Open and FAIR Data

Lunch&Learn Open Science, April 6

Rachel Heyard, Center for Reproducible Science, University of Zurich

What is Open Research Data?

Definition from the Concordat on Open Research Data (by HEFECE, UKRI, Universities UK and the Wellcome Trust)

What is Open Research Data?

Definition from the **Concordat on Open Research Data** (by HEFECE, UKRI, Universities UK and the Wellcome Trust)

Research data **are the evidence** that underpins the answer to the research question, and can be used to validate findings regardless of its form (e.g. print, digital, or physical). These might be quantitative information or qualitative statements collected by researchers in the course of their work by experimentation, observation, modelling, interview or other methods, or information derived from existing evidence. Data may be **raw or primary** (e.g. direct from measurement or collection) or **derived from primary data** for subsequent analysis or interpretation (e.g. cleaned up or as an extract from a larger data set), or derived from existing sources where the rights may be held by others. [...]

What is Open Research Data?

Definition from the **Concordat on Open Research Data** (by HEFECE, UKRI, Universities UK and the Wellcome Trust)

Research data **are the evidence** that underpins the answer to the research question, and can be used to validate findings regardless of its form (e.g. print, digital, or physical). These might be quantitative information or qualitative statements collected by researchers in the course of their work by experimentation, observation, modelling, interview or other methods, or information derived from existing evidence. Data may be **raw or primary** (e.g. direct from measurement or collection) or **derived from primary data** for subsequent analysis or interpretation (e.g. cleaned up or as an extract from a larger data set), or derived from existing sources where the rights may be held by others. [...]

They may include, for example, statistics, collections of digital images, sound recordings, transcripts of interviews, survey data and fieldwork observations with appropriate annotations, an interpretation, an artwork, archives, found objects, published texts or a manuscript.

The primary purpose of research data is to provide the information necessary to support or validate a research project's observations, findings or outputs.

What is Open Research Data?

Definition from the **Concordat on Open Research Data** (by HEFECE, UKRI, Universities UK and the Wellcome Trust)

Research data **are the evidence** that underpins the answer to the research question, and can be used to validate findings regardless of its form (e.g. print, digital, or physical). These might be quantitative information or qualitative statements collected by researchers in the course of their work by experimentation, observation, modelling, interview or other methods, or information derived from existing evidence. Data may be **raw or primary** (e.g. direct from measurement or collection) or **derived from primary data** for subsequent analysis or interpretation (e.g. cleaned up or as an extract from a larger data set), or derived from existing sources where the rights may be held by others. [...]

They may include, for example, statistics, collections of digital images, sound recordings, transcripts of interviews, survey data and fieldwork observations with appropriate annotations, an interpretation, an artwork, archives, found objects, published texts or a manuscript.

The primary purpose of research data is to provide the information necessary to support or validate a research project's observations, findings or outputs.

Open research data are those research data that can be freely accessed, used, modified, and shared, provided that there is appropriate acknowledgement if required.

Required by funders

Required by funders

SNSF policy on Open Research Data

The SNSF values research data sharing as a fundamental contribution to the impact, transparency and reproducibility of scientific research. In addition to being carefully curated and stored, the SNSF believes research data should be shared as openly as possible.

The SNSF therefore expects all its funded researchers

- *to store the research data they have worked on and produced during the course of their research work,*
- *to share these data with other researchers, unless they are bound by legal, ethical, copyright, confidentiality or other clauses, and*
- *to deposit their data and metadata onto existing public repositories in formats that anyone can find, access and reuse without restriction.*

Research data is collected, observed or generated factual material that is commonly accepted in the scientific community as necessary to document and validate research findings.

Required by universities

Required by universities

UZH Open Science Policy

UZH expects that all publicly funded scholarly output – including, e.g. publications, research data and code – is made openly available.

UZH expects output of all publicly funded research to be made FAIR (Findable, Accessible, Interoperable and Reusable). The FAIR principles apply to data and metadata as well as to software, code, algorithms, and workflows/protocols that lead to that data.

The FAIR principles

www.nature.com/scientificdata

SCIENTIFIC DATA

Amended: Addendum

OPEN

SUBJECT CATEGORIES

» Research data

» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*[#]

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

The FAIR principles

Findable

Accessible

Interoperable

Reusable

The FAIR principles

Findable

- ➔ (meta)data should have a globally unique and persistent identifier (doi)
- ➔ well described metadata
- ➔ (meta)data should be registered (zenodo, osf, …)

Accessible

Interoperable

Reusable

The FAIR principles

Findable

- ➔ (meta)data should have a globally unique and persistent identifier (doi)
- ➔ well described metadata
- ➔ (meta)data should be registered (zenodo, osf, …)

Accessible

- ➔ (meta)data and protocols should be retrievable, open and free
- ➔ metadata should stay accessible, even when data not available

