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

Venice, March 29th 2019
elena.giglia@unito.it
The road ahead today

CREATE AWARENESS AMONG RESEARCHERS!!!!!!!

...why should we care about data?

...managed, FAIR, Open: it’s different

...DMPs: a pillar, not an administrative burden

...useful tools + ideas to start/support DM
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 soooooooo time consuming
6) But the benefits are huuuuuuuuuuuge
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 / 1

Figure 1: Causes of Data Loss

DEFINITION OF «BACKUP»:
WHAT YOU HAD TO DO BEFORE

S. Aliprandi, Sicurezza dati e privacy (le norme) 2017
Scientists losing data at a rapid rate
Decline can mean 80% of data are unavailable after 20 years.

Elizabeth Gibney & Richard Van Noorden
19 December 2013

80% will be lost in 20 years

MISSING DATA
As research articles age, the odds of their raw data being extant drop dramatically.

...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 LOOSE THEM
- WHEN DATA ARE ORGANIZED, YOUR RESEARCH IS MORE EFFICIENT
- TO ALLOW FOR CHECKS AND VALIDATION
- TO IMPROVE RESEARCH INTEGRITY
- (IF OPEN) TO BE MORE VISIBLE
- (IF OPEN) TO ALLOW REUSE
- (IF OPEN) TO BOOSTER COLLABORATIONS

Some kind of data are «unique events» (meteoreology, sismology...)

«The coolest thing to do with your data will be thought of by someone else» [R. Pollock]
Why should we care about data?

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?

...SELF-INTEREST: IT’S MANDATORY. MORE AND MORE JOURNALS REQUIRE DATA AVAILABILITY

*Spinal Cord* is committed to increasing data sharing. Now, when submitting their review, peer reviewers are asked to reflect on this issue and respond to the following question:

**Do the authors address data availability?**

This reminds peer reviewers and thus the authors that sharing of data with researchers, clinicians, study participants, and patients is an obligation. This approach is likely to be effective: all improvement over 2015 came in 2018, when this question was introduced. *Spinal Cord* now also makes it clear in its decision letters that suitability for publication will be influenced by authors’ willingness to share data. This issue of *Spinal Cord* contains an article addressing in detail the why and how of making one’s data available, without infringing on the rights of privacy.
The Vienna Declaration on the European Open Science Cloud
Vienna, 23 November 2018

BECAUSE EOSC IS HERE TO STAY

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. Realising this requires that we, as a community, reach out over time to all stakeholders and stakeholders, reaching out over time to all stakeholders and creating a roadmap and the federated infrastructure.

5. Recall that the Council of European Ministers committed the Council of Ministers to providing the necessary conditions for the implementation of the European Open Science Cloud. Emphasize that the European Open Science Cloud is 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.

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.

Note that the 2018 EOSC Summit (held on 11 June 2018) 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.
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 innovation⁹. The success of the EOSC depends upon it.
Why should we care about data?

K. Birney, 2015

Il debito pubblico deprime la crescita? Il clamoroso errore di Carmen Reinhart e Kenneth Rogoff 2013

Herndon, 2013

Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff

Thomas Herndon, Michael Ash, Robert Pollin
April 15, 2013

Herndon, 2013

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 and ‘ex’ 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 RK, 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 methods)

...RETRACTIONS IN HIGH IMPACT FACTOR JOURNALS

Fang, Casadevall 2011
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.

WaveLab and Reproducible Research

Jonathan B. Buckheit and David L. Donoho

Stanford University, Stanford CA 94305, USA

1995

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

To me, data are like footnotes. I might not always read them, but I get suspicious if they are not there.

https://twitter.com/alastairdunning/status/968453078218395648
Data? 

ANYTHING UNDERPINNING A SCIENTIFIC ASSERTION

Wilma van Wezenbeek @wvanwezenbeek

#osc2018 Wolfram Horstmann wants us to talk about datadiversity, like we do with biodiversity #openculture
Research data is the material underpinning a research assertion.
**Data jargon**

A variety of organisations and perspectives on data has led to different definitions. In the course we use the definitions below.

<table>
<thead>
<tr>
<th>Data archive</th>
<th>A data archive is a facility which moves data to an environment for long-term retention. A data archive is indexed and has search facilities, enabling data to be retrieved.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data format</td>
<td>The way in which data or information is coded and stored. A data format (or file format) gives information on how to process the data.</td>
</tr>
</tbody>
</table>

**RDF**

RDF is a standard model for data interchange on the Web (see [http://www.w3.org/RDF/](http://www.w3.org/RDF/)).

**Research data**

Data are facts, observations or experiences on which an argument or theory is based. (see [http://ands.org.au/guides/what-is-research-data.pdf](http://ands.org.au/guides/what-is-research-data.pdf)).

**Resolver**

A system that brings about the link between a persistent identifier and the location where the object is currently situated.

