Data curation checklist YODA

Lena Karvovskaya, Research Data Manager UB Utrecht University

ORCID iD https://orcid.org/0000-0001-7777-5603

May 2, 2019 – version 1.0 created

This is version 5

DOCUMENT HISTORY[1]

NAME	DATE	VERSION	DESCRIPTION
Lena Karvovskaya	2019-05-02	1.0	First draft created
Danny de Koning, Vincent Brunst, Frans Liagre de Böhl, Lena Karvovskaya	2019-06-28	2.0	The second draft created: two checklists, for archiving and for publication. Sections of the checklist: Authorization, Documentation, Metadata, Files and Folders
Ton Smeele and Danny de Koning	2019-09-20	3.0	The third draft created. The questions are reformulated into quality properties. The checklist took the shape of a 3-column table
Ton Smeele	2019-10-02	4.0	Additional check for publication: for every creator and contributor, a PID is mandatory.
Danny, Lena	2019-10-21	5.0	Language, style

[1] See the list of changes https://docs.google.com/spreadsheets/d/1NT-bMkCByELqI9yRrKTdvyvY3WpCYiSkjhpy7pAch2Q/edit?usp=sharing

1. Introduction

This checklist is created for the data managers working with Yoda[2][3]. Data managers evaluate requests for data to be archived or published trough Yoda to ensure that the data is suitable for storage according to, among others, FAIR principles. The data manager assesses whether the data is well described through different forms of metadata, has a good folder structure, follows naming conventions, and whether the data uses preferred formats.[4]

When evaluating data archiving requests, we distinguish between two types of data packages:

- Archival package (can be used for verification/replication) – data meant to be archived but not necessarily shared with others,
- publication package – data intended for publication.

In both cases, the data has to contain enough information to be understandable and useful for other researchers in the discipline. The difference between the two packages is that the publication package should not contain any information that cannot be openly shared in a legal way.

The checklists are meant to be filled in for every data deposit in Yoda through archiving (submitting to the vault) and publishing. The checklists ensure thata baseline level of quality control is performed by data managers in the same way for all data deposits in Yoda. The checklists may also aid in keeping track of the data deposited in Yoda and to determine why certain decisions were made years after the deposition.

We are considering the situation in which the researcher already has access to a Yoda instance. In this case, the researcher can submit a dataset from the research environment to be archived in the vault and send the data manager a notification.

2 This checklist is based on a checklist created by Lena Karvovskaya for EPOS data deposition procedures. It is heavily inspired by the CURATE checklist created by Data Curation Network https://docs.google.com/document/d/1RWt2obXOOeJRRFmVo9VAkl4h41cL33Zm5YYny3hbPZ8/edit

3 For an overview of the terminology used see Yoda's glossary https://yoda.uu.nl/ and LCRDM Glossary https://www.edugroepen.nl/sites/RDM_platform/LCRDM%20glossary/LCRDM%20Glossary.aspx

4 In contrast to most publishers, Yoda does not have a person responsible for the publication, an official Editor or Data Curator. The data managers only have advisory role wrt to the data being published. The responsibility for the quality of the publication lies solely with the researcher. In the future, this aspect might be made more explicit with a pop-up window that appears before the dataset publication. The data manager can point the researcher to the problems with the dataset and advise the researcher on how to improve it. However, the data manager does not have the power to change the dataset for the researcher or to ban the researcher from publishing in case of difference of the opinion. The data managers are advised to contact the responsible research directors in case there is a problem.

2. Archiving checklist

2.1 Authorization.

Question/Additional information	Checklist item	DM notes
Who is in charge of the data? Identify the rights holders.	The stakeholders behind the dataset are identified and documented for internal administration. [5]	
Is the researcher submitting the data the creator of the dataset?	The creator of the dataset is documented. The relation between the creator and the person submitting the dataset is established.	
Are there multiple creators of the dataset? Is the individual submitting responsible/in charge of the work?	The names and affiliations of all creators of the dataset are documented; it is confirmed that the person submitting the work is authorized to deposit the data (see also "documentation and metadata").	
Is the dataset re-use of already existing data? If yes, where is the data from? (See "documentation" checklist)	The origins of the data are documented in case the dataset presents re-use of already existing .	
Who is the funder? (see also "metadata" checklist)	The funder behind the research is documented with the grant number.	
Are there any special regulations with respect to rights holders of the data? For example, is an external funder the rights holder of the data collection it funds?	Any special regulations with respect to rights holders of the data are documented, if known.	

