moresi¹ ¹ Australian National University

Abstract

We introduce the concept of F.A.I.R. data, and the idea that you have to consciously and deliberately make data available to others. To do this you need to have a data plan as part of your project planning. F.A.I.R. data underpins the idea of reproducible research and this applies to computation as well as to measurements.

Plain Language Summary

Data does not just share itself, to make science open and reproducible, the original data need to be shared and distributed.

Keywords Data Management, Reproducible Research, Reusable Research, F.A.I.R. data, Tutorial

The nature of science is to question everything and to look for the unusual and unexpected. When something unexpected is observed we always have to ask if our understanding of the phenomena need to change or if we just need to be more careful with our measurements !

Extraordinary claims require extraordinary evidence (a.k.a., the Sagan standard) was a phrase made popular by Carl Sagan.

Its roots are much older, however, with the French mathematician Pierre-Simon Laplace Wikipedia stating that: “...*the more extraordinary a fact, the more it needs to be supported by strong evidence*” Also, David Hume wrote in 1748: “A wise man ... proportions his belief to the evidence”, and “No testimony is sufficient to establish a miracle, unless the testimony be of such a kind, that its falsehood would be more miraculous than the fact which it endeavors to establish.” and Marcello Truzzi says: “An extraordinary claim requires extraordinary proof.”

Rational Wiki: Extraordinary Claims ...

At the heart of this quote is the concept of **reproducibility**. That is, if I am sceptical of your claim, I should do the experiment again and see for myself. Peer review is supposed to be a certification that this reproducibility has been validated.

There are a number of ways in which scientific research results can be positively repeated (reproduced, replicated) and the following definitions are widely understood, if not universal.

For my work to be **reproducible**, it means that if you follow my detailed instructions, and use the same experimental setup as me, then you will obtain the same answer as me. This follows from the most elementary principle, that any scientific result that is published has already been repeated by the author sufficiently to expect it to be reproducible by anyone, and that the presence of experimental error has been considered and accounted for in the way that the results are presented. A more demanding form of positive repeatability is the replication of research results. To **replicate** my results, it should be possible to use any equivalent experimental setup, not specifically the one I used, and still obtain the same answer.

Neither of these two forms of repeatable experiments demand that you and I agree on the interpretation of the experiments, only that they can be positively repeated by either one of

LECTURE NOTES

Published Mar 04, 2024

Key Points

- Data management is a requirement of funding
- Data need to be F.A.I.R.
- Reproducible science builds on F.A.I.R. principles

Correspondence to

Louis Moresi
louis.moresi@anu.edu.au

Data Availability

The original document is located on Github at on [GitHub](#)

Funding

N/A

Competing Interests

The authors declare no competing interests.

us. You are also free to disagree with my assertion that the results are meaningful, useful, or interesting !

It is also worth remembering that repeating the work in this way does not automatically make it straightforward to expand upon the ideas in the research and be in a position to modify and build upon the experiment. This is the concept of *re-use* of research results and it requires more effort than reproducibility or replicability.

DATA MANAGEMENT

What do we need to do to make it possible for people to trust our scientific reasoning. They may not agree with our conclusions, what can we do to ensure that they cannot undermine our argument on mere technical grounds and are forced to address our reasoning instead ?

We are going to take a discuss data management from the perspective of *advancing the scientific enterprise* — a high level goal that we can all broadly subscribe to, but which is easy to ignore or forget when we are pressed for time.

Data management starts with the idea of “information management” and we can think about what this means if our information is a computer program or a document. In reality, though, it is much more general and applies to measurements or observations of any kind that we use to justify our scientific arguments.

The Australian Research Council (ARC - which funds much of the academic research in non-medical topics in Australia) introduced a requirement that all research projects need to include a **data management plan**. When this was introduced, it took most of us by surprise, and not many people had thought about what this means or how to devise a plan.

WHAT IS MEANT BY F.A.I.R. DATA ?

F.A.I.R. is an acronym (Findable, Accessible, Interoperable, Reusable) that tries to describe the data life-cycle in [Figure 1](#). At the same time it has an upbeat feel that suggests we can all trust data that is F.A.I.R..



Figure 1: The *data lifecycle* is the starting point for understanding a data management plan and to understand what F.A.I.R. data really means

For a dataset to be considered F.A.I.R. it needs to be

- **Findable:** The data has sufficiently rich metadata and a unique and persistent identifier to be easily discovered by others. This includes assigning a persistent identifier (like a

DOI or Handle), having rich metadata to describe the data and making sure it is findable through disciplinary local or international discovery portals.