**Text- and data mining**

The computer-based process of deriving or organising information from text or data. It works by copying large quantities of material, extracting the data, and recombining it to identify patterns, trends and hypotheses or by providing the means to organise the information mined. (see [www.ipo.gov.uk/ipreview-doc-t.pdf](http://www.ipo.gov.uk/ipreview-doc-t.pdf)).
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.
Part I

Five Ways To Think About Research Data

Science has progressed by ‘standing on the shoulders of giants’ and for centuries research and knowledge has been shared through the publication and dissemination of books, papers and scholarly communications. Moving forward much of our understanding builds on (large scale) data sets which have been collected or generated as part of this scientific process of discovery. How will this be made available for future generations? How will we ensure that, once collected or generated, others can stand on the shoulders of the data we produce?

Deciding on how to look after data depends on what your data looks like and what needs to be done with it. You should find out if your discipline already has standard practices and use them. We hope that this brief introduction will give some templates of what is already being done in a few disciplines and enable you to start thinking about what you might do with your research data to make it accessible to others.

Further University of Southampton guidance can be found on the library’s web site http://library.soton.ac.uk/researchdata. Any research data management questions can be emailed to researchdata@soton.ac.uk.

This part of the guide introduces five ways of looking at research data.

### 1 Research data collection

The first way of thinking about research data is where it comes from (Research Information Network, 2008). Each of the case studies in Part II illustrates one of these categories.

| Reference data: | Example: the reference human genome sequence in Case Study 1. A data set that can be used for validation, comparison or information lookup. |
| Scientific experiments: | Example: materials engineering fatigue test in Case Study 2. Data generated during a scientific experiment. |
| Models or simulations: | Example: CFD helicopter rotor wake simulation in Case Study 3. Data generated on computer by an algorithm, mathematical model, or the simulation of an experiment. A computer simulation can help when experiments are too expensive, time consuming, dangerous or even impossible to perform. |
| Derived data: | Example: chemical structures in chemistry in Case Study 4. A data set created by taking existing data and performing some manipulation to it. Each data set requires careful curation because the original data may be needed to understand the new data. |
| Observations: | Example: archaeological dig in Case Study 5. Data generated by recording observations of a specific, possibly unrepeatable, event at a specific time or location. |

<table>
<thead>
<tr>
<th>Types of research data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research can come in many different forms, some electronic and some physical. Here are some examples:</td>
</tr>
<tr>
<td>- Electronic text documents, e.g. text, PDF, Microsoft Word files</td>
</tr>
<tr>
<td>- Spreadsheets</td>
</tr>
<tr>
<td>- Laboratory notebooks, field notebooks and diaries</td>
</tr>
<tr>
<td>- Questionnaires, transcripts and code-books</td>
</tr>
<tr>
<td>- Audiotapes and videotapes</td>
</tr>
<tr>
<td>- Photographs and films</td>
</tr>
<tr>
<td>- Examination results</td>
</tr>
<tr>
<td>- Specimens, samples, artefacts and slides</td>
</tr>
<tr>
<td>- Digital objects, e.g. figures, videos</td>
</tr>
<tr>
<td>- Database schemas</td>
</tr>
<tr>
<td>- Database contents</td>
</tr>
<tr>
<td>- Models, algorithms and scripts</td>
</tr>
<tr>
<td>- Software configuration, e.g. case files</td>
</tr>
<tr>
<td>- Software pre-process files, e.g. geometry, mesh</td>
</tr>
<tr>
<td>- Software post-process files, e.g. plots, comma-separated value data (CSV)</td>
</tr>
<tr>
<td>- Methodologies, workflows, standard operating procedures and protocols</td>
</tr>
<tr>
<td>- Experimental results</td>
</tr>
<tr>
<td>- Metadata (data describing data), e.g. environmental conditions during experiment</td>
</tr>
<tr>
<td>- Other data files, e.g. literature review records, email archives</td>
</tr>
</tbody>
</table>

### 3 Electronic storage

The third way to think about research data is how it is stored on a computer. Here are some of the categories of electronic data:

<table>
<thead>
<tr>
<th>Textual, e.g.</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Flat text files</td>
</tr>
<tr>
<td>- Microsoft Word</td>
</tr>
<tr>
<td>- PDF</td>
</tr>
<tr>
<td>- RTF</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Numerical, e.g.</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Excel</td>
</tr>
<tr>
<td>- CSV</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Multimedia, e.g.</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Image (JPEG, TIFF, DICOM)</td>
</tr>
<tr>
<td>- Movie (MPEG, AVI)</td>
</tr>
<tr>
<td>- Audio (MP3, WAV, OGG)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Structured, e.g.</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Multi-purpose (XML)</td>
</tr>
<tr>
<td>- Relational (MySQL database)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Software code, e.g.</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Java</td>
</tr>
<tr>
<td>- C</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Software specific, e.g.</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Mesh</td>
</tr>
<tr>
<td>- Geometry</td>
</tr>
<tr>
<td>- 3D CAD</td>
</tr>
<tr>
<td>- Statistical model</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Discipline specific, e.g.</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Flexible Image Transport System (FITS) in astronomy</td>
</tr>
<tr>
<td>- Crystallographic Information File (CIF) in chemistry</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Instrument specific, e.g.</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Olympus Confocal Microscope Data Format</td>
</tr>
<tr>
<td>- Carl Zeiss Digital Microscopic Image Format (ZVI)</td>
</tr>
</tbody>
</table>

