DCN IMLS Workshop

Worksheet for evaluating data quality: Simulation data

L. Wynholds, Marsolek, et al, 9/20/2023

The 'CURATE(D)' CHECK step revolves around inventorying and reviewing the contents of the dataset and verifying that it is appropriate for the repository. This often includes:

- Review to ensure data is in scope for the repository
- Inventory the contents of the data files (e.g., open and sample the files or code)
- Verify all metadata provided by the researcher; check available documentation

The data curator should read the readme.txt and associated metadata and decide whether they provide enough information to answer the following questions. If not, consideration should be made about having a discussion with the researcher about whether the readme.txt or related metadata needs additional information[1].

Getting Started:

1) Download the readme.txt file for the dataset
2) Open the readme file
3) Open the repository download page for the dataset
4) See if you can answer the following questions for the dataset (see checklist on following page)
    a. Actual metadata may vary depending on the specific dataset and type

[1] The checklist below was adapted from 'Checklist developed by the NIH DMSP Guidance for Data Support Services Working Group'

Metadata Checklist:	
Data Type	a. What is the data type, format, size, and number of files? b. Whether the dataset is a subset, part, or related to other datasets **c. Does the data description include the level of aggregation, de-identification, processing, or cleaning that was done in the data?** **d. Does the metadata include the source(s) of any secondary data or previously collected data used in the production of the dataset (e.g. inputs, calibration, reference datasets, ground truthing datasets, etc)** e. Metadata available that describes the above adequately for others to use (e.g. readme.txt)
Related Tools, Software, Code & Publications	i. State whether specialized tools are needed ii. State which code versions and which OS were used to produce the data iii. State how to access the code/software/tools (e.g. citation to code or download website) iv. State whether the dataset was used in articles or other publications, and if so, citations to published or in-process title/author/journal *v. Reference code group mat'ls (link)*
Data Standards	Are there data standards used in the dataset? If so, which ones and where are they specified? (cite which standard) a. Metadata schemas b. Standard terminologies/controlled vocabularies c. Content/encoding standards d. Common data elements e. Identifiers (PIDs)
Data Access	i. Is the DOI and/or other PID listed in the metadata? ii. Are there any legal, technical or ethical considerations with sharing the data? iii. Did the funders require that the data be made accessible to the public? If so is that information provided to the end users with consideration for future use? iv. Did the funders require that the data be available for a set period? If so, how long/what dates?
Data Restrictions	a. Are there limitations to use? (e.g. IP restrictions, embargos, requires human subject IRB review for use, etc) b. Are there instructions for gaining access to restricted datasets? c. Embargo period? d. Other access controls?
Data Stewardship Data Provenance	i. Who was responsible for producing/QC for the data? ii. Who can be contacted with questions regarding the data? (usually the first author unless otherwise specified) iii. Funding acknowledgements – Who funded the research, what is the grant number? If there is a compliance requirement for open access, does the metadata specify who is responsible for ensuring that the data is available once deposited? (e.g. the archive, the researcher, or the funding body) iv. Are all the authors, funders and other contributors acknowledged in the metadata? v. Are there any license agreement requirements or IP restrictions that downstream users should know about?