Criteria for appraisal and assessment of research data

upon submission to a data repository

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA Research Data Management Helpdesk

Roman Gerlach, Bettina Färber, Birgitta König-Ries

MichaelStifelCenterJena for Data-Driven and Simulation Science

http://www.researchdata.uni-jena.de/

Motivation

Submitting a dataset to a data repository is the process of transferring a data object from the private domain to the shared or public domain (cf. domain model by Treloar & Klump 2019). The data provider's intention is to preserve the data and in most cases also to make it accessible to a broad audience. Data repositories receiving the data need to make a number of decisions on how to treat the submitted data to fulfill the expectations of the data provider. To our experience, terms and conditions of repositories often do not cover all aspects needed. Additionally, in many cases the process of verification of compliance with these terms and conditions is based on the individual expertise and the experience of the data curation personnel and no formal and transparent process is in place.

Objectives

- Provide data managers and data curators with a criteria catalogue to **evaluate data submissions**.
- Provide a practical guide **specifying information requirements** that need to be collected at submission time (e.g., retention period and responsibility for data disposal are typically not part of standard metadata).
- Support and guide data providers to **prepare data accordingly upfront** and to provide all information needed at submission time.
- In result, **increase effectiveness and efficiency as well as transparency** of data repositories.

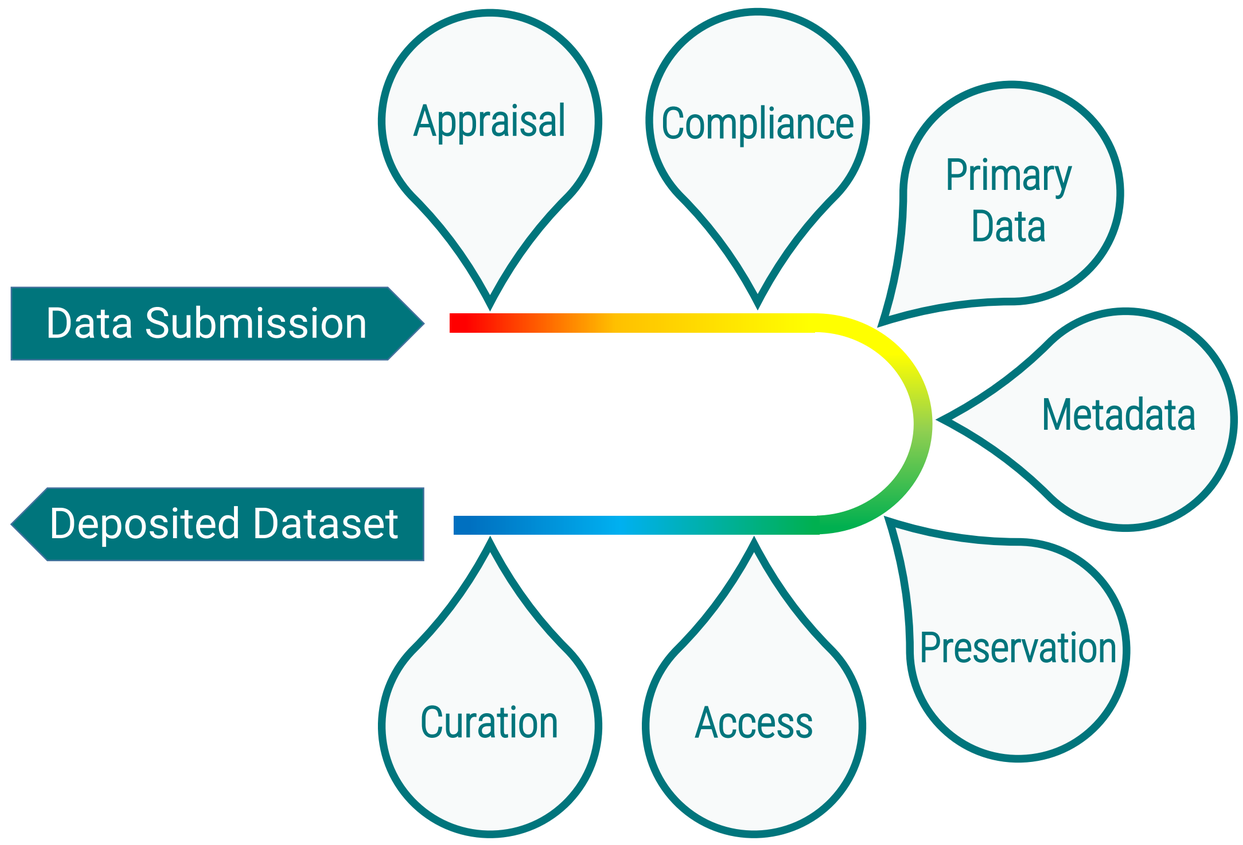
Approach

For an institutional repository, a draft catalogue had been designed previously by the first author based on the criteria of Whyte & Wilson (2010). It served as a starting point for a 1-day workshop with data management support staff, data managers and data curators in which the criteria were discussed, complemented and re-structured. This work was continued and finished collaboratively using an online document. The 90 criteria in the catalogue are phrased as questions, and potential answers are provided (paper in preparation).

References

Treloar, A. & Klump, J. (2019): Updating the Data Curation Continuum: Not Just Data, Still Focussed on Curation, More Domain-Oriented. IJDC 14 (1), pp. 87-101. DOI: 10.2218/ijdc.v14i1.643

Whyte, A. & Wilson, A. (2010): How to Appraise and Select Research Data for Curation. Digital Curation Centre: Edinburgh. http://www.dcc.ac.uk/resources/how-guides/appraise-select-data



Assessment criteria for data curators/data managers (selection)

No.	Criterion / Question
1 Appraisal	
1.1.1	Does the data conform to the **repository**'s purpose and **scope**?
1.4.2	Is the data **relevant** enough to be published?
3 Primary Data	
3.1.1	Does the submitted package contain **software/scripts?**
3.2.1	Are files free of any **file protection mechanism** (e.g. DRM, password, copy or print protection?