Data can be born digitally, such as a simulation, or ingested into a computer, such as scanning a photograph. Some data can remain in a non-digital format.
Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle.

Data curation: for long term preservation

Data management: managing data through the entire lifecycle

Data steward: expert of the field, who support in data management

http://www.dcc.ac.uk/resources/curation-lifecycle-model
The digital curation lifecycle

Digital curation and data preservation are ongoing processes, requiring considerable thought and the investment of adequate time and resources. You must be aware of, and undertake, actions throughout the data lifecycle.

Conceptualise: conceive and plan the creation of digital objects, capture methods and storage options.

Create: produce digital objects and assign unique identifiers and technical archival metadata.

Access and use: ensure that designated users are able to access data on a day-to-day basis. Some digital objects may be password protected.

Appraise and select: evaluate digital objects for short-term and long-term curation and preservation. Adhere to relevant standards and legal requirements.

Ingest: transfer digital objects to an archive, trusted digital repository, data centre or similar, again adhering to documented guidance, policies and legal requirements.

Preservation action: undertake actions to ensure the long-term preservation and retention of the authoritative nature of digital objects.

Reappraise: return digital objects that fail validation procedures for further appraisal and reselection.

Store: keep the data in a secure manner as outlined by relevant standards.

Access and reuse: ensure that data are accessible to designated users for first time use and reuse. Some material may be publicly available, whilst other data may be password protected.

Transform: create new digital objects from the original, for example, by migration into a different form. This may involve conversion or documentation of preservation. Documented guidance, policies and legal requirements may require the secure destruction of these objects.
3 pillars

http://www.dcc.ac.uk/resources

https://www.ands.org.au/

https://dans.knaw.nl/en
... 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

**Data Stewardship for Open Science: Implementing ...**

Data Stewardship for Open Science: Implementing FAIR Principles has been written with the intention of making scientists, funders, and innovators in all disciplines and...

crcpress.com

In conclusion to this paragraph, my statement in 2005: Text-nining? Why bury it first and then mine it again? [Mons, 2005] is still frighteningly relevant.

**A good data steward publishes data with a supplementary article(Data+)**.
Costs of data curation

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.
Impact on...
FAIR data impact on research

**Research Development**
Diverse set of dynamic tasks including writing grant proposals, project submission process, etc.

**Planning of the research project**
The creation of a data management plan is undertaken.

**Collection of the data**
Raw data (i.e., observational or experimental) and/or existing data is collected.

**Analysis of the data**
Data is analysed in order to identify patterns, draw some conclusions, etc. Data analysis can be done using specific tools such as algorithms, etc.

**Integration of the data**
Data from disparate sources are aggregated to form one homogeneous dataset that can be readily analysed.

**Pre-processing and data cleansing**
Data is cleansed and prepared for transformation, integration or analysis.

**Redaction of the report**
Tasks where the researcher is editing and combining multiple sources around a similar theme.

**Registration and publication**
(Meta)data is registered online and published using adequate channels.

**Peer review**
Peer appraisal of the work to be published (i.e., comments from people with relevant expertise or experience).

Activities NOT impacted by FAIR

Activities impacted by FAIR

Cost of not having FAIR data, 2018
...one step beyond...
RESEARCH DATA IS NOT «MINE»
NO COPYRIGHT AS NO CREATIVITY
ON DATA PER SE
How do Open, FAIR & RDM intersect?

- Open
  - External
  - Community benefit
- FAIR data
- Managed data

Sarah Jones, OS2018, March 2018
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
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

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

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

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

★★★★★★ use URIs to denote things, so that people can point at your stuff\(^4\)

★★★★★★★ link your data to other data to provide context\(^5\)
MANAGING DATA

ALL ART HAS BEEN CONTEMPORARY
“A dataset often isn’t just a single Excel worksheet — it’s much more complex”

So what exactly is the problem with keeping research data in the My Documents folder on your computer or in an email box? According to Nuijen, a lot can go wrong. If not stored and preserved properly, data can easily become corrupted, uninterpretable or lost. “A research project takes months or years to complete, so you can imagine that you collect huge amounts of data over such a long period of time. A data set is often not a single Excel worksheet, but a complex set of different files, documents and versions,” she says. “Before you start collecting data, you need to think of how you’ll store all that information in such a way that you and others will still be able to access and make sense of it — whether in five months or five years.”

