FAIR and Open Data

Silvio Peroni
silvio.peroni@unibo.it – https://orcid.org/0000-0003-0530-4305 – @essepuntato@scholar.social

Open Science (A.Y. 2024/2025)
Second Cycle Degree in Digital Humanities and Digital Knowledge
Alma Mater Studiorum - Università di Bologna

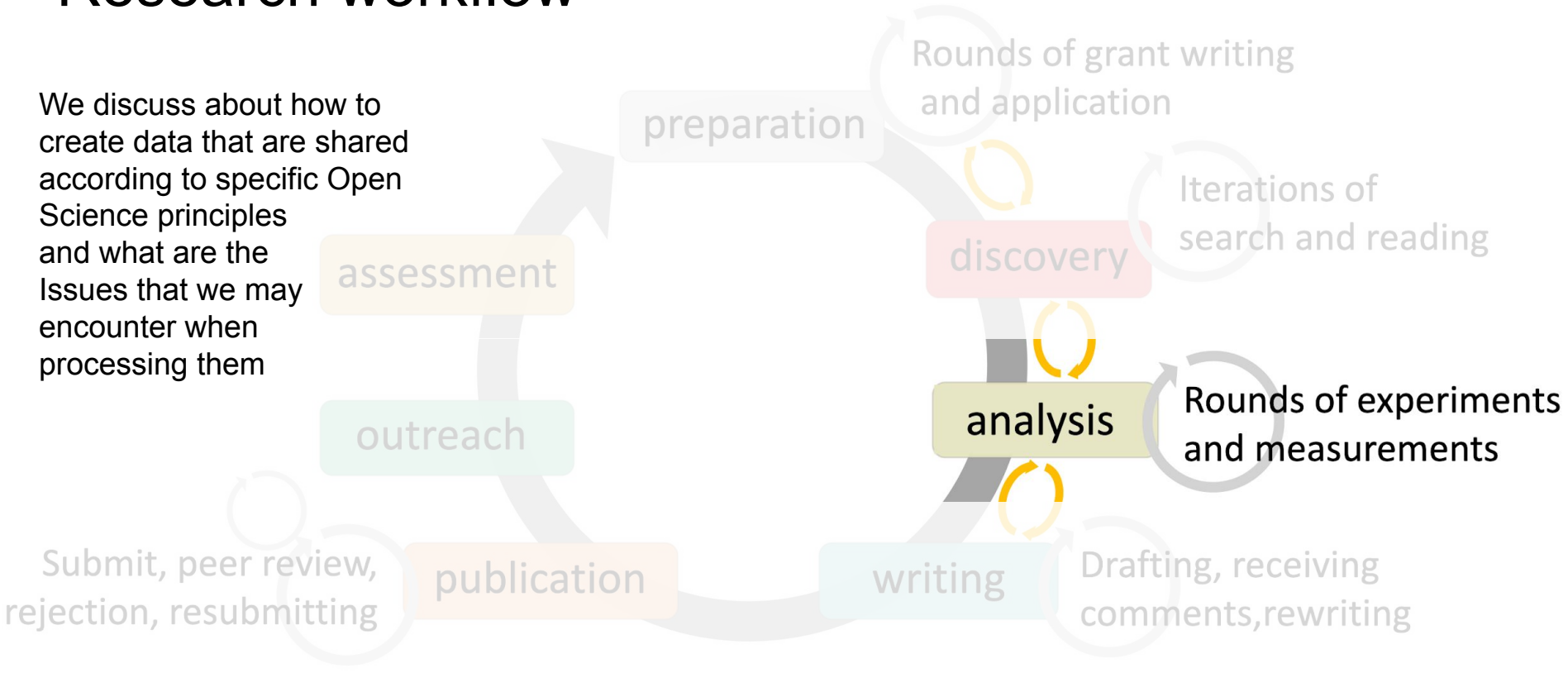
DIPARTIMENTO DI FILOLOGIA CLASSICA E ITALIANISTICA

Discussion

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE Principles for Indigenous Data Governance. Data Science Journal, 19, 43. https://doi.org/10.5334/dsj-2020-043

