Open Science 5
DATA MANAGEMENT IN A «FAIR» AND OPEN ENVIRONMENT

Elena Giglia
Torino, Sept. 23-24, 2019
elena.giglia@unito.it

@egiglia
We’ll learn
1. Why should we take care of our data
2. Useful tools to properly manage your data

Take home messages
• “managed”, “FAIR”, “Open”, they are different
• A Data Management Plan (DMP) is not an administrative burden, it’s an how-to manage your data
... I see...

1) Managing data is not an easy task
2) There is no recipe as data are unique
3) So many aspects to be considered
4) So many **tools** to get used to
5) It seems sooooooooo time consuming
6) But the benefits are huuuuuuuuuuuuuge
...let’s talk about data

HOW MANY OF YOU WERE IN TROUBLE IF THE PC BROKE DOWN NOW?

HOW MANY OF YOU WERE IN TROUBLE IF THE USB STICK GOT LOST?

HOW MANY OF YOU WERE IN TROUBLE IF THE FILES ON G-DRIVE DISAPPEARED?

HOW MANY OF YOU WERE IN TROUBLE IF DROPBOX BECAME FOR FEE?

DEFINITION OF «BACKUP»:
WHAT YOU HAD TO DO BEFORE
Why should you take care of your data?

https://www.youtube.com/watch?v=N2zK3sAtr-4&ecver=2

... this is the data steward’s nightmare:
- no backup
- no software
- no data legend
data are fragile

80% will be lost in 20 years

...THAT’S WHY YOU NEED A DATA MANAGEMENT PLAN.
IT’S NOT JUST AN ADMINISTRATIVE BURDEN
WHERE DO YOU STORE YOUR DATA?
Why should you take care of your data?

- Not to lose them
- When data are organized, your research is more efficient
- To improve research integrity
- To allow for checks and validation
- (If open) to be more visible
- (If open) to allow reuse
- (If open) to booster collaborations
- Some kind of data are «unique events» (meteorology, seismology...)
- To be reproducible

«the coolest thing to do with your data will be thought of by someone else» [R. Pollock]

How and why you should manage your research data: a guide for researchers
An introduction to engaging with research data management processes.

JISC Guide
Executive summary

Introduction

Researchers are creating, gathering and using data in hitherto-unimagined volumes. These vast data resources dramatically increase the capacity of science to infer patterns in phenomena, whether physical, chemical, biological or human, or in the complex systems that are at the heart of most global challenges.

SCIENCE IS DATA INTENSIVE. PERIOD

The challenge is clear to us: if we do not act, there might be a looming crisis on the horizon. The vast majority of all data in the world (in fact up to 90%) has been generated in the last two years. Computers have long surpassed individuals in their ability to perform pattern recognition over large data sets. Scientific data is in dire need of openness, better handling, careful management, machine actionability and sheer re-use. One of the sobering conclusions of our consultations was that research infrastructure and communication appear to be stuck in the 20th century paradigm of data scarcity. We should see this step-change in science as an enormous opportunity and not as a threat. The EOSC is a positive 'Cloud on the Horizon' to be realised by 2020. Ultimately, actionable knowledge and translation of its benefits to society will be handled by humans in the 'machine era' for decades to come, machines are just made to serve us.
Why should we care about data?

Data creates a bridge between traditional disciplines, spawning discovery and innovation from the humanities to the hard sciences. Data dissolves barriers, opening up new channels of communication, lines of research, and commercial opportunities. Data will be the engine, the spark to create a better world for all.

Why should we care about data?

The Vienna Declaration on the European Open Science Cloud
Vienna, 23 November 2018

BECAUSE EOSC IS HERE TO STAY

Vienna, Nov.23, 2018

We, Ministers, delegates and other participants attending the launch event of the European Open Science Cloud (EOSC):

1. Recall the challenges of data driven research in pursuing excellent science as stated in the “EOSC Declaration” signed in Brussels on 10 July 2017.

2. Reaffirm the potential of the European Open Science Cloud to transform the research landscape in Europe. Confirm that the vision of the European Open Science Cloud is that of a research data commons, inclusive of all disciplines and Member States, sustainable in the long-term.

3. Recognise that the implementation of the European Open Science Cloud is a process, not a project, by its nature iterative and based on constant learning and mutual alignment. Highlight the need for continuous dialogue to build trust and consensus among scientists, researchers, funders, users and service providers.

4. Highlight that Europe is well placed to take a global leadership position in the development and application of cloud services for Science. Realise that the EOSC is a roadmap and open to the world, reaching out over time to all stakeholders.

5. Recall that the Council of Ministers and other participants agreed to the creation of an EOSC, now the European Open Science Cloud.

9. Call for the European Open Science Cloud to provide all researchers in Europe with seamless access to an open-by-default, efficient and cross-disciplinary environment for storing, accessing, reusing and processing research data supported by FAIR data principles.

6. Note that the 2016 EOSC Summit (held on 11 June 2016) called for acceleration towards making the European Open Science Cloud a reality, hinting at the need to further strengthen the ongoing dialogue across institutions and with stakeholders, for a new governance framework to be launched in Vienna, on 23 November 2018.
THE EUROPEAN OPEN SCIENCE CLOUD?
SOME NUANCES AND DEFINITIONS

Imagine a federated, globally accessible environment where researchers, innovators, companies and citizens can publish, find and re-use each other's data and tools for research, innovation and educational purposes. Imagine that this all operates under well-defined and trusted conditions, supported by a sustainable and just value for money model. This is the environment that must be fostered in Europe and beyond to ensure that European research and innovation contributes in full to knowledge creation, meet global challenges and fuel economic prosperity in Europe. This we believe encapsulates the concept of the European Open Science Cloud (EOSC), and indeed such a federated European endeavour might be expressed as the European contribution to an Internet of FAIR Data and services.

The European Open Science Cloud is a supporting environment for Open Science and not an 'open Cloud' for science.