3.3.1	Does the provided **file format** correspond to the file content (e.g. table in a word file)?
3.3.2	Is the file format part of the "white-list" of accepted formats of the repository?
3.4.1	Which kind of **file encoding** is used?
3.5.1	Do **file names** describe the data in each file sufficiently?
3.6.1	If the data is **tabular**, is there more than one table on one spreadsheet in e.g. MS Excel?
3.7.1	Are **variable** names unique within the dataset?
3.7.3	Are the variables linked to a controlled vocabulary?
3.7.5	Are the variables in the dataset consistent in terms of data type?
3.7.6	Are the variables in the dataset consistent in terms of units?
3.8.1	Does the primary data contain the **units**?
3.8.2	Are the units linked to a controlled vocabulary?
3.9.1	Are missing **values** clearly denoted (distinguished from null values)?
3.9.2	Are there redundant information in the data: (e.g., multiple variables containing the same information)?
3.9.3	Are there any derived variables, calculated solely from other variables in the same dataset?
3.9.4	Are values of individual variables reasonable (e.g. within a realistic range)?
4 Metadata	
4.1.1	Is there a **readme**.txt **file**?
4.1.2	What is the content of the readme.txt file?
4.3.1	Is the **metadata** provided in one **language** only?
4.3.2	In which language(s) is the metadata provided?
4.4.1	Does the **title** describe the dataset precisely? (What? Why? Where? When? How?)
4.4.2	Is the title comprehensible by non-domain experts?
4.5.1	Is there at least one corresponding **author** with a valid email address or telephone number?
4.5.3	Are authors identified by e.g. ORCID/GND/VIAF?
4.6.1	Does the **abstract** describe the dataset sufficiently and comprehensive? (What? Why? Where? When? How?)
4.6.2	Are limitations specified in the abstract?
4.6.3	Is the instrumental setup described in sufficient detail?
4.6.4	Is the methodology (procedure) described in sufficient detail?
4.8.1	Is there precise information about the **time of data collection** or generation?
4.8.2	Does this time fit with times in primary data?
4.8.3	If applicable, is there information about the **spatial coverage** of the dataset?
4.9.1	Is the structure of **primary data** described in sufficient detail?
4.9.2	Are the variables described in sufficient detail?
4.9.3	Are the units described in sufficient detail?
4.9.4	Are data types clearly specified? (e.g., using terminologies/ontologies)
4.9.7	Is the precision of primary data defined?
4.10.2	**Quality**: Is the metadata rich enough?
5 Preservation	
5.2.1	What kind of **digital migration actions** must be applied to the dataset?
5.2.3	What kind of digital migration actions can be applied to the dataset?
6 Access	
6.1.3	Is the length of the **embargo period** reasonable?
7 Curation	
7.1.1	Which **curation level** does the dataset belong to, according to metadata?
7.1.2	Which curation level does the dataset belong to, according to primary data?
7.3.1	What is the **review status** of the submitted data?
7.4.1	Does the dataset contain different **distribution packages** (e.g., aggregations)?

