WDS/RDA Publishing Data Interest Group
WDS/RDA Certification of Digital Repositories Interest Group

Assessment of Data Fitness for Use

WG Charter

The increasing availability of research data and its evolving role as a first class scientific output in the scholarly communication requires a better understanding of and the possibility to assess data quality, which in turn can be described as conformance of data properties to data usability or fitness for use. These properties are multifaceted and cover various aspects related to data objects, access services, and data management processes such as the level of annotation, curation, peer review, and citability or machine readability of datasets. Moreover, the compliance of a data repository or data center providing datasets - for example with certification requirements - could serve as a useful proxy.

Currently, there is a fairly good understanding on how to certify the quality of a data center / repository as a whole, but there is no generally acknowledged concept for assessment of data usability (or fitness for use) of individual datasets. Some of the properties describing data usability are not available or not transparent to users and requirements for other properties cannot be matched with standards. Furthermore, current certifications and accreditations of data repositories only allow limited conclusions on the re-usability of individual datasets. Thus assessing the fitness for purpose and making a decision whether to reuse a dataset is not straightforward. This situation reduces the chances of shared data being reused and in case of reuse could decrease the reliability of research results.

Firstly, a concept of data fitness requires assessment of quality criteria to be included as well as the weighing of each of those criteria. The process should preferably lead to the development of a corresponding metric. Secondly, we want to find effective ways to expose and communicate this metric, for e.g. by using a labelling or tagging system whereby different usability levels are importantly made explicit.

The proposed working group would work towards the following deliverables:

- The definition of criteria and procedures for assessment of fitness for use
- The development of a system of badges/labels communicating fitness for use of individual datasets

Criteria would be used such as:

- Trustworthiness of the data centers/repositories (such as assessed through existing certifications: DSA-WDS, DIN, ISO 16363 etc.)
- Data accessibility in terms of discoverability, openness, interoperability etc.
- Level of curation applied (citability, metadata completeness, data harmonization, machine readability etc.)

Value Proposition

The following stakeholders would benefit:

- **Researchers that deposit data** can visibly improve and communicate the quality of their datasets, thereby increasing reuse and citation, which provides researcher with additional metrics showing their productivity.
- **Researchers that reuse data** can more easily assess the quality of a dataset and in particular its fitness for their reuse. This makes reuse of data safer and more efficient.
- **Data centers/repositories** can offer better quality data publication services - such as more transparent curation - thus increasing the overall usage of services which in turn might lead to improving the facility's financial base.
- **Science publishers** can better integrate referenced data into the editorial process and improve the review of articles and related datasets as well as citations and cross-linking of datasets and literature as a result of more transparency about data usability.
- **Funders** can make provisions for funded data archiving and publication services in accordance with their funding requirements and expectations in terms of data fitness for use (and reuse).

Overall impacts:

- Improved and standardized data publication services

- Improved communication of data fitness for use
- Improved reliability and efficiency in the reuse of research data

Engagement with existing work in the area

Data fitness for use has been addressed in literature over the last 20 years. The topic received more attention with the general increase of data production. The following gives a brief overview of selected publications. In 1998 Tayi and Ballou stated that the concept of data quality is relative with quality being dependant on users and applications. Some authors concentrated on special aspects as for example assessment of accuracy of geospatial data (de Bruin 2001) or de-duplication relevant for example to data mining approaches (Christen & Goiser 2007). A further aspect is preservation of usability of sensitive data (Bhumiratana & Bishop 2009). In 2007, the OECD underlined the importance of efficiency in reusing data (OECD 2007). For example efficient compilations of data from multiple providers require harmonized and machine readable data, in particular for data with high volumes. Correspondingly the FAIR Data Publishing group supplies a set of principles for publishing data and emphasizes machine readability of data as one of the major challenges (Wilkinson et al. 2016). More recently authors also started to investigate data usability with respect to big data approaches (Jianzhong 2013). The effect of peer-review on data quality, respectively usability was stressed by Lawrence et al (Lawrence 2011) and an editorial in the Nature Scientific Data Journal (2016). Costello linked data fitness for use with the data publication concept (Costello 2013). Also worthy to note is the ISO/IEC 25012 data quality model (ISO/IEC 2008) and the ISO 8000 Requirements for Quality Data (ISO 2009). The W3C Data on the Web Best Practices Working Group elaborated vocabularies needed to describe data quality and highlights the importance of data provenance (W3C 2016), which – if applicable — should include also detailed information about physical samples, for example in the case of biocollections (Bishop 2016). Finally, fitness for use of datasets should be transparent and comprehensive to users. The effectiveness of using badges or labels for this purpose was shown by Kidwell et al (Kidwell 2016).

Besides works published in the literature the WG can build on a wide range of activities that are relevant to the aims and scope of the group. In particular:

- The Working Group would operate under the umbrella of the RDA-WDS Data Publishing IG and RDA/WDS Certification of Digital Repositories IG

- This Working Group will follow up on the work of the RDA/WDS Data Publishing Workflows WG and assess the impact of workflows on fitness for use
- This Working Group will follow up on the work of the Repository Audit and Certification DSA–WDS Partnership WG and develop a related certification system for individual datasets
- The Working Group would incorporate the criteria defined by the FAIR working Group (Wilkinson 2016) as a starting point.
- The Working Group will collaborate with the NIH Commons FAIR metrics group to elaborate on the FAIR criteria (NIH 2016)
- This Working Group would incorporate the W3C data quality vocabulary to define quality processes (W3C 2016).