The EOSC aims to accelerate the transition to more effective Open Science and Open Innovation in a Digital Single Market by removing the technical, legislative and human barriers to the re-use of research data and tools, and by supporting access to services, systems and the flow of data across disciplinary, social and geographical borders. The term European Open Science Cloud requires some reflection to dispel incorrect associations and clarify boundaries; in fact the term 'cloud' is a metaphor to help convey the idea of seamlessness and a commons.
[EOSC is not only based on data, but also on data stewardship]

The number of people with these skills needed to effectively operate the EOSC is, we estimate, likely exceeding half a million within a decade. As we further argue below, we believe that the implementation of the EOSC needs to include instruments to help train, retain and recognise this expertise, in order to support the 1.7 million scientists and over 70 million people working in innovation\textsuperscript{9}. The success of the EOSC depends upon it.
## Data steward Level A - Institutional, coordinating, policy

<table>
<thead>
<tr>
<th>Responsibilities</th>
<th>Activities / tasks</th>
<th>Mappings (see below for reference)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1) Policy / strategy</strong></td>
<td>Responsible for advice on and development, implementation and monitoring of a research data management (RDM) policy and strategy for the research institute, which includes the complete research data life cycle, and supports FAIR data and Open Science, in alignment with the relevant stakeholders and with financial and legal constraints, within the institute and in the context of the institute. The policy is the basis for (project) data management plans (DMP).</td>
<td><img src="image" alt="Mappings" /></td>
</tr>
<tr>
<td></td>
<td>• Develops, implements and monitors the institute's RDM policy.</td>
<td><img src="image" alt="Mappings" /></td>
</tr>
<tr>
<td></td>
<td>• Advises the institute's management on short- and long-term actions to advance RDM in the institute.</td>
<td><img src="image" alt="Mappings" /></td>
</tr>
<tr>
<td></td>
<td>• Assesses and monitors institute’s time and financial investments in relation to the institute’s needs for RDM.</td>
<td><img src="image" alt="Mappings" /></td>
</tr>
<tr>
<td></td>
<td>• Explores new needs, opportunities and trends in RDM.</td>
<td><img src="image" alt="Mappings" /></td>
</tr>
<tr>
<td><strong>2) Compliance</strong></td>
<td>Responsible for compliance of the RDM policy to the Netherlands Code of Conduct for Academic Practice and the General Data Protection Regulation (GDPR), as well as continuous alignment with legal and ethical standards.</td>
<td><img src="image" alt="Mappings" /></td>
</tr>
<tr>
<td></td>
<td>• Ensures compatibility of the RDM policy and monitors compliance.</td>
<td><img src="image" alt="Mappings" /></td>
</tr>
<tr>
<td></td>
<td>• Contacts the institute’s privacy officer, legal advisors or ethical board in case of questions regarding compliance.</td>
<td><img src="image" alt="Mappings" /></td>
</tr>
<tr>
<td></td>
<td>• Translates policies from legal/privacy officer to the institute’s practice.</td>
<td><img src="image" alt="Mappings" /></td>
</tr>
<tr>
<td></td>
<td>• Develops and/or guides standard solutions for recurring data issues and for data classification, including input for the DPIA.</td>
<td><img src="image" alt="Mappings" /></td>
</tr>
</tbody>
</table>

## Data steward Level B – Project focused, operational, supporting

<table>
<thead>
<tr>
<th>Responsibilities</th>
<th>Activities / tasks</th>
<th>Mappings (see below for reference)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1) Policy / strategy</strong></td>
<td>Responsible for the development and implementation of a data management plan (DMP) for departments, projects or data collections within the institute, in line with the institute's RDM policy and within financial and legal constraints, that supports FAIR data and Open Science.</td>
<td><img src="image" alt="Mappings" /></td>
</tr>
<tr>
<td></td>
<td>• Develops DMP templates tailored for the departments, projects or data collections within the institute.</td>
<td><img src="image" alt="Mappings" /></td>
</tr>
<tr>
<td></td>
<td>• Writes and/or supports researchers (including PhD students and students) in writing a DMP for departments, projects and data collections, in line with the institute's RDM policy.</td>
<td><img src="image" alt="Mappings" /></td>
</tr>
<tr>
<td></td>
<td>• Implements RDM as a regular aspect of doing research.</td>
<td><img src="image" alt="Mappings" /></td>
</tr>
<tr>
<td><strong>2) Compliance</strong></td>
<td>Responsible for monitoring compliance of the project or data collection with the DMP, and the Netherlands Code of Conduct for Academic Practice, the General Data Protection Regulation (GDPR), as well as continuous alignment with legal and ethical standards.</td>
<td><img src="image" alt="Mappings" /></td>
</tr>
<tr>
<td></td>
<td>• Monitors and supervises the execution of a project or data collection in line with the DMP.</td>
<td><img src="image" alt="Mappings" /></td>
</tr>
<tr>
<td></td>
<td>• Identifies gaps and takes action if needed.</td>
<td><img src="image" alt="Mappings" /></td>
</tr>
<tr>
<td><strong>3) Alignment with FAIR data principles</strong></td>
<td>Responsible for facilitating and supporting FAIR data.</td>
<td><img src="image" alt="Mappings" /></td>
</tr>
<tr>
<td></td>
<td>• Advises, supports and provides guidelines to researchers (including PhD students and students) on the findability (F) of data, including adequate metadata, persistent identifiers and rich data descriptions.</td>
<td><img src="image" alt="Mappings" /></td>
</tr>
</tbody>
</table>
Mappings to existing data stewardship principles and approaches:
(*: indicates that this topic involves potentially all areas, with emphasis on the indicated areas)

**Research data lifecycle**: [https://www.ukdataservice.ac.uk/manage-data/lifecycle](https://www.ukdataservice.ac.uk/manage-data/lifecycle)

1. creating data; 2. processing data; 3. analysing data; 4. preserving data; 5. giving access to data; 6. reusing data.

**FAIR principles**: [http://www.nature.com/articles/sdata201618](http://www.nature.com/articles/sdata201618)

- findable; A. accessible; 1. interoperable; R. reusable data

**Purdue competence areas** [https://docs.lib.purdue.edu/lib_fsdocs/136](https://docs.lib.purdue.edu/lib_fsdocs/136)

1. databases and data formats; 2. discovery and acquisition of data; 3. data management and organization; 4. quality assurance; 5. data conversion and interoperability; 6. metadata; 7. curation and re-use; 8. cultures of practice; 9. data preservation; 10. data analysis; 11. data visualization; and 12. ethics, including citation of data.

**DAMA knowledge areas** [https://dama.org/content/body-knowledge](https://dama.org/content/body-knowledge)

1. data governance; 2. data architecture; 3. data modelling and design; 4. data storage and operations; 5. data security; 6. data integration and interoperability; 7. documents and content; 8. reference and master data; 9. data warehousing and business intelligence; 10. metadata; and 11. data quality.
Why should we care about data

DATA AB INITIO
K. Birney, 2015

We replicate Reinhart and Rogoff (2010a and 2010b) and find that coding errors, selective exclusion of available data, and unconventional weighting of summary statistics lead to serious errors that inaccurately represent the relationship between public debt and GDP growth among 20 advanced economies in the post-war period. Our finding is that when properly calculated, the average real GDP growth rate for countries carrying a public-debt-to-GDP ratio of over 90 percent is actually 2.2 percent, not −0.1 percent as published in Reinhart and Rogoff. That is, contrary to RR, average GDP growth at public debt/GDP ratios over 90 percent is not dramatically different than when debt/GDP ratios are lower.