Research workflow

We discuss about how to create data that are shared according to specific Open Science principles and what are the Issues that we may encounter when processing them



Kramer, B., & Bosman, J. (2015, June 18). The good, the efficient and the open—Changing research workflows and the need to move from Open Access to Open Science. CERN Workshop on Innovations in Scholarly Communication (OAI9), University of Geneva, Geneva, Switzerland. https://www.slideshare.net/bmkramer/the-good-the-efficient-and-the-open-oai9

FAIR data principles

Findability, Accessibility, Interoperability, and Reusability: these four principles, proposed by the FORCE11 community, serve to guide data producers and publishers for helping them to maximize the added-value gained by contemporary, formal scholarly digital publishing

They represent goals and desiderata of good data management and stewardship

Even if they have been devised originally for data, they have been proposed considering all scholarly digital research objects in mind, since all components of the research process must be available to ensure transparency, reproducibility, and reusability

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

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 knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with 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 detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

Different types of data

The agnosticism of generic principles like FAIR requires the various communities' specificities (i.e. proper to Informatics, Humanities, Medicine, Statistics, etc.) to be taken into account

General rule: within the FAIR framework, everything is data

Data should be "intended to be as universal as possible, including datasets, publications, software, etc."

Avanço, K., Balula, A., Błaszczyńska, M., Buchner, A., … Wieneke, L. (2021). Future of Scholarly Communication—Forging an inclusive and innovative research infrastructure for scholarly communication in Social Sciences and Humanities (p. 46). Digital Humanities Centre at the Institute of Literary Research of the Polish Academy of Sciences. https://doi.org/10.5281/zenodo.5017705

Some examples: what is data in Informatics?

A number of research communities and groups have been considering how to apply aspects of FAIR to research software since 2017

The adoption of FAIR principles enables the transparency, reproducibility, and reusability of research – and part of this story applies also to software that needs to be well-described (i.e. metadata), inspectable, documented and appropriately structured so that it can be executed, replicated, built-upon, combined, reinterpreted, reimplemented, and/or used in different settings

F: Software, and its associated metadata, is easy for both humans and machines to find.
F1. Software is assigned a globally unique and persistent identifier. • F1.1. Components of the software representing levels of granularity are assigned distinct identifiers. • F1.2. Different versions of the software are assigned distinct identifiers. F2. Software is described with rich metadata. F3. Metadata clearly and explicitly include the identifier of the software they describe. F4. Metadata are FAIR, searchable and indexable.
A: Software, and its metadata, is retrievable via standardized protocols.
A1. Software is retrievable by its 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 software is no longer available.
I: Software interoperates with other software by exchanging data and/or metadata, and/or through interaction via application programming interfaces (APIs), described through standards.
I1. Software reads, writes and exchanges data in a way that meets domain-relevant community standards. I2. Software includes qualified references to other objects.
R: Software is both usable (can be executed) and reusable (can be understood, modified, built upon, or incorporated into other software).
R1. Software is described with a plurality of accurate and relevant attributes. • R1.1. Software is given a clear and accessible license. • R1.2. Software is associated with detailed provenance. R2. Software includes qualified references to other software. R3. Software meets domain-relevant community standards.

Chue Hong, N. P., Katz, D. S., Barker, M., Lamprecht, A.-L., Martinez, C., Psomopoulos, F. E., Harrow, J., Castro, L. J., Gruenpeter, M., Martinez, P. A., & Honeyman, T. (2022). FAIR Principles for Research Software (FAIR4RS Principles) [Recommendations with RDA Endorsement in Process]. Research Data Alliance. https://doi.org/10.15497/RDA00068

Some examples: what is data in the Humanities?