Interoperable

Reusable

The FAIR principles

Findable

- ➔ (meta)data should have a globally unique and persistent identifier (doi)
- ➔ well described metadata
- ➔ (meta)data should be registered (zenodo, osf, …)

Accessible

- ➔ (meta)data and protocols should be retrievable, open and free
- ➔ metadata should stay accessible, even when data not available

Interoperable

- ➔ (meta)data should use a formal, accessible, shared and broadly applicable language

Reusable

The FAIR principles

Findable

- ➔ (meta)data should have a globally unique and persistent identifier (doi)
- ➔ well described metadata
- ➔ (meta)data should be registered (zenodo, osf, …)

Accessible

- ➔ (meta)data and protocols should be retrievable, open and free
- ➔ metadata should stay accessible, even when data not available

Interoperable

- ➔ (meta)data should use a formal, accessible, shared and broadly applicable language

Reusable

- ➔ (meta)data with clear and accessible usage license and detailed provenance

Why is data sharing so important for science?

Why is data sharing so important for science?

- Facilitates sharing of knowledge and resources

Why is data sharing so important for science?

- Facilitates sharing of knowledge and resources
- Makes science more efficient
- Makes science more transparent

Why is data sharing so important for science?

- Facilitates sharing of knowledge and resources
- Makes science more efficient
- Makes science more transparent
- Increases trust in science
- Increases fairness towards tax-payers

Why is data sharing so important for science?

- Facilitates sharing of knowledge and resources
- Makes science more efficient
- Makes science more transparent
- Increases trust in science
- Increases fairness towards tax-payers
- Facilitates self regulation of science

Why is data sharing so important for the individual researcher?

Why is data sharing so important for the individual researcher?

- Makes you / the researcher more attractive for collaboration

Why is data sharing so important for the individual researcher?

- Makes you / the researcher more attractive for collaboration
- Increases credibility
- Increases citations / outreach
- Increases visibility

Why is data sharing so important for the individual researcher?

- Makes you / the researcher more attractive for collaboration
- Increases credibility
- Increases citations / outreach
- Increases visibility
- Makes review process more efficient

REPRODUCIBILITY
NOTES

When should data and code be made available?

Sharing data and code as part of a research publication is crucial for ensuring the computational reproducibility of scientific work. But sharing should be done at the article submission stage, not after publication as it is now, say **Rachel Heyard** and **Leonhard Held**. Statisticians and data scientists have the skills and tools to make this change and lead by example, though there are obstacles to overcome

Open and FAIR data sharing does not come without obstacles

1. Data privacy issues

Open and FAIR data sharing does not come without obstacles

1. Data privacy issues
   - Pseudo-anonymisation
   - Statistical methods to ensure anonymisation

Open and FAIR data sharing does not come without obstacles

1. Data privacy issues
   - Pseudo-anonymisation
   - Statistical methods to ensure anonymisation
2. Recognition, being acknowledged for data sharing

Open and FAIR data sharing does not come without obstacles

1. Data privacy issues
   - Pseudo-anonymisation
   - Statistical methods to ensure anonymisation
2. Recognition, being acknowledged for data sharing
   - Data Journals
   - Data with DOI can be linked to ORCiD

Open and FAIR data sharing does not come without obstacles

1. Data privacy issues
   - Pseudo-anonymisation
   - Statistical methods to ensure anonymisation
2. Recognition, being acknowledged for data sharing
   - Data Journals
   - Data with DOI can be linked to ORCiD
3. Long-term data maintenance

Open and FAIR data sharing does not come without obstacles

1. Data privacy issues
   - Pseudo-anonymisation
   - Statistical methods to ensure anonymisation
2. Recognition, being acknowledged for data sharing
   - Data Journals
   - Data with DOI can be linked to ORCiD
3. Long-term data maintenance
   - Training and data stewardships

Open and FAIR data sharing does not come without obstacles

1. Data privacy issues
   - Pseudo-anonymisation
   - Statistical methods to ensure anonymisation
2. Recognition, being acknowledged for data sharing
   - Data Journals
   - Data with DOI can be linked to ORCiD
3. Long-term data maintenance
   - Training and data stewardships
4. Open and FAIR data sharing *is not enough*

Open and FAIR data sharing does not come without obstacles

1. Data privacy issues
   - Pseudo-anonymisation
   - Statistical methods to ensure anonymisation
2. Recognition, being acknowledged for data sharing
   - Data Journals
   - Data with DOI can be linked to ORCiD
3. Long-term data maintenance
   - Training and data stewardships
4. Open and FAIR data sharing *is not enough*
   - Additionally share code, software, and all research material

How does it work in practice?