We also show how the relationship between public debt and GDP growth varies significantly by time period and country. Overall, the evidence we review contradicts Reinhart and Rogoff’s claim to have identified an important stylized fact, that public debt loads greater than 90 percent of GDP consistently reduce GDP growth.
Why should we care about data

Box 1. Some Research Practices that May Help Increase the Proportion of True Research Findings

- Large-scale collaborative research
- Adoption of replication culture
- Registration (of studies, protocols, analysis codes, datasets, raw data and results)
- Sharing (of data, protocols, materials, software, and results)

Retractions in high impact factor journals

Fang, Casadevall 2011

http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001747
Why should we care about data?

An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.

A PAPER WITHOUT DATA OR SOFTWARE IS MERELY ADVERTISING
No data?

Is withholding your data simply bad science, or should it fall under scientific misconduct?

A recent study sent data requests to 200 researchers, and received 194 responses, of which 63% said ‘data available upon request’. Most data is witheld, and as a community think about those withholding their data, Nicole Janz argues that if you need a good reason to withhold data, you need to be able to justify that reason. Classifying data secrecy as misconduct.

My first talk of the year! Message is going to be that the opposite of ‘open science’ isn’t ‘closed science’ - it’s bad science.

Gold Standard Research Integrity

Open data
Open code
Pre-registration
Version control

Questionable Research Practices

P-hacking
Sloppy statistics
Peer review abuse
Inappropriate research design
Not answering to replicators
Lying about authorships

Scientific Misconduct

Fabrication
Falsification
Plagiarism
Data?

ANYTHING UNDERPINNING A SCIENTIFIC ASSERTION

Wilma van Wezenbeek
@wvanwezenbeek

#osc2018 Wolfram Horstmann wants us to talk about datadiversity, like we do with biodiversity #openscience

https://twitter.com/wvanwezenbeek/status/973527086685093893
5 WAYS TO THINK OF DATA:
- THE WAY DATA ARE COLLECTED
- THEIR FORM
- THEIR FORMAT
- THEIR SIZE/VOLUME
- THE WORKFLOW PHASE THEY ARE IN

- The way the data is collected.
  - By experimenting, simulations, observations, derived data, reference data.

- The data forms.
  - For example text documents, spreadsheets, lab journals, logs, questionnaires, software code, transcripts, code books, audio and video recordings, photos, samples, slides, artefacts, models, scripts, databases, metadata, etc.

- The formats for electronic storage of the research data.
- The size (volume) of the data files.
- The research lifecycle phase the data is in.
3 pillars

http://www.dcc.ac.uk/resources
... and a master

Data Stewardship for Open Science
Implementing FAIR Principles

the worst way imaginable to communicate the outcome of the scientific process. If science has become indeed data driven and data is the oil of the 21st century, we better put data centre stage and publish data as first-class research objects, obviously with supplementary narrative where needed, steward them throughout their life cycle, and make them available in easily reusable format.

Yet another recent study claimed that only about 12% of NIH funded data finds its way to a trusted and findable repository. Philip Bourne, when associate director for data science at the U.S.A. National Institutes of Health coined the term dark data for the 88% that is lost in amateur repositories or on laptops. When we combine the results of the general reproducibility related papers and the findability studies,

Monsense and more... @barendmons · 2 h
Finally! Tomorrow the book goes to the printer: Data Stewardship for Open Science: Implementing FAIR Principles

A good data steward publishes data with a supplementary article(Data+).
...fast track

https://vidensportal.deic.dk/RDMelearn

eScience

Take the course

Module 1: Introduction

Module 2: FAIR principles

Module 3: Data Management Plans

90% of the world’s data was created within the last two years.
THERE ARE COSTS. IT’S TRUE.

BUT HOW MUCH COULD IT COST NOT TO CURATE AND PRESERVE DATA?
Following this approach, we found that the annual cost of not having FAIR research data costs the European economy at least €10.2bn every year. In addition, we also listed a number of consequences from not having FAIR which could not be reliably estimated, such as an impact on research quality, economic turnover, or machine readability of research data. By drawing a rough parallel with the European open data economy, we concluded that these unquantified elements could account for another €16bn annually on top of what we estimated. These results relied on a combination of desk research, interviews with the subject matter experts and our most conservative assumptions.
...one step behind...
RESEARCH DATA IS NOT «MINE»
NO COPYRIGHT AS NO CREATIVITY ON DATA PER SE

Repeat with me: research data is not mine
Seldom do I see something that truly shakes me at work. You know, work is work, I am no neurosurgeon, no médecin sans frontières nor am I a social

Lusoli, Apr. 2017
the 3 steps

Open

FAIR

Managed

How do Open, FAIR & RDM intersect?
1. Data should be managed

Data management is an active process by which digital resources remain discoverable, accessible and intelligible over the longer term, a process that invests data and datasets with the potential to accrue value as assets enjoying far wider use than their creators may have anticipated. In the world of research, such a value-adding process is a significant contributor to the much desired achievement of impact.
2. Data should be FAIR

TO BE FINDABLE:
F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

TO BE ACCESSIBLE:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1 the protocol is open, free, and universally implementable.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
A2 metadata are accessible, even when the data are no longer available.

TO BE INTEROPERABLE:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for their content.
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data include qualified references to other (meta)data.

TO BE RE-USABLE:
R1. meta(data) have a plurality of accurate and relevant attributes.
R1.1. (meta)data are released with a clear and accessible data usage license.
R1.2. (meta)data are associated with their provenance.
R1.3. (meta)data meet domain-relevant community standards.

«ACCESSIBLE» DOES NOT MEAN «OPEN». DATA CAN BE CLOSED, PROVIDED YOU – AND MACHINES - KNOW WHERE TO FIND THEM AND AT WHAT ACCESS CONDITIONS

https://www.force11.org/group/fairgroup/fairprinciples
3. Data COULD be Open

Tim Berners-Lee, the inventor of the Web and Linked Data initiator, suggested a 5-star deployment scheme for Open Data. Here, we give examples for each step of the stars and explain costs and benefits that come along with it:

1. ★ make your stuff available on the Web (whatever format) under an open license

2. ★★★ make it available as structured data (e.g., Excel instead of image scan of a table)

3. ★★★★ make it available in a non-proprietary open format (e.g., CSV instead of Excel)

4. ★★★★★ use URIs to denote things, so that people can point at your stuff

5. ★★★★★★ link your data to other data to provide context
MANAGING DATA

DCC because good research needs good data
Add a "version management" tab to your spreadsheet.

Now, let me expand on this idea.