2.2 Documentation and metadata.

2.2.1 Documentation

Question/Additional information	Checklist item	DM notes
Does the dataset include a file with documentation? By documentation we understand a readme file in pdf or txt format	The data documentation is included in the dataset.	

[5] Intellectual property is a complex area, especially when applied to data. As a data manager, you are not expected to be a legal expert on IP.

and a codebook[6]. Documentation provides context for the data and explains how the data should be read. Documenation includes information about the software used to create/open the files, including the version of the software.		
Are the discipline-specific aspects of the dataset considered when reviewing the documntation?	Depending on the discipline and the nature of the dataset, the following aspects might be importnat for contextualization: □the setup of the whole research project □experimental set up, if there are experiments □the variables of the dataset □self-defined abbreviations should be mentioned. □for tabular data, headers should be defined in the documentation □the units of measurement □the instruments □for special file formats, software required to open the files, including version (see “files and folders” checklist) □sampling method □sample size □algorithms and/or transformation scripts that derived secondary data from raw data □if primary data is not contained in the dataset, there should be a link or a reference to the primary data (see “authorisation” checklist) □setup of the folder structure (which files can be found where in the data package). □explanations of what scripts and code do	
Is the dataset complete? Completeness of the	□Complete list of files is present in the accompanying documentation.	

6 A codebook describes the contents, structure, and layout of a data collection. A well-documented codebook "contains information intended to be complete and self-explanatory for each variable in a data file": https://www.icpsr.umich.edu/icpsrweb/content/shared/ICPSR/faqs/what-is-a-codebook.html

dataset is verified by checking the submitted data files and the accompanying documentation. Are there missing parts or parts that are not mentioned in the documentation?	□All the files listed in the data documentation are in the dataset. □There are no files that are not listed in the documentation.	

2.2.2 The metadata fields in the schema

Question/Additional information	**Checklist item**	**DM notes**
Different instances of Yoda have different metadata schemes. The dataset should only be archived if the mandatory fields of the relevant scheme are filled.	□Structured *metadata* is provided. □Mandatory fields are filled.	
Name convention followed.	The names of the contributors follow the convention: **LastName, Firstname**[7]	
Ask the researcher about ORCID, SCOPUSID, RESEARCHERID. For publication, providing persistent identifiers for every contributor is mandatory (see publication checklist).	Author's identifier(s) like ORCID are provided if available[8].	
Is the contact information provided? Is the research program mentioned? The research program can also be a discipline.	□A contact person with the contact information is added to the metadata. □Reference to the research program is added[9].	

7 Currently, the name is entered as a free string. Therefore, it is important to make sure that all entered names follow this convention. In the nearest future, two separate fields for the first and second name will be implemented in Yoda

8See ORCID: (publisher neutral): https://orcid.org/orcid-search/quick-search/?searchQuery=
SCOPUSID: (Elsevier) https://www.scopus.com/search/form.uri?display=authorLookup
RESEARCHERID: (Thomson Reuters): http://www.researcherid.com/ViewProfileSearch.action

9 There is an ongoing discussion as to what is the best suitable point of contact for an archived or published package. It is problematic to leave contact details of one specific person, as this person might leave the UU, the country or even pass away. It could be a department or a laboratory. In case there are structural solutions on these questions for a certain community, the data manager is expected to follow them. If there is a structural solution for a given discipline the data manager is expected to follow this solution and make sure that the contact information is correct.

In some metadata schemes, adding a contact person will require repeatedly adding the contributor with the contributor type “contact person”. Reference to the research program should be provided using the contributor type "Project Leader" and name of the program.		
The minimal retention time for the dataset depends on the discipline and the type of data. For instance, medical data may need to be kept significantly longer than the 10 years required for reproducing research. Data managers should have a list of data types with minimal retention times as a point of reference[10]	The retention time for the dataset meets the required minimum.	
Data managers should have a list of standards/preferred list of keywords (Tags) for their disciplines[11].	□Keywords are not combined in a single field. □Keywords comply with standards/preferred list used in the discipline.	