- **Interoperable:** The associated data and metadata uses a ‘formal, accessible, shared, and broadly applicable language for knowledge representation’. This involves using community accepted languages, formats and vocabularies in the data and metadata. Metadata should reference and describe relationships to other data, metadata and information through identifiers.
- **Accessible:** The data is retrievable by humans and machines through a standardised communication protocol, with authentication and authorisation where necessary. The data does not necessarily have to be open. Data can be sensitive due to privacy concerns, national security or commercial interests. When it’s not able to be open, there should be clarity and transparency around the conditions governing access and reuse.
- **Reusable:** The associated metadata provides rich and accurate information, and the data comes with a clear usage licence and detailed provenance information. Reusable data should maintain its initial richness. For example, it should not be diminished for the purpose of explaining the findings in one particular publication. It needs a clear machine readable licence and provenance information on how the data was formed. It should also use discipline-specific data and metadata standards to give it rich contextual information that will allow reuse.

Read more about F.A.I.R. data at the [Australian Research Data Commons F.A.I.R. data](#) page.

You can see parallels between the ideas of F.A.I.R. data and the open-source software community. Both are built upon the premise that more gets done if everyone shares the load. Open source software is normally described as a community working together (a positive vibe) whereas sometimes (often) reproducible science starts from the idea that untrustworthy results are weeded out (a little bit more negative, in my mind).

DISCUSSION

The Australian Research Council (ARC - which funds much of the academic research in non-medical topics in Australia) introduced a requirement that all research projects need to include a **data management plan**. When this was introduced, it took most of us by surprise, and not many people had thought about what this means or how to devise a plan.

We are going to take a discuss data management from the perspective of *advancing the scientific enterprise* – a high level goal that we can all broadly subscribe to, but which is easy to ignore or forget when we are pressed for time.

Data management starts with the idea of “information management” and we can think about what this means if our information is a computer program or a document. In reality, though, it is much more general and applies to measurements or observations of any kind that we use to justify our scientific arguments.

What to we need to do to make it possible for people to trust our scientific reasoning. They may not agree with our conclusions, what can we do to ensure that they cannot undermine our argument on mere technical grounds and are forced to address our reasoning instead ?

ACTIVITY (1)

VIDEO / Presentation (reproducible research)

QUESTIONS

- Why should our data be freely available ?
- Are there any reasons to keep thing secret ?

- Why should our methods / software be open ?
- Is there something like open source for research equipment ?

ACTIVITY (2)

I would like you to understand a little about how we can share data, measurements and written information. We can start simply by considering how we share and collaborate on a document.

We'll look at the source of this particular document and how we can share it on GitHub (www.github.com). The discussion that we have while we do this should allow you to appreciate the following:

- Provenance: How can we track the data or the document back through time to see all the steps along the path to its present state.
- Version control / Revision control: Formalising the way we track changes to our information (in documents or software we track versions and the changes between them). What does this mean for a dataset ?
- Meta-data: The description of the dataset or the document is considered to be a form of data in its own right and is usually called “meta-data”. A good example that you will be familiar with is to pull up the information on a photo from your phone.

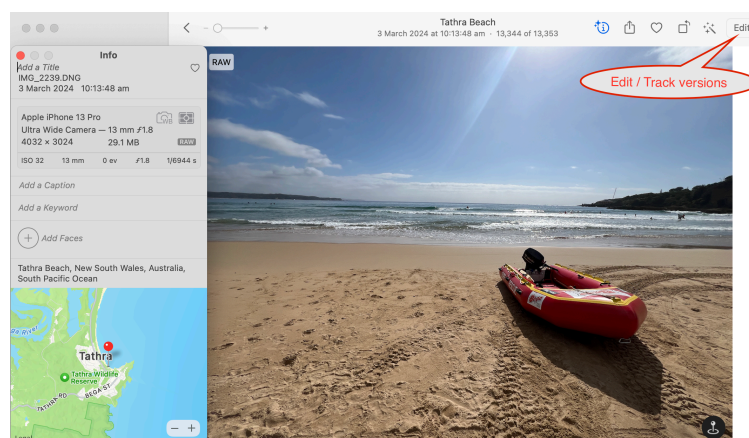


Figure 2: An example of meta-data that you will be familiar with already.

- DOI: a “digital object identifier” is a persistent URL that directs you to a piece of information. This can be a dataset, a publication, a blog post, what is important is that it can be guaranteed to work in perpetuity (at least in theory).

EXAMPLES OF OPEN DATA STATEMENTS

AUSTRALIAN RESEARCH COUNCIL

Effective data management is an important part of ensuring open access to publicly funded research data. Data management planning from the beginning of a research project helps to outline how data will be collected, formatted, described, stored and shared throughout, and beyond, the project lifecycle.

Since February 2014, the ARC has required researchers to outline how they plan to manage research data arising from ARC-funded research. From 2020, this requirement forms part of the agreement for funding under the National Competitive Grants Program.