Start by adding an extra "version management" tab to a new spreadsheet. In this sheet, carefully write down a version name (name of the file, typically) in the first column, in the second column the date, and in a third column an explanation of all changes you made to the sheet. Carefully fill out this sheet every single time you move something around, or tinker with the sheet.

If you’re a starting PhD student, start doing this the very next time you build a new sheet. Thank me later.

If you already have multiheaded monstrous sheets: start by managing them in this way, and take a few extra hours to redefine the logic behind what you did earlier. Your dissertation writing self will thank you.
Plan

In this introductory tour, you will become aware of what data management and a data management plan (DMP) are and why they are important. General concepts such as social science data and FAIR data will be explained. Based on our recommendations and good practice examples, you will be able to start writing your DMP.

Organise & Document

If you are looking for good practices in designing an appropriate data file structure, naming, documenting and organising your data files within suitable folder structures, this chapter is for you.

Store

To be able to plan a storage and backup strategy, you will learn about different storage and backup solutions and their advantages and disadvantages. Also, measures to protect your data from unauthorised access with strong passwords and encryption will be explained.

Protect

This chapter highlights your legal and ethical obligations and shows how a combination of gaining consent, anonymising data, gaining clarity over who owns the copyright to your data and controlling access can enable the ethical and legal sharing of data.

Archive & Publish

When you arrive at this chapter you will have learnt to differentiate between currently available data publication services. You will also find a number of stepping stones on how to promote your data.

Discover

How can you discover and reuse existing or previously collected datasets?
Some [practical] support

AT THE END OF EACH STEP, THERE IS A SECTION «ADAPT YOUR DMP» ACCORDING TO WHAT YOU HAVE JUST LEARNT

Adapt your DMP: part 6

This is the sixth 'Adapt your DMP' section in this tour guide. To adapt your DMP, consider the following elements and corresponding questions:

Versioning

Interoperability

In order to be able to link your work to other research, it might be useful to build on established terminologies as well as commonly uses coding and soft- and hardware wherever this is possible.

- Which software and hardware will you use? How does this relate to other research?

If applicable:

- Will established terminologies/ontologies (i.e. structured controlled vocabularies) be used in the project? If not, how does yours relate to established ones?
- Which coding is used (if any)? How does this relate to other research?

Deposit your data

- Will the data you produce and/or used in the project be useable by third parties, in particular after the end of the project?
- Which data and associated metadata, documentation and code will be deposited?
- What methods or software tools are needed to access the data?
- Is documentation about the software needed to access the data included?
- Is it possible to include the relevant software (e.g. in open source code)?
- What data quality assurance processes will you apply?
...and the SSH?

https://training.parthenos-project.eu/sample-page/manage-improve-and-open-up-your-research-and-data/
File naming conventions

The conventions comprise the following 13 rules. Follow the links for examples and explanations of the rules.

1. Keep file names short, but meaningful.
2. Avoid unnecessary repetition and redundancy in file names and file paths.
3. Use capital letters to delimit words, not spaces or underscores.
4. When including a number in a file name always give it as a two-digit number, i.e. 01-99, unless it is a year or another number with more than two digits.
5. If using a date in the file name always state the date ‘back to front’, and use four digit years, two digit months and two digit days: YYYYMMDD or YYYYMM or YYYY-MM.
6. When including a personal name in a file name give the family name first followed by the initials.
7. Avoid using common words such as ‘draft’ or ‘letter’ at the start of file names, unless doing so will make it easier to retrieve the record.
8. Order the elements in a file name in the most appropriate way to retrieve the record.
9. The file names of records relating to recurring events should include the date and a description of the event, except where the inclusion of any of these elements would be incompatible with rule 2.
10. The file names of correspondence should include the name of the correspondent, an indication of the subject, the date of the correspondence and whether it is incoming or outgoing correspondence, except where the inclusion of any of these elements would be incompatible with rule 2.
11. The file name of an email attachment should include the name of the correspondent, an indication of the subject, the date of the correspondence, ‘attach’, and an indication of the number of attachments sent with the covering email, except where the inclusion of any of these elements would be incompatible with rule 2.
12. The version number of a record should be indicated in its file name by the inclusion of “V” followed by the version number and, where applicable, ‘Draft’.
Data versioning

What do we mean by the term ‘data versioning’?

A version is “a particular form of something differing in certain respects from an earlier form or other forms of the same type of thing”. In the research environment, we often think of versions as they pertain to resources such as manuscripts, software or data. We may regard a new version to be created when there is a change in the structure, contents, or condition of the resource.

In the case of research data, a new version of a dataset may be created when an existing dataset is reprocessed, corrected or appended with additional data. Versioning is one means by which to track changes associated with ‘dynamic’ data that is not static over time.

Why is data versioning important?

Increasingly, researchers are required to cite and identify data to support research reproducibility and trustworthiness. It is particularly challenging where the data to be cited are accessed via a web service.

Numbering system 1

Data versioning follows a similar path to software versioning, usually applying a two-part numbering rule: Major.Minor (e.g. V2.1). Major data revision indicates a change in the formation or content of the dataset that may bring changes in scope, context or intended use. For example, a major revision may increase or decrease the statistical power of a collection, require change of data access interfaces, or enable or disable answering of more or less research questions. A Major revision may incorporate:

- substantial new data items added to/deleted from a collection
- data values changed because temporal and/or spatial baseline changes
- additional data attributes introduced
- changes in a data generation model
- format of data items a changed
- major changes in upstream datasets.

Minor revisions often involve quality improvement over existing data items. These changes may not affect the scope or intended use of initial collection. A Minor revision may include:

- renaming of data attribute
- correction of errors in existing data
- re-running a data generation model with adjustment of some parameters
- minor changes in upstream datasets.

What tools are available for data versioning?

There is no one-size-fits-all solution for data versioning and tracking changes. Data come in different forms and are managed by different tools and methods. In principle, data managers should take advantage of data management tools that support versioning and track changes.

Example approaches include:

Git (and Github) for Data (with size <10Mb or 100k rows) which allows:
- effective distributed collaboration – you can take my dataset, make changes, and share those back with me (and different people can do this at once)
- provenance tracking (i.e. what changes came from where)
- sharing of updates and synchronizing datasets in a simple, effective, way.

Data versioning at ArcGIS

- Users of ArcGIS can create a geodatabase version, derived from an existing version. When you create a version, you specify its name, an optional description, and the level of access other users have to the version. As the owner of the version, you can change these properties or delete a version at any time.
# Good Practice and Guidance – Document Version Control Chart (Draft)

1. Create Document/File
   - Save the document according to file naming guidance/good practice.

2. Document Identification
   - Identify on the document e.g. in header or footer, the author, filename, page number and date the document is created/revised.