…. How did I try to do it myself? (my interpretation of Open and FAIR data)

Abstract of one of my recent preprints

Abstract

Funding agencies rely on peer review and expert panels to select the research deserving funding. Peer review has limitations, including bias against risky proposals or interdisciplinary research. The inter-rater reliability between reviewers and panels is low, particularly for proposals near the funding line. Funding agencies are also increasingly acknowledging the role of chance. The Swiss National Science Foundation (SNSF) introduced a lottery for proposals in the middle group of good but not excellent proposals. In this article, we introduce a Bayesian hierarchical model for the evaluation process. To rank the proposals, we estimate their expected ranks (ER), which incorporates both the magnitude and uncertainty of the estimated differences between proposals. A provisional funding line is defined based on ER and budget. The ER and its credible interval are used to identify proposals with similar quality and credible intervals that overlap with the funding line. These proposals are entered into a lottery. We illustrate the approach for two SNSF grant schemes in career and project funding. We argue that the method could reduce bias in the evaluation process. R code, data and other materials for this article are available online.

First Problem

Data privacy issues

→ data on grades given to research proposals by panel members!!!

First Problem

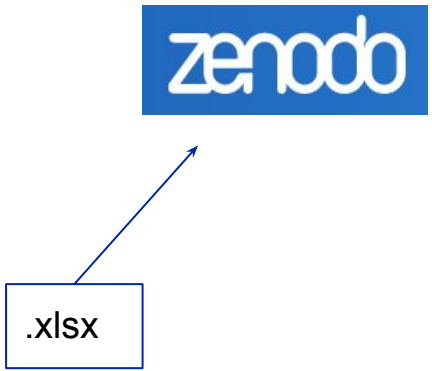
Data privacy issues

→ data on grades given to research proposals by panel members!!!

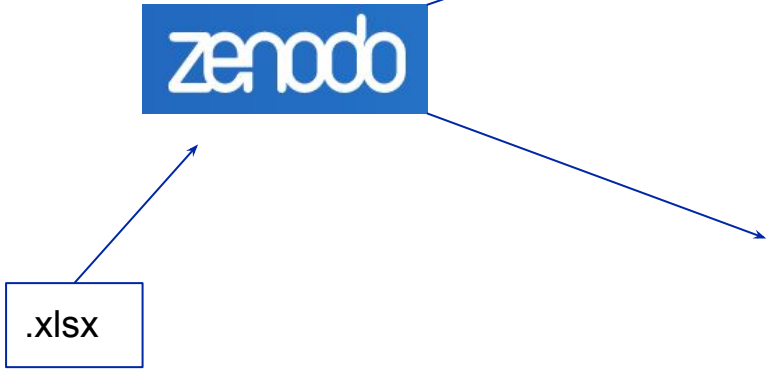
→ Consulted the SNSF legal department to ensure that data is (pseudo)anonymous.

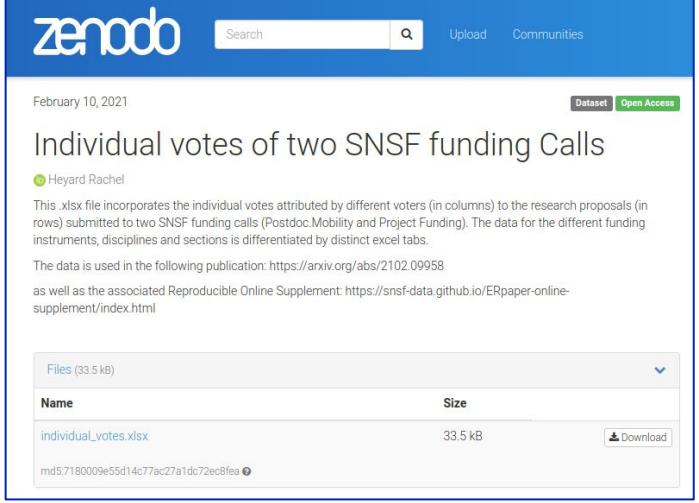
Where? → public data registry

Where? → public data registry



Where? → public data registry





Findable, Interoperable, …

Supplemental Materials and Data

An online fully reproducible supplement is provided which uses an `R` (`ERforResearch`) package with the implementation of the above presented methodology (see snsf-data.github.io/ERpaper-online-supplement/). The data used in the case studies can be downloaded from Zenodo: https://doi.org/10.5281/zenodo.4531160.

Findable, Interoperable, …

Supplemental Materials and Data

An online fully reproducible supplement is provided which uses an `R` (`ERforResearch`) package with the implementation of the above presented methodology (see snsf-data.github.io/ERpaper-online-supplement/). The data used in the case studies can be downloaded from Zenodo: https://doi.org/10.5281/zenodo.4531160.

+ The data set is linked to my ORCiD-Profile
+ Online supplement uses / loads data-set directly from Zenodo to reproduce results from paper using functionalities implemented in an R-package (shared on github)

CRS-UZH ReproducibiliTea

April 07, 2022
16:00-16:45h

With
Jessie Baldwin, Postdoc @UCL

Protecting against researcher bias in secondary data analysis: challenges and potential solutions

https://www.crs.uzh.ch/en/training/ReproducibiliTea.html

Thank you.

Question?

Comments?