The requirement is consistent with the responsibilities outlined in the Australian Code for the Responsible Conduct of Research 2018, which include the

proper management of research data and primary materials by researchers, along with institutional policies addressing data ownership, storage, retention and “appropriate access...by the research community”.

The OECD Principles and Guidelines for Access to Research Data from Public Funding (2007) also provide guidance on the management of data and primary materials. The ARC notes that Australia, as an OECD member, is expected (not legally bound) to implement these principles and guidelines.

The ARC’s requirement is designed to encourage researchers to consider the ways in which they can best manage, store, disseminate and reuse data. Researchers, in consultation with institutions, have a responsibility to consider the management and future potential of their research data, taking into account the particular approaches, standards and uses for data that may exist in different institutions, disciplines and research projects. Some institutions may have infrastructure and/or processes in place for storing, managing and sharing data – these are valuable resources that should be utilised.

The ARC does not require that full, detailed data management plans be submitted for assessment, but from 2020 will require that such plans are in place prior to the commencement of the project. Currently, the ARC does not mandate open access to data.

THE ROYAL SOCIETY

The Royal Society supports science as an open enterprise and is committed to ensuring that data outputs from research supported by the Society are made publicly available in a managed and responsible manner, with as few restrictions as possible. Data outputs should be deposited in an appropriate, recognised, publicly available repository, so that others can verify and build upon the data, which is of public interest. To fully realise the benefits of publicly available data they should be made intelligently open by fulfilling the requirements of being discoverable, accessible, intelligible, assessable and reusable.

The Royal Society does not dictate a set format for data management and sharing plans. Where they are required, applicants should structure their plan in a manner most appropriate to the proposed research. The information submitted in plans should focus specifically on how the data outputs will be managed and shared, detailing the repositories where data will be deposited. In considering your approach for data management and sharing, applicants should consider the following:

- What data outputs will be generated by the research that are of value to the public?
- Where and when will you make the data available?
- How will others be able to access the data?
- If the data is of high public interest, how will it be made accessible not only for those in the same or linked field, but also to a wider public audience?
- Specify whether any limits will be placed on the data to be shared, for example, for the purposes of safeguarding commercial interests, personal information, safety or security of the data.
- How will datasets be preserved to ensure they are of long-term benefit?

NATIONAL SCIENCE FOUNDATION

NSF's data sharing policy:

NSF-funded investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF awards.

See: [NSF: Dissemination and Sharing of Research Results](#)

Plans for data management and sharing of the products of research. Proposals must include a supplementary document of no more than two pages labeled Data Management Plan . This supplement should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results (see AAG Chapter VI.D.4), and may include:

1. the types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project;
2. the standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies);
3. policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements;
4. policies and provisions for re-use, re-distribution, and the production of derivatives; and
5. plans for archiving data, samples, and other research products, and for preservation of access to them.

A valid Data Management Plan may include only the statement that no detailed plan is needed, as long as the statement is accompanied by a clear justification.

RESOURCES

- <https://ardc.edu.au/resource/good-data-practices/>
- GitHub (<https://www.github.com>) — a web-based repository for information, metadata and version control. GitHub is not intended for massive datasets, but it does offer long-lived archives and change tracking.
- Zenodo (<https://www.zenodo.org>) - a data archive that guarantees permanence for your data and provides a DOI for anything you upload. It is possible to have versions of data but every version is treated as a unique artefact and is considered whole and complete. You can use Zenodo to create a snapshot of your GitHub repository (e.g. a release) and give it a DOI, a fact that might help you to understand the difference in these two approaches to data management (and when you would choose one over the other). [<https://zenodo.org/communities/auscope/records?q=&l=list&p=1&s=10&sort=newest>](Auscope's zenodo) page. It's a bit dry and library-like but that's intentional.
- Figshare (<https://www.figshare.org>) - started out as a place to keep large posters and maps but is now *"a home for papers, F.A.I.R. data and non-traditional research outputs that is easy to use and ready now"*. This is another way to apply for a DOI for your data.
- ArXiv / EarthArXiv - *"arXiv* is a free distribution service and an open-access archive for nearly 2.4 million scholarly articles in the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and

systems science, and economics.” / “Active since 2017, [EarthArXiv](#) is a preprint server devoted to open scholarly communication. EarthArXiv publishes articles from all sub-domains of Earth Science and related domains of planetary science. The EarthArXiv platform assigns each submission a Digital Object Identifier (DOI), therefore assigning provenance and making it citable in other scholarly works.”

- <https://www.geo-down-under.org.au> - this is an Earth Science blog that you are welcome to write articles for if you want. It is a little bit unusual in that all blog posts are given a DOI. Do you think blog posts deserve a DOI ? What about tweets or text messages ?