“Repositories make your data sustainable”

More importantly, there is a growing awareness that good data management benefits individual scientists as well as science as a whole.

Petra Ploeg notices that some Tilburg University researchers are hesitant to deposit their data in a repository because they are concerned they will lose ownership or control over their data. They are afraid researchers will not have a lot of control over what they share, and worry about what will happen to their data or the hard work they’ve done. But if you deposit your data in a repository, that doesn’t mean you’re giving it away. Quite the contrary — repositories make your data sustainable, ensuring your hard work is not wasted.”
... from the beginning

Research Data Management: Get it right from the beginning

Good RDM = Higher quality, efficiency and value for your research

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.
Some support

Data Management Expert Guide

1. Plan
2. Organise & Document
3. Process
4. Store
5. Protect
6. Archive & Publish
7. Discover

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?

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 guides which provide actionable information about how to advance practices as necessary or desired, are intended to be easily customizable to meet the needs of a 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 (51.82 kb)
Some support

Research Data Management (RDM) Support will guide you in managing, sharing, and preserving your precious research data. This website provides info and guidance during all stages of the data lifecycle. From the proposal phase and setting up your research, including creating data management plans, to publishing your data at the end of your research project. Read more about the importance of RDM...

Welcome

Research lifecycle

- Setting up research
- Investigating & Experimenting
- Publishing & Outreach
- Training & Awareness
- Data support at your faculty
- FAQ

Store your data

TU Delft offers its employees several options for storing and exchanging research data safely. Solutions range from basic storage to storage where you can store, organize and exchange data with colleagues all over the world.

The table below summarizes the current storage possibilities. The solution that fits your research data depends on your specific (security) needs.

Goal and solution

- I want to easily store data and share it selected others (in and outside of TU Delft).
- I want to store and backup personal data.
- I want to store, share and backup data with faculty colleagues.
- I want to store raw data.
- I want to store, backup and share data with other colleagues from TU Delft.
- I want to regularly exchange data internationally between groups.
- I want to share project information.

Diverse

DataverseNL is specifically designed to store, back-up, organise, annotate and share research data with colleagues all over the world. With this open source application you can grant multiple individuals controlled access to your data.

Some support

How does humanities data tend to be different?

There are problems with sharing and managing the humanistic data, however. First of all, much of it is not digital. Humanists still tend to gravitate toward multimodal knowledge creation systems, hybrid digital and technical worlds that resist norms of deposit and reuse. Second, the semiotic systems of humanities data can be quite personal and individual: we prepare our sources to be useful for us, and what works for our research questions and personal epistemic instruments may not work at all for anyone else. Finally, and perhaps most importantly, cultural data is seldom if ever ‘raw’ and seldom, if ever, under the sole ownership of the researcher him or herself. The records of human activity and creativity belong to everyone and no one, they are often preserved and curated by dedicated public institutions or private publishers. Whatever humanities data is, it is not simple!

https://training.parthenos-project.eu/sample-page/manage-improve-and-open-up-your-research-and-data/
Some support

https://www.ed.ac.uk/information-services/research-support/research-data-service
Some support

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

[Image of the MANTRA website]

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

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: YYYYMDD or YYYYMMDD
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, and to accurately indicate exactly which version of a dataset was used. Determining what version of a dataset to use can be especially challenging where the data to be cited are accessed via a web service.

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.

**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 and/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.
Data management ABC – Versioning

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).
Data management ABC – Versioning

<table>
<thead>
<tr>
<th>University of Leicester</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good Practice and Guidance – Document Version Control Chart (Draft)</td>
</tr>
</tbody>
</table>

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
Data management ABC – Data entry

Data Management Expert Guide

1. Plan
2. Organise & Document
3. Process
   - Data entry and integrity
     - Quantitative coding
     - Qualitative coding
     - Weights of survey data
     - File formats and data conversion
     - Data authenticity
     - Wrap up: Data quality
     - Adapt your DMP: part 3
     - Sources and further reading
4. Store
5. Protect
6. Archive & Publish

- 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
Online training
Research Data Management and Sharing - MOOC

This free five-week Coursera MOOC - created by the Universities of Edinburgh and North Carolina - is designed to reach learners across disciplines and continents.

