FAIR data for reproducible research

Lunch&Learn Open Science, September 23

Rachel Heyard, Center for Reproducible Science, University of Zurich

Goals of the Center for Reproducible Science (CRS)

1. Teaching and training to improve the overall reproducibility and quality of empirical research

- **Good Research Practice** (GRP) courses
- **Primers** in Good Research Practice
- **ReproducibiliTea** journal club
- **Reproducibility Lab Pitches**

2. Promote, support and conduct original research in reproducibility and methodology

- Design and analysis of **replication studies**
- **Meta-research**

more infos, see
www.crs.uzh.ch

What is Research Data?

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

What is 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 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 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.

What is 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.

+ code, software, protocols and methods

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

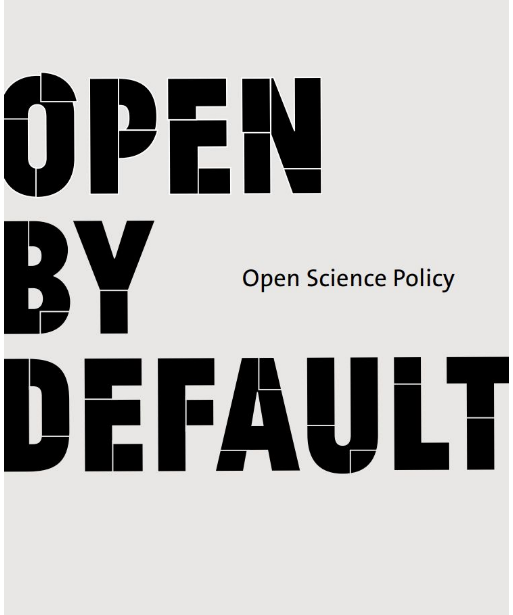
Open research data are those research data that can be freely accessed, used, modified, and shared, provided that there is appropriate acknowledgement if required.

University requirement

UZH Open Science Policy

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

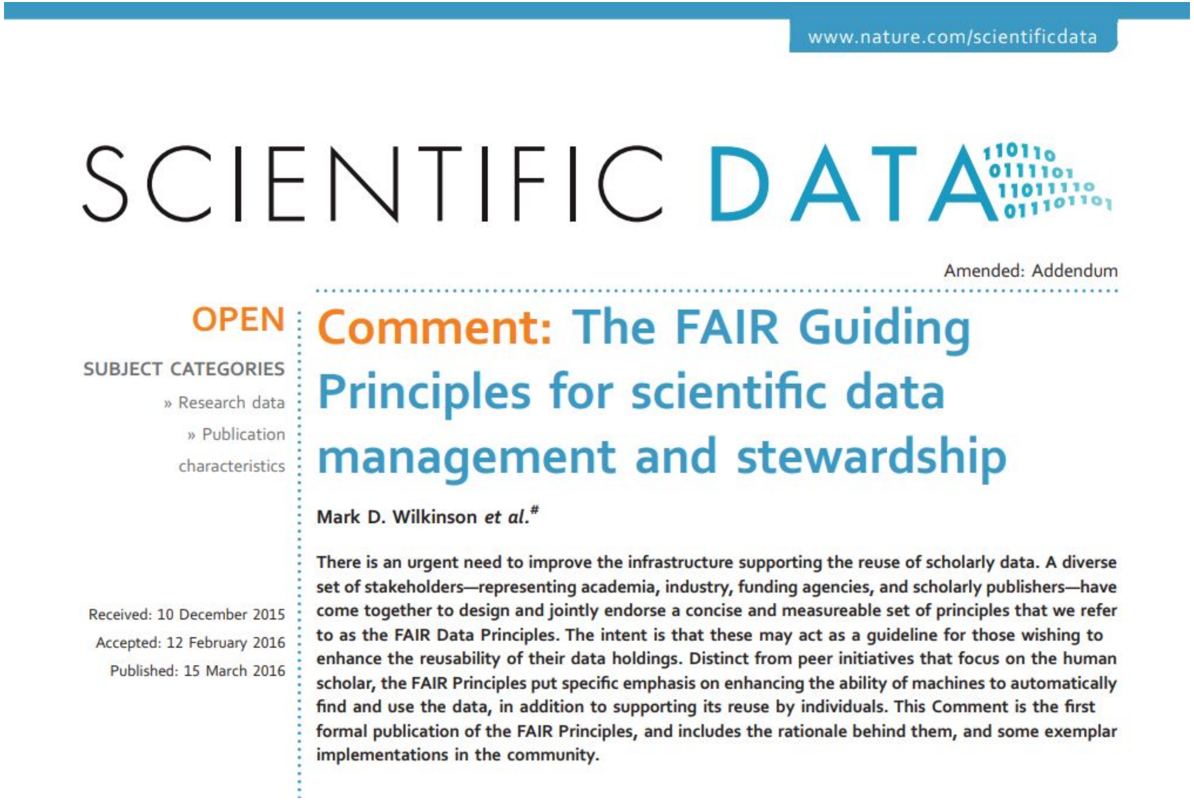
UZH expects output of all publicly funded research to be made FAIR (Findable, Accessible, Interoperable and Reusable). The FAIR principles apply to data and metadata as well as to software, code, algorithms, and workflows/protocols that lead to that data.