Humanities (at least, those related to studies in philology, literary criticism, language, linguistics, history of art and archival and library studies) uses and/or produces several kinds of data

Probably, one of the most surprisingly data types is events, meaning “any one-off gathering of people organised as a result of a research project, to share ideas, offer training, or present something to the public”, including:

- Conferences
- Exhibitions
- Webinars
- Guided tours
- Teacher training

Data types	Produced by (tot. 19)	Used by (tot. 19)
publications	18	15
other primary sources (e.g., manuscripts, artworks)	0	18
digital representation of cultural objects or events (e.g., facsimiles, photo, video)	4	8
catalogues, databases and other search tools	2	9
events (e.g., conferences, exhibitions)	6	0
static websites	4	0
software	2	2
documentation	3	0
digital infrastructures (e.g., mobile apps, web platforms)	3	0
personal data	2	0
corpora	2	0
standards	0	2
born-digital artifacts (e.g., tags, associations, texts)	1	1

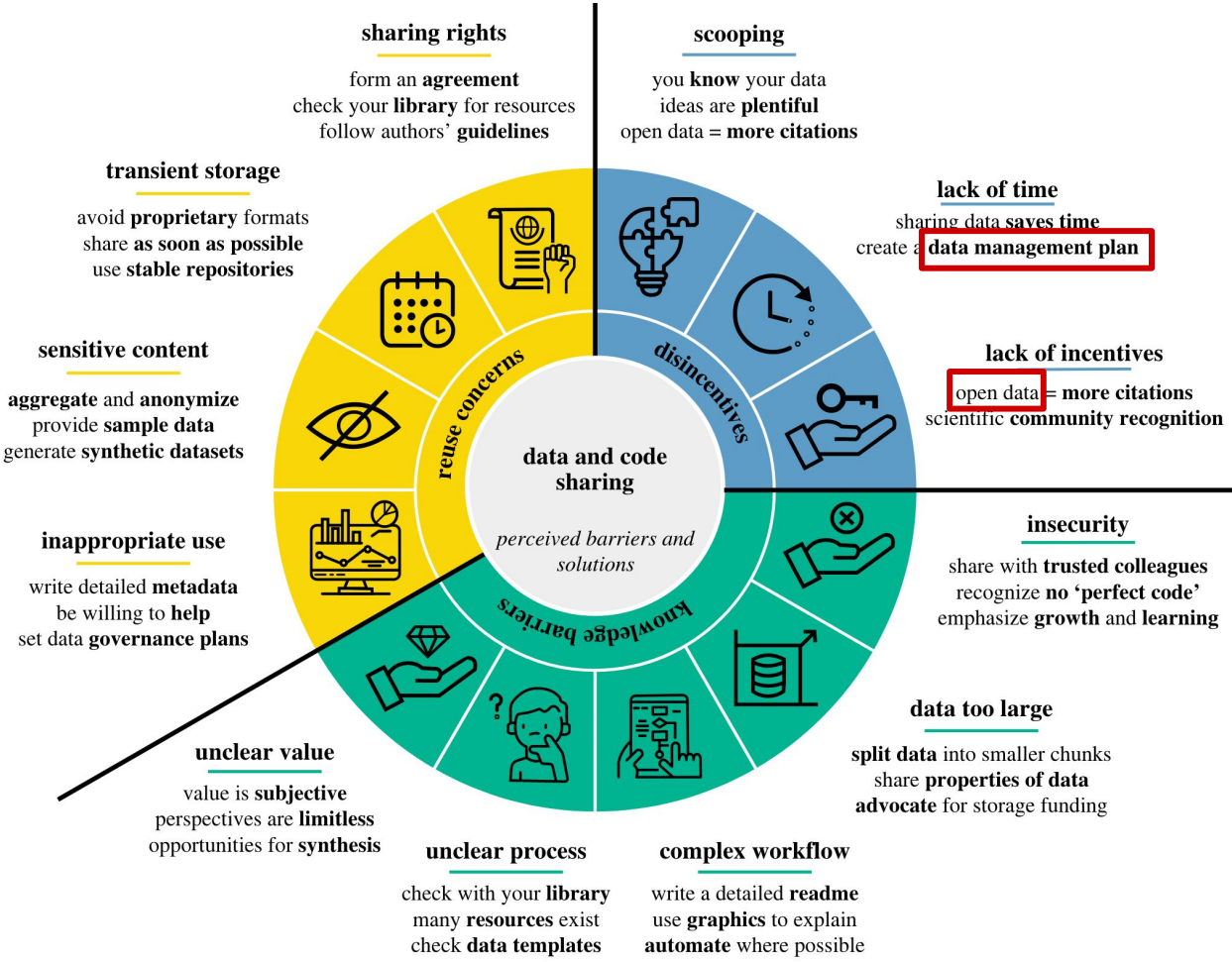
Gualandi, B., Pareschi, L., & Peroni, S. (2022). What do we mean by ‘data’? A proposed classification of data types in the arts and humanities. Journal of Documentation, 79(7), 51–71. https://doi.org/10.1108/JD-07-2022-0146