3. Version Control Table
   - Versions and changes documented with Version Control Table where significant/formal/project based.

4. Version Number
   - Current version number identified on the first page and where appropriate, incorporated into the header or footer of the document.
   - Version number is included as part of the file name.

5. First Draft Version
   - Named as version "0-1" (no full stops in electronic file names).
   - Subsequent draft versions 0-2, 0-3, 0-4 ...

6. First Final/Approved Version
   - When document is final/approved it becomes version 1-0.

7. Changes to Final Version
   - Changed/revised final version becomes x-1.
   - Subsequent drafts to Final version become e.g. 1-1, 1-2, 1-3 etc.

8. Further Final/Approved Documents
   - Version number increased by "1-0" e.g. 1-0, 2-0, 3-0 etc.
   - e.g. Amendments to Final 1-0 are 1-1, 1-2, 1-3 and as approved becomes 2-0.

[https://www2.le.ac.uk/services/research-data/documents/UoL_VersionControlChart_d0-1.pdf](https://www2.le.ac.uk/services/research-data/documents/UoL_VersionControlChart_d0-1.pdf)
Version control

Version control can be done through:

- Uniquely identifying different versions of files using a systematic naming convention, such as using version numbers or dates (date format should be YYYY-MM-DD, see 'File naming');
  - Record the date within the file, for example, 20010911_Video_Twintowers;
  - Process the version numbering into the file name, for example, HealthTest-00-02 or HealthTest_v2;
  - Don't use ambiguous descriptions for the version you are working on. Who will know whether MyThesisFinal.doc, MyThesisLastOne.doc or another file is really the final version?
- Using version control facilities within the software you use;
- Using versioning software like Subversion (2017);
- Using file-sharing services with incorporated version control (but remember that using commercial cloud services as the Google cloud platform, Dropbox or iCloud comes with specific rules set by the provider of these services. Private companies have their own terms of use which applies for example to copyrights);
- Designing and using a version control table. In all cases, a file history table should be included within a file. In this file, you can keep track of versions and details of the changes which were made. Click on the tab to have a look at an example which was taken from the UK Data Service (2017c).
Support Your Data: A Research Data Management Guide for Researchers

John A Borghi, Stephen Abrams, Daniella Lowenberg, Stephanie Simms, John Chodacki

Abstract

Researchers are faced with rapidly evolving expectations about how they should manage and share their data, code, and other research materials. To help them meet these expectations and generally manage and share their data more effectively, we are developing a suite of tools which we are currently referring to as “Support Your Data”. These tools, which include a rubric designed to enable researchers to self-assess their current data management practices and a series of short guideswhich provide actionable information about how to advance practices as necessary or desired, are intended to be easily customizable to meet the needs of researchers working in a variety of institutional and disciplinary contexts.

Suppl. material 5: Draft Guide - Preparing

Authors: John Borghi

Data type: OpenDocument Text (.odt) file

Brief description: A draft guide that corresponds with the "Getting your data ready for analysis" row of the RDM rubric. Suggested points of customization are highlighted in yellow (discipline-specific) and red (institution-specific).

Filename: Draft Guide - Preparing.odt
Download file (59.52 kb)

Suppl. material 6: Draft Guide - Analyzing

Authors: John Borghi

Data type: OpenDocument Text (.odt) file

Brief description: A draft guide that corresponds with the "Analyzing your data and handling the outputs" row of the RDM rubric. Suggested points of customization are highlighted in yellow (discipline-specific) and red (institution-specific).

Filename: Draft Guide - Analyzing.odt
Download file (51.82 kb)

Suppl. material 7: Draft Guide - Sharing

Authors: John Borghi

Data type: OpenDocument Text (.odt) file

Brief description: A draft guide that corresponds with the "Sharing and publishing your data" row of the RDM rubric. Suggested points of customization are highlighted in yellow (discipline-specific) and red (institution-specific).

Filename: Draft Guide - Sharing.odt
Download file
Training

MANTRA is a free online course for those who manage digital data as part of their research project.

https://mantra.edina.ac.uk/
Data management ABC – Data entry

- Check the completeness of records
- Reduce burden at manual data entry
- Minimise the number of steps
- Conduct data entry twice
- Perform in-depth checks for selected records
- Perform logical and consistency checks
- Automate checks whenever possible
Learn to manage

https://www.fosteropenscience.eu/node/2328
Learn to protect

What are personal data?

+ What are personal data?
+ Protecting personal data
+ Legal requirements - EU General Data Protection Regulation (GDPR)
+ Legal requirements - GDPR research exemptions

This course covers data protection in particular and ethics more generally. It will help you understand the basic principles of data protection and introduces techniques for implementing data protection in your research processes. Upon completing this course, you will know:

- what personal data are and how you can protect them
- what to consider when developing consent forms
- how to store your data securely
- how to anonymise your data

Start the Free Course

Full details

Level of knowledge: Introductory: no previous knowledge is required

Topics
I. Process lawfully, fairly and transparent

The participant is informed of what will be done with the data and data processing should be done accordingly.

II. Keep to the original purpose

Data should be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes.

III. Minimise data size

Personal data that are collected should be adequate, relevant and limited to what is necessary.

IV. Uphold accuracy

Personal data should be accurate and, where necessary kept up to date. Every reasonable step must be taken to ensure that personal data that are inaccurate are erased or rectified without delay.

V. Remove data which are not used

Personal data should be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed.

VI. Ensure data integrity and confidentiality

Personal data are processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures.
**Privacy**

**Personal Data Protection Acts** are present in all European countries and concern general laws regulating the protection of personal data. They are based on European Directive 95/46/EC. This Directive will be replaced in the near future by the General Data Protection Regulation (GDPR), which all EU Member States will have to implement in their national legislation by May 2018.

**Obligations to Report Data Leakage Acts** are additions to the Personal Data Protection Acts. They deal with the publication of personal data and contain sanctions in the form of penalties.

**Medical Treatment Agreement Acts** regulate the use and preservation of personal (patient) data in and for medical research.

**Scientific Medical Research with Humans Acts** regulate scientific research in the medical field, in particular how to handle personal health-related data. These make ethical reviews compulsory for all medical research projects.

**Intellectual Property Rights**

**Copyright Acts** regulate the rights of the creator of a work. One distinguishes between exploitation rights and personal intellectual rights ("moral rights").

**The Database Rights Act** recognises the investments made in creating and/or compiling a database. It is based on European Directive 96/9/EC.11

**Related Rights Acts or Neighbouring Rights Acts** mostly refer to the rights of performers, phonogram producers, and broadcasting organisations.

