Supporting FAIR data: categorization of research data as a tool in data management

Jessica Parland-von Essen
University of Helsinki
parland@csc.fi
https://orcid.org/0000-0003-4460-3906

Katja Fält
Tampere University of Technology
katja.falt@tut.fi
https://orcid.org/0000-0002-6172-5377

Zubair Maalick
CSC – IT Center for Science
zubair.maalick@csc.fi
https://orcid.org/0000-0002-0975-1471

Miika Alonen
Aalto University
miika.alonen@csc.fi
https://orcid.org/0000-0002-0065-0017

Eduardo Gonzalez
CSC – IT Center for Science
eduardo.gonzalez@csc.fi
https://orcid.org/0000-0003-1400-0995

There are two main aspects of research data management (RDM) that are discussed in this presentation: the researchers' need for reliable data citation and the

Persistent identifier: https://doi.org/10.23978/inf.76084

need for data management at large for research organizations and infrastructures. The existing guidelines such as the The Research Data Alliance (RDA) for citing dynamic data and the FAIR data principles for scientific data management aim at pushing the research culture towards data that is findable, accessible, interoperable and reusable and research that is reproducible. Researcher are requested to create FAIR data and use services that support these guidelines. The demand for implementation of the FAIR data principles gives us a great challenge to fix data citation (FORCE11, 2014; Laine & Nykyri, 2018; UNIFI, 2018; Wilkinson et al., 2016). The RDA guidance is today in many cases in practice impossible for a researcher to adhere to in an efficient way due to lacking infrastructure and services. By analyzing data categorization, it is possible to organize RDM services in an adequate way on both the local level and at a larger scale. Planning research data management services today needs to take into account new requirements in a systematic way. In practice this means a mindful usage of persistent identifiers and a data management that takes into account different types of data as well as the diverse needs of the users.

To make the aims meet we need analysis of the properties of the data and as an outcome a more extensive data categorization. This can be done in three concurrent ways. We argue that it is important to make a distinction between technical, contextual and inherent traits of the data. Especially the inherent properties, that are not purely of a technical character, are the most important and currently often under-appreciated in RDM and metadata schemas etc, even though they are important for both citation and planning infrastructures. We focus particularly on the aspect of stability, because this is the most important for making data FAIR and enabling solutions for trustworthy data citation.

We suggest that contextual traits of data are more clearly separated from its inherent qualities. We propose a tripartite research data categorization as part of describing these inherent properties, which would make lifecycle and architecture planning and managing heterogeneous data resources within organizations more easy. The categories are operational data, generic research data and research data publications. Generic research data is validated data and can be cumulative, i.e. data can be added explicit without versioning or by versioning a database. (CEOS Data Stewardship Interest Group, 2017; Matthiesen & Dieckmann, 2018) It should be separated from immutable dataset publications that are published for reasons of reproducibility of specific research results. Managing and citing cumulative datasets could be handled differently from discrete research data publications. By addressing conflicts between the data deluge in dynamic research data and the traditional static, archival way of looking at data (where a new version always constitutes a complete new copy of a dataset etc), we hope to achieve more appropriate ways to handle the need for trustworthy

but efficient data management and the researchers' need of citation and scientific reproducibility.

Metadata formats and research data architecture should be developed further to better answer the diverse needs in data management and research. Data categorization is an important tool that is currently underutilized in creating infrastructures and services that enable wider deployment of FAIR data compliant practices.

References

CEOS Data Stewardship Interest Group. (2017). Persistent identifier best practices. Version 1.2. CEOS/wgiss/dsig/pidbp. http://ceos.org/document_management/Working_Groups/WGISS/Documents/WGISS/%20Best/%20Practices/CEOS/%20Persistent/%20Identifier/%20Best/%20Practices_v1.2.pdf

FORCE11. (2014, September). The fair data principles. *FORCE11*. https://www.force11.org/group/fairgroup/fairprinciples

Laine, H., & Nykyri, S. (2018). Dataviittaamisen tiekartta tutkijalle. *Informaatiotutkimus*, *37*(2). https://doi.org/10.23978/inf.72999

Matthiesen, M., & Dieckmann, U. (2018). Versioning with persistent identifiers. *CLARIN Annual Conference 2018 in Pisa, Italy.* https://www.clarin.eu/clarin-annual-conference-2018-abstracts

UNIFI. (2018). *Avoin tiede ja data. Toimenpideohjelma suomalaiselle tiedeyhteisölle*. Suomen yliopistojen rehtorineuvosto UNIFI ry. http://urn.fi/URN:NBN:fi-fe2018052424593

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016, March). The fair guiding principles for scientific data management and stewardship. *Scientific Data*. Comments and Opinion. https://doi.org/10.1038/sdata.2016.18