Sharing: some guidelines

Many researchers do not share their data simply because they do not know how

Data should be published alongside detailed metadata

Data sharing will increase opportunities for collaboration



Gomes, D. G. E., Pottier, P., Crystal-Ornelas, R., Hudgins, E. J., Foroughirad, V., …, & Gaynor, K. M. (2022). Why don't we share data and code? Perceived barriers and benefits to public archiving practices. Proceedings of the Royal Society B: Biological Sciences, 289(1987), 20221113. https://doi.org/10.1098/rspb.2022.1113

What about Open?

Warning: FAIR data does not imply Open Data

Open as defined in the Open Definition:

> "Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness)"

License	Attribution (i.e. preserving provenance)	Forcing openness
Creative Commons Attribution (CC-BY)	yes	no
Creative Commons Attribution - Share-Alike (CC-BY-SA)	yes	yes
Creative Commons CCZero (CC0)	no	no
Open Data Commons Public Domain Dedication and Licence (PDDL)	no	no
Open Data Commons Attribution License (ODC-BY)	yes	no
Open Data Commons Open Database License (ODbL)	yes	yes

Open Knowledge Foundation. (2015). Open Definition 2.1. https://opendefinition.org/od/2.1/en/

Issues to the path towards the Open: personal data

Personal data is any information that enable the identification of a natural person (i.e. the data subject), in particular by reference to an identifier such as a name, an identification number, etc.

Personal data can be gathered under specific conditions and cannot be published as Open Data – derogation to the this rule may exist, e.g. personal data in bibliographic information (authors' names, ORCIDs, etc.)

The principles of data protection do not apply if

- Anonymisation – rendering personal data anonymous in such a manner that the data subject is not or no longer identifiable
- Pseudonymisation – rendering personal data in such a manner that they can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), no. L119, Official Journal of the European Union (2016). http://data.europa.eu/eli/reg/2016/679/oj

How to publish data

Generally speaking, data should be published

as open as possible, as closed as necessary

In choosing the license, in particular if considering or not the attribution clause, one has to think about the final goal to reach – that, within the Open Science ecosystem, should be fostering maximum reuse

Recently, the European Union has provided a clear guideline to follow for what it concerns the publication of data:

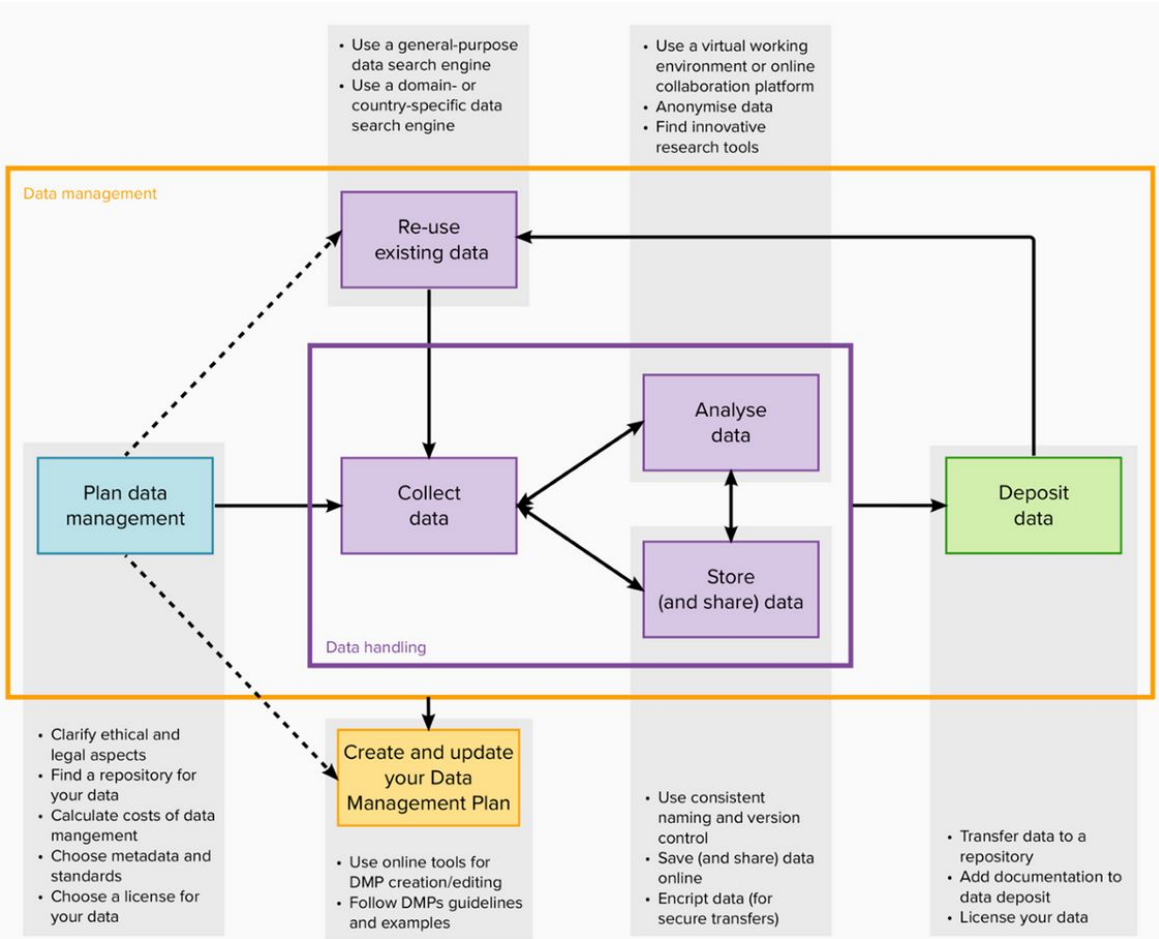
> "raw data, metadata or other documents of comparable nature may alternatively be distributed under the provisions of the Creative Commons Universal Public Domain Dedication deed (CC0 1.0)"

Landi, A., Thompson, M., Giannuzzi, V., Bonifazi, F., Labastida, I., da Silva Santos, L. O. B., & Roos, M. (2020). The "A" of FAIR – As Open as Possible, as Closed as Necessary. Data Intelligence, 2(1–2), 47–55. https://doi.org/10.1162/dint_a_00027

FAIR and RDM

FAIR principles should be applied during the entire data life cycle

- When creating the DMP
- When you analyse data
- When you store data
- When you deposit data
- When you re-use existing data



Caldoni, G., Gualandi, B., & Marino, M. (2022). Research Data Management: Decision Tree [Decision tree]. Alma Mater Studiorum - Università di Bologna. https://doi.org/10.5281/zenodo.7190005
Gualandi, B., Caldoni, G., & Marino, M. (2022). Research Data Management: Data Lifecycle [Diagram]. Alma Mater Studiorum - Università di Bologna. https://doi.org/10.5281/zenodo.7249051

Mandatory text to read for the next lecture

Chigbu, U. E., Atiku, S. O., & Du Plessis, C. C. (2023). The Science of Literature Reviews: Searching, Identifying, Selecting, and Synthesising. Publications, 11(1), 2. https://doi.org/10.3390/publications11010002

End

FAIR and Open Data

Silvio Peroni
silvio.peroni@unibo.it – https://orcid.org/0000-0003-0530-4305 – @essepuntato@scholar.social

Open Science (A.Y. 2024/2025)
Second Cycle Degree in Digital Humanities and Digital Knowledge
Alma Mater Studiorum - Università di Bologna

DIPARTIMENTO DI FILOLOGIA CLASSICA E ITALIANISTICA