**Patent Acts** are for the protection of patents. Publication of research results (including data) is restricted during the application stage of a patent.

**Public data**

**Public Records Acts** (Public Archives Acts) oblige all public administration offices and services to preserve their documents and transfer these, after appraisal and selection, to public archives.

**Public Sector Information Acts** (concerning re-usability of public data) are based on European Directive 2013/37/EU12 that focuses on the economic aspects of the re-use of public information. It encourages Member States to make as much of this information as possible available for re-use. This also covers content held by museums, libraries, and archives, but does not apply

**Freedom of information Acts** regulate and enable citizen access to documents held by public authorities or companies carrying out work for a public authority. They do not specifically deal with access to research data.

**Heritage Acts** are relevant for archaeological research data in so far as that they regulate ownership of documentation (data) from archaeological excavations.

**Statistical Information Acts** regulate the competencies of the statistics authorities in data gathering as well in access to data.

**Land Registry Acts** (cadastral information) regulate the competencies of the national land registries and access to their data, with special provisions concerning personal data contained in their various databases.

**Codes of Conduct/Ethical Issues**

**Codes of Conduct**, where these exist on a national level or in an institution, should be taken into account in DMPs. They contain the general principles of good academic teaching and research.

**Codes of Practice** for the use of personal data in scientific and scholarly research are based on the Personal Data Protection Acts11 and prescribe how to handle personal data in research practice.

**Codes of Conduct** for Medical Research regulate how researchers should handle medical personal data. They may be based on Medical Treatment Agreement Acts.
LIBER Webinar: GDPR & What It Means For Researchers

The Privacy Impact Assessment (PIA) Route Planner for Academic Research
Inspired by Harry Beck’s London Metro Map

Webinar Video: GDPR & What It Means For Researchers
The Privacy Impact Assessment (PIA) Route Planner for Academic Research
Inspired by Harry Beck’s London Metro Map

No processing of personal data in your research

Processing (special categories of) personal data of (vulnerable) individuals in your research

No legal ground for processing

Legal ground for processing

No high risk processing

High risk processing

Mitigate risks with appropriate measures

Prior consultation with the supervisory authority

Demonstrate compliance with the GDPR

Implement appropriate technical and organisational measures

Conduct Research

No high risk processing

Stop Research

Re-design Research

Research Design

Erasmus University Rotterdam
marlon.domingus@eur.nl
February 2018

https://surfdrive.surf.nl/files/index.php/s/BPRxchnZ44NZAgW
The Logic of a Privacy Impact Assessment (PIA) for Academic Research

Q1. Do you process (special categories of) personal data of (vulnerable) individuals in your research?

YES

NO Proceed - no measures required for safeguarding privacy.

"Personal Data" (GDPR*, Article 4):
Any information relating to an identified or identifiable natural person: a name, an identification number, location data, an online identifier, one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

"Special Categories of Personal Data (Sensitive Data)“ (GDPR, Article 9):
Data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, the processing of genetic data, biometric data for the purpose of trade union membership, the processing of opinions, religious or philosophical beliefs, or economic, cultural or social identity of that natural person.

Q2. What is the legal ground for this processing?

Lawfulness of Processing (GDPR*, Article 6, 89):
1. The individuals participating in your research have freely given their explicit consent for one or more specific purposes.
2. Your research contributes to a legitimate interest, yet results in no high risks for the individuals participating in the research.
3. Your research has a scientific, historical or statistical purpose, yet results in no high risks for the individuals participating in the research.

Q3. Is this processing a high risk processing?

Criteria for high risk processing (WP29 - DPIA Guideline**):
1. Evaluation or scoring
2. Automated-decision making with legal or similar significant effect
3. Systematic monitoring
4. Sensitive data or data of a highly personal nature
5. Data processed on a large scale
6. Matching or combining datasets
7. Data concerning vulnerable data subjects
8. Innovative use or applying new technological or organisational solutions
9. When the processing itself prevents data subjects from exercising a right or using a service or a contract

** Article 29 Data Protection Working Party: Guidelines on data protection impact assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679. Adopted on 4 April 2017. As last Revised and Adopted on 4 October 2017. Online available at: https://ec.europa.eu/newsroom/document.cfm?doc_id=47711

Principles relating to processing of personal data (GDPR*, Article 5):
Demonstrate compliancy with the principles: lawfulness, fairness, transparency, purpose limitation, data minimisation, accuracy, storage limitation, integrity, confidentiality and accountability.

Records of processing activities (GDPR*, Article 30):
The university shall maintain a digital record of the processing activities in your research to demonstrate compliancy to the GDPR.

This register contains:
1. The name and contact details of the researcher, the research partners and service providers;
2. The purposes of the processing;
3. A description of the categories of data subjects and of the categories of personal data;
4. The categories of recipients to whom the personal data have been or will be disclosed.
5. The operations performed on the personal data,
6. The safeguards and security measures in place to protect the personal data,
7. The recipients or categories of recipients or countries or international organisations to which the personal data are transferred, and the safeguards in place in those countries or international organisations,
8. The period for which the personal data will be stored, or, if that is not possible, the criteria used to determine that period,
9. Any further information the data subject is entitled to know under the provisions of this Regulation (GDPR*, Article 15).

Action

Data protection by design and by default (GDPR*, Article 25):
Implement appropriate technical and organisational measures:

1. Individual participating in your research (data subject). Is the participant well informed, aware of possible risks for her/him and aware of the purpose of the research?
2. Data. Is the data de-identified and encrypted?
3. Access Management. How is access managed and controlled for the PI / team (expanded) / public?
4. Software / Platform. Are the Terms of Service for used software / platform checked (where is the data and who has access and has which usage rights)?
5. Devices. Are devices used safe? Encrypted drive, encrypted communication, strong password / two factor authentication.
6. Partners. Are the research partners / service partners trusted and are appropriate legal agreements made, with regards to roles, rights and responsibilities?
7. Safe and secure collaboration. Is the (cross border) communication to, in and from the) collaboration platform end to end encrypted, are roles and permissions defined and implemented, is logging and monitoring implemented?
8. Risk definition and mitigation. Are risks defined and mitigated? Is a risk audit procedure started?

Action

Prior consultation (GDPR*, Article 36):
1. The Data Protection Officer shall, on behalf of the researcher, consult the supervisory authority, prior to the processing (the research) when the processing would result in a high risk in the absence of measures to mitigate the risk.