2.3 Files and Folders.

2.3.1 File naming

Question/Additional information	Checklist item	DM notes
File and folder naming.	□The file and folder naming are logical □Files and folders are named in a consistent and descriptive manner[12]	

10 A list with data types and retention times needs to be created

11 For the existing Yoda environments, one should be able to see the lists with preferred discipline-specific keywords

12Currently, there are no general standards for Yoda. The data manager can make some suggestions according to the best practices. For example, the filename can included include version, date, project abbreviation,

	If there are any file naming conventions for the discipline in question, the file naming follows these conventions.	
Are there spaces and unusual symbols in the names of the files? In general, special characters should be avoided to ensure that files can be read by any operating system workstation.	There are no special characters[13] in filenames.	
Is it the case that the names of some files or folders only differ from each other by the use of upper/lowercase letters? For example, Windows does not always differentiate between upper/lowercase in filenames.	Upper/lowercase letters do not contribute to the meaning differences in file names.	
	Advise the researcher to adjust the names of files and folders if necessary (e.g. versioning of files should not include the words final, old, new, etc but instead -> date_v01 etc.)	

2.3.2 File formats

Question/Additional information	**Checklist item**	**DM notes**
Are the files in the dataset in future-proof formats[14]? Use Yoda 1.5+ feature to check the data folder for compliance with DANS and 4TU file types.	If possible, files are in open, non-proprietary, future-proof formats.	
If proprietary formats or specialized formats are chosen, is feasible to make derivatives of the files in preferred formats as additional files?	□If feasible, for proprietary formats, derivatives of the files in preferred formats are added to the data package. (xls -> xls and its txt/csv derivate) □If feasible, for specialized formats from specialized equipment, derivatives of the	

abbreviation of contents, etc. For more examples, Stanford Libraries provides some advice in their best practices for file naming: https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-naming

13 See https://en.wikipedia.org/wiki/Filename for a list of special characters.

14 https://dans.knaw.nl/en/deposit/information-about-depositing-data/before-depositing/file-formats?set_language=en

(xls -> xls (copy) and csv/txt of the same file, etc.).	files in preferred formats are added to the data package.	
Is it clear which software will be needed to open files with specialized formats?	Documentation specifies which software was used and is required to read the files (see "documentation" checklist).	

2.3.3 Folder structure

Question/Additional information	**Checklist item**	**DM notes**
	If there is a folder structure recommended for the given discipline, this folder structure is obeyed[15].	
Is raw data separated from processed and analyzed data?	Raw data is separated from processed and analyzed data, unless there are good reasons not to do so.	
	There are no empty folders.	
Are the pathnames long? The maximum length of a pathname is limited depending on the operating system. and the files should be able to be read on various systems. Max pathname must be less than 4096 characters including Yoda prefix such as zone name, home, and research group name.	The nesting of files and folders is not too deep.	
Advise the researcher to delete hidden files like desktop.ini, and indexing files like Apple ._, DS_Store, etc.	There are no hidden files like Apple ._, DS_Store, etc	
Does the data set contain parts which can be considered as sensitive? In general, data can be classified into three types of datapackages:	□Sensitive data should be separated from non-sensitive data. □Sensitive data should be stored in separate folder structures and preferably be deposited as separate data packages. This allows for both data packages to be reused separately.	

15 At the moment, there are no general templates for folder structures for Yoda

1) anonymous data 2) pseudonized data (typically shared as "restricted use") 3) privacy/patent/etc - sensitive data (typically "restricted use" or "closed")		

2.3.4 Data validity

Question/Additional information	Checklist item	DM notes
Can you assess the data validity ?	Software/scripts have been used to make sure the files are not corrupt[16].	