OPEN BY DEFAULT

Open Science Policy

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.*#

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.

Received: 10 December 2015

Accepted: 12 February 2016

Published: 15 March 2016

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, Open Science Framework, …)

Can I find it, also in a year from now?

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, Open Science Framework, …)

Can I find it, also in a year from now?

Accessible

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

Can my colleague access it, without paywall, or other obstacles?

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, Open Science Framework, …)

Can I find it, also in a year from now?

Accessible

➔ (meta)data and protocols should be retrievable, open and free

➔ metadata should stay accessible, even when data not available

Can my colleague access it, without paywall, or other obstacles?

Interoperable

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

Can my colleague interact with the data, do they understand and find all relevant information?

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, Open Science Framework, …)

Can I find it, also in a year from now?

Accessible

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

Can my colleague access it, without paywall, or other obstacles?

Interoperable

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

Can my colleague interact with the data, do they understand and find all relevant information?

Reusable

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

Will a potential future collaborator be able to re-use my data, without contacting me?

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
   - FAIR data does not need to be open

Open and FAIR data sharing does not come without obstacles

1. Data privacy issues
   - Pseudo-anonymisation
   - Statistical methods to ensure anonymisation
   - FAIR data does not need to be open
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
   - FAIR data does not need to be open
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
   - FAIR data does not need to be open
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
   - FAIR data does not need to be open
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
   - FAIR data does not need to be open
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
   - FAIR data does not need to be open
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

What is Metadata?

Metadata is **data about data**, that

is machine readable.

makes data FAIR - *findable, accessible, interoperable, and reusable*.

facilitates data reuse and discovery.

contains a data-dictionary or codebook defining and explaining variable in the data.

Start planning your data management

Start planning your data management

Data management plans (DMPs) are more and more required (funders, publishers, …).

UZH Library provides lots of support.

Start planning your data management

Data management plans (DMPs) are more and more required (funders, publishers, …).

UZH Library provides lots of support.

From researcher to researcher - this will save you lots of trouble along the way!

Data management planning

What goes in a DMP?:

Data management planning

What goes in a DMP?:

- What type of data will be reused / generated?
- What is the purpose of data reuse / generation?
- What is the data provenance and origin?
- What is the expected size?

Data management planning

What goes in a DMP?:

- What type of data will be reused / generated?
- What is the purpose of data reuse / generation?
- What is the data provenance and origin?
- What is the expected size?
- To whom might it be useful?
- What type of metadata will be shared with the data? Any metadata standards to follow? Other documentation?

Data management planning

What goes in a DMP?:

- What type of data will be reused / generated?
- What is the purpose of data reuse / generation?
- What is the data provenance and origin?
- What is the expected size?
- To whom might it be useful?
- What type of metadata will be shared with the data? Any metadata standards to follow? Other documentation?
- Strategy for sharing? Where? What?
- Strategy for quality control?
- Ethical considerations and data privacy?

Data management planning - My suggestion

Start with defining data, and list the types of data you generate and reuse.

Data management planning - My suggestion

Start with defining data, and list the types of data you generate and reuse.

I *generate*

- R scripts
- research protocols
- papers and preprints
- teaching material

I *reuse*

- R code and packages
- openly accessible data

Data management planning - My suggestion

Quality control?

I *generate*

- R scripts
- research protocols
- papers and preprints
- teaching material

I *reuse*

- R code and packages
- openly accessible data

Data management planning - My suggestion

Quality control?

I *generate*

- R scripts: code review and version control
- research protocols: follow reporting guidelines and templates
- papers and preprints: reporting guidelines, use community approved vocabulary, engage with pre-/post-publication peer review
- teaching material: option for feedback / test material

I *reuse*

- R code and packages
- openly accessible data

Data management planning - My suggestion

Strategy for sharing?

I *generate*

- R scripts: via git, with snapshot on Zenodo for DOI
- research protocols: preregistration on the Open Science Framework (with DOI)
- papers and preprints: post preprint before submission, prefer OA journals
- teaching material: upload to the OSF (with DOI)

I *reuse*

- R code and packages: *properly cite all packages and sources*
- openly accessible data: *ensure data has a license*

Data management planning - My suggestion

Metadata? Documentation?

I *generate*

- R scripts: comments in code, add examples for usage, add README and license
- research protocols: is documentation + minimal metadata via OSF + license
- papers and preprints: …
- teaching material: description of learning objectives, add license

I *reuse*

- R code and packages
- openly accessible data

Thank you.

Question?

Comments?