Subjects covered in the 5-week course follow the stages of any research project. They are:

- Understanding Research Data
- Data Management Planning
- Working with Data
- Sharing Data
- Archiving Data

The MOOC (The Massive Open Online Course) uses the Coursera on-demand format to provide short, video-based lessons and assessments across a five-week period, but learners can proceed at their own pace. Although no formal credit is assigned for the MOOC, Statements of Accomplishment will be available to any learner who completes a course for a small fee.

https://www.coursera.org/learn/data-management
Learning

The structure of this MOOC is still under

1. Open Principles
2. Open Collaboration
3. Reproducible Research and Data Analysis
4. Open Research Data
5. Open Research Software and Open Source
6. Open Access to Research Papers
7. Open Evaluation
8. Public Engagement with Science
9. Open Educational Resources
10. Open Advocacy

https://opensciencemooc.eu/
Learn to manage

Data-driven research is becoming increasingly common in a wide range of academic disciplines, from Archaeology to Zoology, and spanning Arts and Science subject areas alike. To support good research, we need to ensure that researchers have access to good data. Upon completing this course, you will:

- understand which data you can make open and which need to be protected
- know how to go about writing a data management plan
- understand the FAIR principles
- be able to select which data to keep and find an appropriate repository for them
- learn tips on how to get maximum impact from your research data

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
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.
applicable laws

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.
- **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/EU 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 Acts 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.
Webinar Video: GDPR & What It Means For Researchers

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

Erasmus University Rotterdam
marlon.domingus@eur.nl
February 2018
Q1. Do you process (special categories of) personal data of (vulnerable) individuals in your research?

**YES**

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.

**NO**

Proceed - no measures required for safeguarding privacy.

Q3. Is this processing a high risk processing?

**YES**

Criteria for high risk processing (WP29 - DIA 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

**NO**

Stop research or redefine research.

---

"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 uniquely identifying a natural person, data concerning health or data concerning a natural person’s sex life or sexual orientation.

---

**Action**

Records of processing activities (GDPR*, Article 30):

The university shall maintain a digital record of the processing activities in your research to demonstrate compliance 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.

**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**

Principles relating to processing of personal data (GDPR*, Article 5):

Demonstrate compliance with the principles: lawfulness, fairness, transparency, purpose limitation, data minimisation, accuracy, storage limitation, integrity, confidentiality and accountability.

---


** 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]
Understanding the Implications of the GDPR on Research

Dr Scott Summers
University of Essex

Ensuring Compliance with the GDPR in Higher Education
1st February 2018

http://www.insidegovernment.co.uk/uploads/2018/02/Presentation-Scott-Summers-Final.pdf

https://www.insight.mrc.ac.uk/2018/04/16/gdpr-research-changes
anonymizing data

http://catalogue.openaire.eu/service/openaire.amnesia
FAIRIFYING DATA
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
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 FAIR principle.

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

F2. Data are described with rich metadata

F3. Metadata clearly and explicitly includes (meta)data identifiers

F4. (Meta)data are registered or indexed in one or more public repositories

Accessible
Once the user finds the required data, they should be able to access it. This involves the provision of authentication and authorisation, including access control, permissions, and persistency of the data.

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

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

A1.2 The protocol allows for an automated access

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).

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). 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 11-13). 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 become invalid over time. Registry services guarantee capability of that link into the future.
F1. (meta)data are assigned a globally unique and globally accessible identifier.

There are many resources created by the ARDC on the topic of metadata:
- Metadata guide
- Data versioning

The ARDC has information on persistent identifiers on three different levels:
- Persistent identifiers: awareness level
- Persistent identifiers: working level
- Persistent identifiers: expert level

It is also a provider of services for minting persistent identifiers of many different levels of the data being identified:
- Digital Object Identifier (DOI) System for research data
- Handle minting Service (Identify My Data)
- International Geo Sample Numbers (IGSN)

Complementary to the assignment of persistent identifiers is their proper


Rob Hooft, DTL
...FAIR data

GUIDELINES to FAIRify data management and make data reusable

PARTHENOS

2019

Introduction

Once upon a time, in the beautiful kingdom of Datamania lived a prince named Prince Fairhair. Though he was gentle as a fox 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 MyFair 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 wing houses 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 gather the chests and bring them to the castle.

Findable #1:

Metadata are assigned globally unique and persistent identifiers

The elves returned done 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 were removed, just leaving holes in the ground. Fairhair 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 Fairhair to the data wizard. "I have this strange code 10.123456789 and I don't know what it means?"

"Oh, those are very useful indeed" said the data wizard. "We can look up the codes in these huge books. Now let me see. 123456789 is the great country of Datamania, and we should look in the house number 1234. He showed a map to Fairhair in the book. This is where you should go.

"Are you sure it's still there?" said Fairhair, 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." said Fairhair, 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 representation, such as RDF. While these may share some of the goals of the FAIR principles.