3. Publication checklist

3.1 Authorization.

Question/Additional information	Checklist item	DM notes
Does the data set contain parts which can be considered as sensitive? In general, data can be classified into three types of datapackages: 1) anonymous data 2) pseudonized data (typically shared as "restricted use") 3) privacy/patent/etc - sensitive data (typically "restricted use" or "closed")	The dataset does not contain data that raises questions wrt to: □Privacy issues (personal data) □Commercial issues (data can be provided by third party, there might be patent involved, etc.) □Political issues □Legal issues	
Who is in charge of the data? Identify the rights holders.	It is clear who is the creator(s) of the data. □The contact person is authorized to publish the data[17] □Any special regulations with respect to ownership of the data are documented if known.	

16 Some automatization to assist the data managers is being developed by Research IT team.

17 The data manager is not expected to contact the head of the department for each submission. The expectation is that the data manager asks the contact persons if they are authorized and receives an oral confirmation.

3.2 Completeness of the metadata.

Question/Additional information	Checklist item	DM notes
	The provided structured *metadata* meets the criteria that the YODA community agreed on.	
	'Related Data package' field is filled whenever possible.	
Are related publications included in the metadata? (See for instance Related Data package metadata field)	For some Yoda communities, if there are any related publications, such as journal articles or data reports based on the data, describing the data, etc. the relevant PIDs are included in the metadata	
Ask the researcher about ORCID, SCOPUSID, RESEARCHERID. Persistent identifiers are crucial to link the data package to researchers in Pure in an automated way.	Persistent identifier(s) like ORCID are provided for the creator and every contributor[18].	
Is the contact information provided? Is the research program mentioned? The research program can also be a discipline. In some metadata schemes, adding a contact person will require repeatedly adding the contributor with the contributor type "contact person". Reference to the research program should be provided using the contributor type "Project Leader" and name of the program.	The contact information of the contact person/organization is added to the metadata.	
	Valid license type is used	

[18]See ORCID: (publisher neutral): https://orcid.org/orcid-search/quick-search/?searchQuery=
SCOPUSID: (Elsevier) https://www.scopus.com/search/form.uri?display=authorLookup
RESEARCHERID: (Thomson Reuters): http://www.researcherid.com/ViewProfileSearch.action

Is embargo date reasonably defined? For example, when the datapackage is to be stored for 10 years and the embargo date expires a day before the retention date of the datapackage, that will not be considered 'reasonable'	If an embargo date is defined in the metadata, it represents a reasonable period.	

4. (Optional) researcher's awareness

To ensure that the **content** of the dataset complies with the quality parameters, the data manager has to rely on the researcher or the researcher's PI, the domain specialist. By contrast to the data manager, the domain specialist can provide a quality assessment of the data itself, not just completeness and presentation.

Below we sketch the list of controls that cannot be expected by default from a general data manager. The researchers who want to have a high-quality data publication can consider asking other domain-specialists for a peer-review of their datasets.[19]

- Research. The domain specialist can evaluate the validity of data and the adequacy of the data selection. If the data package is supplement to a journal article, this evaluation is indirectly done by the peer reviews: any anomalies with the data would be noticeable in the text of the article. Only domain specialists can evaluate such parameters of the data package as:
    - Scientific validity,
    - Veracity,
    - Accuracy,
    - Completeness
- Documentation (see documentation above). While the data manager can evaluate the documentation for completeness, the domain specialist can determine if the documentation of the data is sufficient to understand and reuse the data. There must be sufficient background information in the documentation. The documentation explains the dataset, including such aspects as:
    - How the data was created,
    - Data selection process,
    - Measurements that were taken,
    - Transformation,
    - Analysis techniques
    - Preservation,
    - Versioning,
    - Methodology,
    - Study aims
    - Standards used.
- The metadata. The fields of the metadata form must be filled correctly (see *data manager* task 3 above). The domain specialist can evaluate, among others, if:

19 There are also curation models that use discipline-specific expertise to enhance the quality of the curated datasets. See, for example the work of the Data Curation Network in the US: https://datacurationnetwork.org/

- The keywords provided are sufficient to contextualize the dataset;
- The key references that appear in the documentation are included in the metadata (‘Related Work’);
- The interpretation of the metadata fields corresponds to the discipline.