Action

Risk definition and mitigation. Are risks defined and mitigated? Is a risk audit procedure started?
anonymizing data

[Image of OpenAIRE's AMNESIA service, which allows users to anonymize sensitive data for sharing with a broad audience. The service offers a trade-off between privacy guarantees and data utility, and can be accessed through a web interface.]

http://catalogue.openaire.eu/service/openaire.amnesia
FAIR data

**Findable**
- discoverable with machine readable metadata, identifiable and locatable by means of a standard identification mechanism

**Accessible**
- available and obtainable to both human and machine

**Interoperable**
- both syntactically parseable and semantically understandable, allowing data exchange and reuse among scientific disciplines, researchers, institutions, organisations and countries

**Reusable**
- sufficiently described and shared with the least restrictive licences, allowing the widest reuse possible across scientific disciplines and borders, and the least cumbersome integration with other data sources
FAIR in a nutshell

**Findable**

The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential principle.

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

**F2. Data are described with rich metadata**

**F3. Metadata clearly and explicitly indicate existence**

**F4. (Meta)data are registered or indexed in community registries**

**Accessible**

Once the user finds the required data, she needs to be able to access the data, including authentication and authorization.

**A1. (Meta)data are retrievable by the protocol**

A1.1 The protocol is open, free, and necessary

A1.2 The protocol allows for an automatically generated access URL

**What does this mean?**

Principle F1 is arguably the most important because it will be hard to achieve other aspects of FAIR without globally unique and persistent identifiers. Hence, compliance with F1 will already take you a long way towards publishing FAIR data (see [10 ways identifiers can help with data integration](https://www.go-fair.org/fair-principles/)).

Globally unique and persistent identifiers remove ambiguity in the meaning of your published data by assigning a unique identifier to every element of metadata and every concept/measurement in your dataset. In this context, identifiers consist of an internet link (e.g., a URL that resolves to a web page that defines the concept such as a particular human protein: [http://www.uniprot.org/uniprot/P98161](http://www.uniprot.org/uniprot/P98161)). Many data repositories will automatically generate globally unique and persistent identifiers to deposited datasets. Identifiers can help other people understand exactly what you mean, and they allow computers to interpret your data in a meaningful way (i.e., computers that are searching for your data or trying to automatically integrate them). Identifiers are essential to the human-machine interoperation that is key to the vision of Open Science. In addition, identifiers will help others to properly cite your work when reusing your data.

Of course, identifiers are one thing, but their meaning is another (see principles I1-I3). F1 stipulates two conditions for your identifier:

1. It must be globally unique (i.e., someone else could not reuse/reassign the same identifier without referring to your data). You can obtain globally unique identifiers from a registry service that uses algorithms guaranteeing the uniqueness of newly minted identifiers.
2. It must be persistent. It takes time and money to keep web links active, so links tend to break anyway—therefore, identifiers that refer to data resources should not depend on web links.
...FAIR data

GUIDELINES to FAIRify data management and make data reusable

PARTHENOS

RDA Webinar with Dr. Michel Dumontier: FAIR principles

Findable Accessible Interoperable Reusable

Principles to enhance the value of all digital resources
data, images, software, web services, repositories,...

Developed and endorsed by researchers, publishers, funding agencies, industry partners.

https://youtu.be/jFekfemq7qU

\[ \text{Introduction} \]

Once upon a time in the beautiful kingdom of Datamania lived a prince named Prince Fairhair. Though he was gentle as ever, and good looking too, his father would not let him choose the love of his life on his own. No, he was destined to marry a woman from the neighbouring kingdom. He did not even know her name, only that she was referred to as My Fair Lady. Before the father of My Fair Lady could accept the marriage, he had a quest for Prince Fairhair. Only by fulfilling the quest, would he be able to marry the princess. His quest was to find out how to turn water into gold. A quest that would require gathering loads of data chests and look for clues that could lead to the recipe.

Luckily, Prince Fairhair was not alone in his quest. One of the castle winged housed a number of wizards who could help him decrypt and investigate the data chests. However, it was impossible for the data wizards to go and hunt for data themselves. Thus to assist them, a huge number of elves were trained to work on the data chests.

Findable #1:

Metadata are assigned globally unique and persistent identifiers

The elves returned one by one to the castle, and some of them were really frustrated. They had been following paths to data chests that had been meticulously described, but somehow the data chests had been removed, just leaving holes in the ground. Fimbie was one of these elves, who came back quite puzzled about some strange codes he had found. He could not decipher them and therefore did not know where to go.

"Look" said Fimbie to the data wizard, "I have this strange code 10.123456789 and I don't know what it means?"

"Oh, these are very useful indeed" said the data wizard.

"We can look up the codes in these huge books. Let me see." S0 is the great country of Datamania, and we should look in the house number 1234. He showed a map to Fimbie in the book. "This is where you should go."

"Are you sure it's still there?" said Fimbie, not wanting to waste a single more step on hunting down data chests he could not find.

"Absolutely. These books are magic. If someone moves the data chest to a new location, the book will know."

"Great" said Fimbie, and took off in a sprint. He soon returned happy carrying a data chest.
3. What FAIR is...

FAIR refers to a set of principles, focused on ensuring that research objects are reusable, and actually will be reused, and so become as valuable as is possible. They deliberately do not specify technical requirements, but are a set of guiding principles that provide for a continuum of increasing reusability, via many different implementations. They describe characteristics and aspirations for systems and services to support the creation of valuable research outputs that could then be rigorously evaluated and extensively reused, with appropriate credit, to the benefit of both creator and user.
4. ...and what FAIR is not

**FAIR is not a standard:** The FAIR guiding principles are sometimes incorrectly referred to as a ‘standard’, even though the original publication explicitly states they are not [25]. The guiding principles allow many different approaches to rendering data and services Findable, Accessible, Interoperable, to serve the ultimate goal: the reuse of valuable research objects. Standards are prescriptive, while guidelines are permissive. We suggest that a variety of valuable standards can and should be developed, each of which is guided by the FAIR Principles. FAIR simply describes the qualities or behaviours required of data resources to achieve – possibly incrementally – their optimal discovery and scholarly reuse.

**FAIR is not equal to RDF, Linked Data, or the Semantic Web** The reference article in Scientific Data [25] emphasises the machine-actionability of data and metadata. This implies (in fact, requires) that resources that wish to maximally fulfil the FAIR guidelines must utilise a widely-accepted machine-readable framework for data and knowledge representations. While there are many such frameworks, we have focused here on the Semantic Web (via RDF and the Web Ontology Language (OWL)) and Linked Data (via the Web Ontology Language (OWL)).