**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 firmly encourage and recommend the dissemination and reuse of these resources.
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?
FAIR data - components

- POLICIES
- DATA MANAGEMENT PLANS
- PERSISTENT IDENTIFIERS
- STANDARDS
- REPOSITORIES

FAIR Digital Objects sit in a wider FAIR ecosystem comprising services and infrastructures for FAIR. The realisation of FAIR relies on, at a minimum, the following essential components: policies, DMPs, identifiers, standards and repositories. In this ecosystem, data policies are issued by several stakeholders and help to define and regulate requirements for the running of data services. Data Management Plans provide a dynamic index that articulates the relevant information relating to a project and linkages with its various FAIR components. Persistent Identifiers are assigned to many aspects of the ecosystem including data, software, institutions, researchers, funders, projects and instruments. Specifications and standards are relevant in many ways, from metadata, vocabularies and ontologies for data description to transfer and exchange protocols for data access, and standards governing the certification of repositories or composition of DMPs. Repositories offer databases and data services and should be certified to ensure trust.
FAIR data – a distributed ecosystem

DATA WILL INCREASINGLY REMAIN AT DIFFERENT LOCATIONS

The future FAIR ecosystem will necessarily be highly distributed. It will require technical mechanisms for linking resources as well as collaboration mechanisms for coordination and for agreement about specifications and standards. EOSC will have an important role to play in each of these mechanisms. For the FAIR ecosystem to work, there need to be registries cataloguing the component services and automated workflows between them. Federations offer a means to establish agreements between repositories or registries to carry out certain tasks collaboratively and therefore will be essential to this distributed system. Data will increasingly remain at different locations for reasons such as the expense of copying data or because of legal or ethical restrictions. Distributed queries, managed by brokering software, will be used to virtually integrate data. The need for such
1.7 Priority recommendations

1.7.1 Step 1: Define – concepts for FAIR Digital Objects and the ecosystem

- Rec. 1: Define FAIR for implementation
- Rec. 2: Implement a model for FAIR Digital Objects
- Rec. 3: Develop components of a FAIR ecosystem

In order to implement FAIR, research communities must define how the FAIR principles and related concepts apply in their context. This will differ based on the data types, the nature of research (e.g., ethical sensitivities or commercial partners) and the level of existing support for data sharing. The process of definition will help to identify points where the FAIR principles need to be supported with additional concepts and policies. To make...
FAIR data – A URGENT NEED

• DIFFUSED DATA SCIENCE SKILLS
• PROFESSIONAL DATA STEWARDS

1.4 Data science and stewardship skills

There is an urgent need to develop skills in relation to FAIR data. These skills fall broadly into two categories: data science and data stewardship. In the context of research, data science skills can be understood as the ability to handle, process and analyse data to draw insights from it. Data stewardship, meanwhile, is a set of skills to ensure data are properly managed, shared and preserved, both throughout the research lifecycle and for long-term preservation.

All researchers need a foundational-level set of data skills in order to make adequate use of available data and technologies. Such data skills should be recognised as intrinsic to research. That said, not all researchers should be expected to become experts in data science or data stewardship; some will become specialists of these domains but generally, research teams should be supported by - or should include - data professionals providing these skillsets.

New job profiles need to be defined and education programs put in place to train the large cohort of data scientists and data stewards required to support the transition to FAIR. Since the skillsets required for data science and data stewardship are varied and rapidly evolving, multiple formal and informal pathways to learning are required. This will help to scale up the cohort of data professionals required and enable a more diverse group of professionals to enter the field.
Smart Data Management Plans for FAIR Open Science

For serious researchers and data stewards

Data Stewardship Wizard

Data integration

Data interpretation

Information and insight

Will you be using any pre-existing data (including other people's data)?

Will you be referring to any earlier measured data, reference data, or data that should be mined from existing literature? Your own data as well as data from others?

Data Stewardship for Open Science: [link]

Do you need to harmonize different sources of existing data?

If you are combining data from different sources, harmonization may be required. You may need to re-analyse some original data.

Data Stewardship for Open Science: [link]

Will you be storing samples?

https://app.dsw.fairdata.solutions/questionnaire
Do all datasets you work with have a license?

It is not always clear to everyone in the project (and outside) what can and cannot be done with a dataset. It is helpful to associate each dataset with a license as early as possible in the project. A data license should ideally be as free as possible: any restriction like 'only for non-commercial use' or 'attribution required' may reduce the reusability and thereby the number of citations. If possible, use a computer-readable and computer actionable license.

- No
- Yes

Will you store the licenses with the data at all time?

It is very likely that data will be moved and copied. Some origins. It can be helpful to have the licenses of the data association with the data.

- Data Stewardship For Open Science: altw
- No
- Yes

How will you keep provenance?

To make your experiments reproducible, all steps in the data detail. The software you used, including version number information together for every step of the analysis is part of some questions regarding this in the chapter on data process.

- All steps will be documented in an (electronic) lab notebook
- Our workflow system documents the provenance automated

Data Stewardship Wizard

FAQR Data management

Will you store licenses with the data?

What's up?