Information requested from data providers (selection)

No.	Criterion / Question
1 Appraisal	
1.2.1	Is there a **data policy** that mandates submitting this dataset to this repository?
1.3.1	What is the **purpose** of this data submission?
1.5.1	Has this dataset been submitted to any **other repository**?
2 Compliance	
2.2.1	**Sensitivity**: Does the dataset contain personal information?
3 Primary Data	
3.3.5	Is the **file format** accepted by the community? (Will it exist in 10 years?)
4 Metadata	
4.2.1	To which **metadata standard** does the data comply?
4.2.2	Does the dataset contain additional metadata (beyond the one requested by the repository)?
4.7.1	Are the (domain specific) **keywords** backed by controlled vocabularies/ontologies?
4.7.2	Which vocabularies/ontologies have been used?
4.13.1	**References**: Is the dataset related to any physical research object?
4.13.2	References: Is the dataset related to any other research object (e.g., paper)?
5 Preservation	
5.1.1	**Retention period**: For how long should the data be stored?
5.1.2	What happens when the retention period terminates?
5.1.3	Who evaluates whether a dataset can be disposed?
5.2.2	What kind of **digital migration actions** should be applied to the dataset?
6 Access	
6.1.2	When should the data be openly available (end of **embargo period**)?
6.2.1	What is the **level of** metadata **sharing** during the (potential) embargo period or if archiving only?
6.2.2	What is the level of primary data sharing (archiving only)?
6.2.3	What is the level of primary data sharing during the (potential) embargo period (publishing)?
6.2.4	What is the level of primary data sharing after the embargo period (publishing)?

Acknowledgement: This work was conducted as part of the eeFDM project funded by the Federal Ministry of Education and Research (Germany). For valuable input by discussing the criteria and the catalogue draft we would like to thank:
Friedrich-Schiller University Jena: Rainer Vock and Volker Schwartze (Research Data Management Helpdesk); Andreas Ostrowski, Jitendra Gaikwad, Franziska Zander and Cornelia Fürstenau (Heinz Nixdorf Chair for Distributed Information Systems); Maria Sorokina and Christoph Steinbeck (Cheminformatics and Computational Metabolomics)
Thüringer Universitäts- und Landesbibliothek: Angela Hammer, Isabel Hoffmann, Thomas Mutschler
As well as: Johannes Biermann (Niedersächsische Staats- und Universitätsbibliothek Göttingen), Janine Felden (MARUM Center for Marine Environmental Sciences), Jens Kattge (Max Planck Institute for Biogeochemistry), David Russell (Senckenberg - Leibniz Institution for Biodiversity and Earth System Research), Dagmar Triebel and Tanja Weibulat (Botanische Staatssammlung München)

eeFDM

Bundesministerium für Bildung und Forschung