**FAIR is not just about humans being able to find, access, reformat and finally reuse data:** The official press release following the publication of the FAIR Principles states the authors’ position clearly: “The recognition that computers must be capable of accessing a data publication autonomously, unaided by their human operators, is core to the FAIR Principles. Computers are now an inseparable companion in every research endeavour”. In recent surveys, the time reportedly spent by PhD students and other researchers in projects dealing with discovering and reusing multiple data sources – so called ‘data munging’ – has been pegged at 80% [19]. Were these colleagues and their machine-assistants only having to deal with FAIR data and services, this wasted time would be reduced to a fraction of what it is today. The avoidance of time-wasting would be a first return on investment in good data stewardship. To serve this potentially enormous cost reduction, FAIR compliant (meta)data and services should be actionable by machines without human supervision whenever and wherever possible.

**FAIR is not equal to Open:** The ‘A’ in FAIR stands for ‘Accessible under well defined conditions’. There may be legitimate reasons to shield data and services generated with public funding from public access. These include personal privacy, national security, and competitiveness. The FAIR principles, although inspired by Open Science, explicitly and intentionally acknowledge that some data may need to be held confidential.
Data can be FAIR or Open, both or neither. The greatest benefits come when data are both FAIR and Open, as the lack of restrictions supports the widest possible reuse, and reuse at scale. To maximise the benefits of making FAIR data a reality, and in the context of Open Science initiatives, the FAIR principles should be implemented in combination with a policy requirement that research data should be Open by default - that is, Open unless there is a good reason for restricting access or reuse. In recent European Commission formulations, the maxim ‘as open as possible, as closed as necessary’ has been introduced, which is a helpful articulation of the principles.

**Rec. 17: Align and harmonise FAIR and Open data policy**

Policies should be aligned and consolidated to ensure that publicly-funded research data are made FAIR and Open, except for legitimate restrictions. The maxim ‘as Open as possible, as closed as necessary’ should be applied proportionately with genuine best efforts to share.
The system of incentives and rewards must also be addressed in a fundamental way. From the perspective of measuring and rewarding research contributions, the full diversity of outputs should be taken into account including FAIR data, code, workflows, models, and other digital research objects as well as their curation and maintenance. In the 21st century, traditional publications and journal articles are far from being the only significant contributions to the advancement of knowledge.

...STOP WORSHIPPING IMPACT FACTOR

...AND JOURNAL ARTICLES

...SIGN DORA?
OPENING DATA
Why Open Data?

Wilma van Wezenbeek
@wvanwezenbeek

#osc2018 @sjDCC I really like what Sarah said just now "There is more risk in losing your data than sharing your data #opencience"

https://twitter.com/wvanwezenbeek/status/97350245711537408

"Open data is like a renewable energy source: it can be reused without diminishing its original value, and reuse creates new value."

https://www.youtube.com/watch?v=HJboOAaJ1I&feature=youtu.be

https://twitter.com/wvanwezenbeek/status/97350245711537408
Behaviours

People will contact me to ask about stuff

Christopher and Alex (C&A) say: “This is usually an objection of people who feel overworked and that [data sharing] isn’t part of their job...” I would add to this that science is all about learning from each other – if a researcher is opposed to the idea of discussing their datasets, collaborating with others, and generally being a good science citizen, then they should be outed by their community as a poor participant.

People will misinterpret the data

C&A suggest this: “Document how it should be interpreted. Be prepared to help and correct such people; those that misinterpret it by accident will be grateful for the help.” From the UK Data Archive: “Producing good documentation and providing contextual information for your research project should enable other researchers to correctly use and understand your data.”

It’s worth mentioning, however, a second point C&A make: “Publishing may actually be useful to counter willful misrepresentation (e.g. of data acquired through Freedom of Information legislation), as one can quickly point to the real data on the web to refute the wrong interpretation.”

My data is not very interesting

C&A: “Let others judge how interesting or useful it is — even niche datasets have people that care about them.” I’d also add that it’s impossible to decide whether a dataset has value to future research. Consider the many datasets collected before “climate change” was a research topic which have now become invaluable to documenting and understanding the phenomenon. From the UK Data Archive:

I might want to use it in a research paper

Anyone who’s discussed data sharing with a researcher is familiar with this excuse. The operative word here is might. How many papers have we all considered writing, only to have them shift to the back burner due to other obligations? That said, this is a real concern.

C&A suggest the embargo route: “One option is to have an automatic or optional embargo; require people to archive their data at the time of creation but it becomes public after X months. You could even give the option to renew the embargo so only things that are no longer cared about become published, but nothing is lost and eventually everything can become open.” Researchers like to have a say in the use of their datasets, but I would caution to have any restrictions default to sharing. That is, after X months the data are automatically made open by the repository.

I would also add that, as the original collector of the data, you are at a huge advantage compared to others that might want to use your dataset. You have knowledge about your system, the conditions during collection, the nuances of your methods, et cetera that could never be fully described in the best metadata.

I’m not sure I own the data

My data is too complicated.

C&A: “Don’t be too smug. If it turns out it’s not that complicated, it could harm your professional [standing].” I would add that if it’s too complicated to share, then it’s too complicated to reproduce, which means it’s arguably not real scientific progress. This can be solved by more documentation.

My data is embarrassingly bad

C&A: “Many eyes will help you improve your data (e.g. spot inaccuracies)… people will accept your data for what it is.” I agree. All researchers have been on the back end of making the sausage. We know it’s not pretty most of the time, and we can accept that. Plus it helps you strive will be at managing and organizing data during your next collection phase.

It’s not a priority and I’m busy

Good news! Funders are making it your priority! New sharing mandates in the OSTP memorandum state that any research conducted with federal funds must be accessible. You can expect these sharing mandates to drift down to you, the researcher, in the very near future (6-12 months).

http://carlystrasser.net/closed-data-excuses-excuses/

Closed Data… Excuses, Excuses
... «as open as possible»...

2/4 "Open as possible, as closed as necessary" is the new principle for all #data from publicly funded #research in Europe #openaccess
Open Science: why just your data?

You can make your workflow more open by ...

- adding alternative evaluation, e.g. with altmetrics
- communicating through social media, e.g. Twitter
- sharing posters & presentations, e.g. at FigShare
- using open licenses, e.g. CC0 or CC-BY
- publishing open access, ‘green’ or ‘gold’
- using open peer review, e.g. at journals or PubPeer
- sharing preprints, e.g. at OSF, arXiv or bioRxiv
- using actionable formats, e.g. with Jupyter or CoCalc
- open XML-drafting, e.g. at Overleaf or Authorea
- sharing protocols & workfl., e.g. at Protocols.io
- sharing notebooks, e.g. at OpenNotebookScience
- sharing code, e.g. at GitHub with GNU/MIT license
- sharing data, e.g. at Dryad, Zenodo or Dataverse
- pre-registering, e.g. at OSF or AsPredicted
- commenting openly, e.g. with Hypothes.is
- using shared reference libraries, e.g. with Zotero
- sharing (grant) proposals, e.g. at RIO