Always consider the use of your data beyond the original purpose. One of the issues with reusing other people’s data is that they cannot be assumed to be reusable from an ethical or legal standpoint without explicit permission. Assuming that unlicensed data are 'free to use for whatever purpose' is intrinsically wrong, and in the case of for instance Pharmaceutical industry can lead to court cases later on. Therefore, whenever you publish a data set or any other kind of information or digital object, it is important to define a license for reuse. For software many licences exist, and for data, increasingly, standard licences are available or under development. Please note that a given license is also a defined concept and therefore deserves a Persistent Identifier and a URI pointing to where the license can be studied (machine readable licenses are also under development in some areas). This means that in the metadata, the license under which the data or the workflow can be reused is 'just another PID in the right place'. Users can then specify in their search or workflow container that 'only data with the following licenses should be included'.

For instance, if you include some data in your analysis that cannot be used for commercial purposes, that decision may render your entire results not usable for commercial purposes (at least in the view of some lawyers). This means that not licensing your data at all, even if you don't care who uses them and for what purpose is very counterproductive and will severely undermine the actual reuse of your data by others and in particular by industry. It will also lower the attribution-rate (usually part of the license conditions) and thus the citation and the impact score of your data.

Do

- Always carefully choose a license to be attached to your data upon publication.
- Include and clearly mark the license URL as a metadata attribute in the metadata.
- Store 'license' as part of the metadata in Open Access environments where search engines can easily find the license, even of the data they describe are not (yet) FAIR or even highly restricted in access. The fact that a data set with a specific license is 'out there' is a first step toward effective reuse of your data or information source.
- Make sure, especially when you restrict use of your data, that you are able to enforce the license you choose. Licenses that are not enforceable make no sense. (please note the enforcement is usually not done by an individual research group but at institutional or repository level)

Don’t

- Never publish data without a license attached or choose a license lightly, without considerations of anticipated reuse of your data.
- Choose a license that is not transitive (i.e. can not be transferred with subsets of the data), but make sure its transitivity does not unduly restrict the reuse of your data.
- Choose an unnecessary complicated license with many clauses and wherever possible one that is already widely adopted in the research community for either software

https://app.dsw.fairdata.solutions/questionnaire
... check your FAIRness

[Image -101x0 to 821x540]

[25x485] …
[65x485] check your FAIRness


FAIR maturity evaluator

The FAIR Maturity Evaluation Service

Public Entry Points:

- Browse existing Maturity Evaluations
- Browse MI Test Collections (Begin New Maturity Evaluation)
- Register a New MI Test Collection
- Browse Maturity Indicator Tests
- Register a new Maturity Indicator Test
- Search for MI Tests and Test Collections

https://linkeddata.systems:3000/

• OBJECTIVE
• MACHINE-ORIENTED ... AS FAIR IS!
Top 10 FAIR Data & Software Things
February 1, 2019

Sprinters:
Reid Otten, Stephanie Lathou, Ryan Johnson, Guillaume Canales, Mia Villas Boas, Anna

Geoscience

Sprinters:
John Brown, Janice Chan, Niamh Quigley (Curtin University)

Audience:
Researchers

Things

Findable
Thing 1: Data sharing and discovery
Thing 6: Vocabularies for data description
Thing 7: Identifiers and linked data
Thing 10: Spatial data

Accessible
Thing 2: Long-lived data: curation & preservation
Thing 3: Data citation for access & attribution


Our own data set: An example where a specific vocabulary (the VUG gridset) was used to markup a dataset can be found here. The dataset is part of a project to reconstruct the domestic market for colonial goods in the Dutch Republic.

ACTIVITIES:
1. Try to find one or two terms that are relevant to your research using the resources that are mentioned above. You can also use Swoogle to search for vocabularies related to your research.
2. Search for a term related to your research in the CIDOC Conceptual Reference Model (CRM) concept search. Were you able to find it? Tip 1: Search for “person” to get an idea of how the thesaurus works. Tip 2: All the terms used can be found in the last release of the model: http://www.cidoc-crm.org/get-last-official-release.

Thing 7: FAIR data modelling

The fourth and fifth star in Berner Lee’s model can be awarded when the data are stored in a format in which the topics their properties and their characteristics are identified using URIs whenever possible. More concretely, it implies that you record your data using the Resource Description Framework (RDF) format. RDF, simply put, is a technology which enables you to publish the contents of a database via the web. It is based on a simple data model which assumes that all statements about resources can be reduced to a basic form.
OPENING DATA
Why Open Data?

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

Reuse of data addresses the issue that sometimes researchers can use their data again. Reference here is not to (semi-)duplicate publication, where the same results are rehashed many times, but serendipity: researchers, thinking about the next research project or reading someone else’s paper, realize that their existing data might be useful in shining some (preliminary) light on a new question.