Work Plan

Work will be along four strands:

1. Descriptions and definitions of data fitness criteria
2. Development of a fitness for use label at the level of datasets
    a. Conceptual model
        i. Selection and evaluation/weighing of criteria with respect to the different aspect of fitness for use such as curation or accessibility
        ii. Considerations for adoption by stakeholders (archives/repositories: for e.g built into workflows, science publishers)
    b. Design of label/badge
3. Development of service components
    a. Investigate how a fitness of use concept can be integrated into current certification procedures for data centers/repositories (WDS/DSA)
    b. Investigate data centers/repositories service components
    c. Setup of a testbed of several data centers/repositories
4. Governance and sustainability:
    a. Concept for a long-term organizational structure to operate elaborated services successfully and in a way that meets the needs of all stakeholder groups. This stream will also deliver a process through which new organizations can connect to the service.

Deliverables

- Document defining fitness for use criteria
- Description and design of fitness for use label (badge system)
- Concept for a certification procedure including the fitness for use aspect
- Concept for a data centers/repositories service components
- Adoption plan including certifying organizations and governance

Milestones

- Fitness for use concept ready
- Identification of certifying organization(s)
- Setup of a testbed with several data centers/repositories and science publishers
- Fitness for use label operational

Mode & frequency of operation

- Telecons every 4 weeks
- Face to face meetings during RDA plenaries and at least one additional workshop. RDA plenaries in particular will be used to engage the wider community and coordinate the work with related groups.
- Additional meetings of subgroups working on particular deliverables including adoption

Timeline

Months	Action	Deliverable
January - March 2017	Define criteria	Overview of criteria, for discussion at 9th plenary meeting
April - September 2017	Pilot assessment of criteria	Report on outcomes of pilot, for discussion at 10th plenary meeting
October - December 2017	Development of badge system and integration with current certification schemes	Guide for repositories
January - July 2018	Integration of data services into framework	Governance structure and adoption plan

Adoption Plan

Members of the proposed working group are planning to carry out a pilot during the 12-18 month timeframe in which they incorporate the insights that come out of the working group. In this pilot, a first assessment of the fitness for use of individual datasets will be carried out. This simultaneous pilot will provide the working group with important information about both benefits of and challenges with adoption which will make it easier for additional organizations to adopt the outcomes of the working group. The goal is that at the end of the 18 month timeframe, a first network of adopters will exist.

Initial Membership

Bradley Wade Bishop (Univ. Tennessee)
Helena Cousijn (Co-Chair)
Michael Diepenbroek (Co-Chair)
Amy Nurnberger (Columbia University Libraries)
Ingrid Dillo (DANS)
Stephane Pesant (MARUM)
Mustapha Mokrane (ICSU-WDS)
Markus Stocker (PANGAEA)
Rob Hooft (DTL)
Peter Doorn (DANS)
Christina Lohr (Elsevier)
Robert R. Downs (CIESIN, Columbia University)
Daniel Fowler (Open Knowledge International)
Martina Stockhause (WDC Climate, DKRZ)
Ian Bruno (CCDC)
Tim Smith (CERN/Zenodo)
Donna Scott (NSIDC)

References

Bhumiratana B & Bishop M (2009) Privacy aware data sharing: balancing the usability and privacy of datasets, in: Proceedings of the 2nd International Conference on PErvasive Technologies Related to Assistive Environments, https://doi.org/10.1145/1579114.1579187

Bishop, B. W. & Hank, C. F. (2016) Fitness for Use in Data Curation Profiles for Biocollections [Presentation] American Society for Information Science and Technology Annual Meeting, October 2016, Copenhagen, Denmark

de Bruin S, Bregt A, van de Ven M (2001) Assessing fitness for use: the expected value of spatial data sets, International Journal of Geographical Information Science, v15, no5, p457-471

Christen P & Goiser K (2007) Quality and Complexity Measures for Data Linkage and Deduplication, in: Guillot FC & Hamilton HJ (eds) Quality Measures in Data Mining, Studies in Computational Intelligence pp 127-151

Costello M et al (2013) Biodiversity data should be published, cited, and peer reviewed, Trends in Ecology & Evolution, p1-8

International Renewable Energy Agency (2013) Data quality for the Global Renewable Energy Atlas – Solar and Wind, https://goo.gl/a8xr1Q

ISO (2009ff) Data quality, https://en.wikipedia.org/wiki/ISO_8000

ISO/IEC (2008) Data quality model, http://www.iso.org/iso/catalogue_detail.htm?csnumber=35736

Kidwell MC, Lazarević LB, Baranski E, Hardwicke TE, Piechowski S, Falkenberg L-S, et al. (2016) Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. PLoS Biol 14(5): e1002456. http://doi.org/10.1371/journal.pbio.1002456

Lawrence, B., Jones, C., Matthews, B., Pepler, S. & Callaghan, S. (2011). Citation and Peer Review of Data: Moving Towards Formal Data Publication. International Journal of Digital Curation 6, 4–37

Li Jianzhong & Liu Xianmin (2013) An important aspect of big data: data usability, Journal of Computer Research and Development, v6

NIH Commons FAIR metrics group (2016) WG interim report, https://goo.gl/n4PpWv

OECD (2007) OECD Principles and Guidelines for Access to Research Data from Public Funding, http://www.oecd.org/sti/sci-tech/38500813.pdf

Scientific Data Journal (2016) Let referees see the data, editorial, Nature Scientific Data Journal, 3, 160033. http://doi.org/10.1038/sdata.2016.33

Tayi GK & Ballou DP (1998) Examining data quality, Communications of the ACM, v41, no2, p54-57

W3C (2016) Data on the Web Best Practices: Data Quality Vocabulary, W3C Working Group Note, https://www.w3.org/TR/vocab-dqv/#mapping-ISOZaveri

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., and Baak, A. **et al.** (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018, http://doi.org/10.1038/sdata.2016.18