Recycle in this context is not much different from Reuse, but is used to designate making one’s data available to other researchers for use in secondary analysis, individual patient data meta-analysis, reanalysis to check on the findings reported by the data creator, or for any other valid scientific purpose. More and more funding agencies, professional

We strongly encourage that all datasets on which the conclusions of the paper rely should be available to readers. We encourage authors to ensure that their datasets are either deposited in publicly available repositories (where available and appropriate) or presented in the main manuscript or additional supporting files whenever possible. If a public repository does not exist, the information must be made available to editors and referees at submission and to readers promptly upon request. Any restrictions on material availability or other relevant information must be disclosed in the manuscript’s Methods section and should include details of how materials and information may be obtained.

In a 2016 editorial, I called attention to the “moral obligation” of Spinal Cord authors to share their data [4], and provided a tally for the year 2015: of 144 primary quantitative studies, just one made a statement (however inadequate) on data availability. All 143 others either claimed “There were no data to deposit” or did not address the issue at all. Recently, I analyzed all papers published in 2016–2018, and things have not improved: of 435 primary quantitative studies published by Spinal Cord, 95% claimed that there were no data to deposit, or made no statement.
Sharing data: good for science, good for you

https://www.youtube.com/watch?v=HJbo-OAaJ1I&feature=youtu.be
Sharing

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 #openscience"
Behaviours

Christopher and Alex (C&A) say: “This 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:

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 status.” 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).
2/4 "Open as possible, as closed as necessary" is the new principle for all #data from publicly funded #research in Europe #openaccess
A side effect of «open»

...YOU AVOID DUPLICATIONS...
ALL THE MORE SO IF YOU PUBLISH ALSO NEGATIVE DATA
Be open

Research Data Management

Prepare your data for publication

Decide what research data you want to make openly accessible for whom and whether you have the rights to do so. Think for example about the following topics:

- Resolve issues around intellectual property, ethics and privacy if you haven’t covered issues surrounding data ownership and data

Planning research

Research Data Management

Be open

Research Data Management

Research Data Management

Raise impact

Many of the tools available for measuring the impact of data are at an early stage of development. Here is what you can personally contribute to raising the impact of your data:

- Deposit data into a trustworthy repository. Trustworthy data repositories like 4TU.Centre for Research Data (with a Data Seal of Approval) provide a stable location for your data. Every dataset in the 4TU.Centre for Research Data archive is provided with a unique,

Research Data Management

Keep your data safe

TU Delft offers its employees several options for storing and exchanging research data safely. Solutions range from basic storage to tailored solutions like an integrated workflow for processing and sharing dynamic geodata in a tailored data lab.

The table below summarizes the current possibilities. The solution that fits your research data depends on your specific (security) needs.

Contact Faculty IT management if you want advice on the software solution and protective measures (like encryption). If you want to set up an environment specifically designed to share research data – like a Dataverse or data lab – contact 4TU.Centre for Research Data for assistance.

<table>
<thead>
<tr>
<th>Goal</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>I want to easily store data and share it selected others (in and outside of TU Delft).</td>
<td>SURFdrive allows you to store, synchronize and exchange 100 GB of research data safely and easily with third parties. You</td>
</tr>
<tr>
<td>I want to store and backup personal data.</td>
<td>Personal data storage offers a H-drive with standard 8 GB and back-up.</td>
</tr>
<tr>
<td>I want to store, share and backup data with faculty colleagues.</td>
<td>Group data storage is a data storage facility for multiple users in the same title, department or research group. The standard storage space for a group folder is 50 GB and back-up.</td>
</tr>
<tr>
<td>I want to store raw data.</td>
<td>With bulk data storage large quantities of temporary or measuring data can be stored default without backup contact your Faculty IT manager for the possibilities and costs of back-ups.</td>
</tr>
</tbody>
</table>
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

Bianca Kramer & Jeroen Bosman https://101innovations.wordpress.com

DOI: 10.5281/zenodo.1147025
Imagine a world where the preponderance of Earth, space, and environmental science data, software, and models are routinely shared in ways that allow easy discovery, recombination, reuse, and to test reliability, and where information about samples, methods, and tools are standardized, available, and linked across publications.

Scholars receive credit and recognition for producing data, developing new techniques and algorithms, and providing key samples. Tools, scripts, and common requirements enable scholars to prepare data, software, and samples efficiently for reuse starting from when they are collected or created.

Scholarly publishers are not the end point of scholarship but rather enable rich connections to these resources as well among researchers to accelerate new investigations within and across disciplines, expanding the research lifecycle.

Scientific repositories are valued for stewardship, data access, improving peer review and digital product quality and are supported and linked to ensure discovery of related data, software, services, and other digital research products.

Society and the public have increased confidence in scientific research, access to the digital research products that underpin research findings, and increased capability to discover and integrate diverse data sets rapidly to plan for sustainability or respond in real time to disasters. Singular observations of the environment or an event could be connected with confidence to understand dynamics and better assess change and impacts from local to global scales.

NEED A BREAK?