This checklist guides humanities researchers and humanities librarian liaisons on key considerations for making their data findable, accessible and clear to interested scholars and institutions.

Element	**Definition**	**Guidance/Examples**
Title*	Typically the name of the file or file set in the repository.	**E.g.** *Art of the Aeneid Dataset* **E.g.** *2014 Survey of Licensed Street Vendors in São Paulo*
Dataset Persistent ID	A dataset persistent identifier is a permanent and unique referral to an online digital object, independent of change in the actual location.	A common identifier is a Digital Object Identifier (DOI). Upon dataset publication, TDR will provide an automatically generated DOI in the ***Dataset Persistent ID*** field. **E.g.** doi:10.18738/T8/L2SJQT
Related Publication	Here, you can cite a related publication to your dataset. This field will prompt you for a citation and a persistent ID.	You may have an article DOI already if you have published an article related to your dataset. **E.g.** http://doi.org/10.5334/johd.13
Description*	Briefly summarize the type of study (or studies) to help others understand the purposes for which the data are being collected or created.	**Questions to consider:** - What is the nature of your research project? - What research questions are you addressing? - For what purpose are the data being collected or created? Descriptions can vary in length. Here is what a description could look like: "The Colección Conflicto Armado contains scans of political posters from the period of the Salvadoran Civil War. The posters were produced in support of anti-government activities in El Salvador as well as solidarity organizations outside the country. The posters are housed at the Museo de la Palabra y la Imagen (MUPI) in San Salvador. This collection also includes metadata created for visualization and a guide for creating visualizations with ImagePlot."
Kinds of Data*	Describe data features, such as type of data, number of files, file types, file size. This should explain what kinds of files will be downloaded.	**Questions to consider:** -Which data formats/types of data do you have? -Do the file formats and software enable easy access to the data, such as non-proprietary software and files?

	These can be files containing "survey data, census data, machine readable text, administrative records, textual data, et al." The National Endowment for the Humanities provides examples of humanities data such as, "citations, software code, algorithms, digital tools, documentation, databases, geospatial coordinates, text corpora, manuscripts, scraps of papyri, artifacts, images."	Click here for a proprietary file and software check. **E.g.** 100 txt files that include the corpora of Latin literature spanning from 65 BC-321 CE **E.g.** Annotations in jsonl; Maps; Plain text annotations[1]
Publication Date	The system automatically generates the publication date when data are published on TDR.	This date field is similar to another automatically generated TDR field, **Deposit Date**. The deposit date is generated when the data are uploaded to TDR, and the publication date is generated when the user selects to publish the data.
Production Date*	Date when the data collection or other materials were produced (not distributed, deposited, published or archived).	Date should be expressed in ISO format (YYYY-MM-DD). Can provide full date, year and month, or year. **E.g**. 2016-01-30 **E.g**. 2016
Author	List anyone who helped to create the dataset (who may or may not be an author of the data paper). Please include primary contact name and email if available.	Please use a Surname, Firstname format. TDR will prompt for optional fields like ORCID ID, roles, affiliations and primary contact. **E.g.** Jemisin, N.K. Molina-Gavilán, Yolanda
Language	Language of the dataset	**E.g.** English Spanish

[1] Palacios dataset example: https://dataverse.tdl.org/dataset.xhtml?persistentId=doi:10.18738/T8/L2SJQT

Vocabulary	For the specification of the keyword controlled vocabulary in use.	Abbreviated name of vocabulary **E.g.** LCSH ULAN
Dataset Sources	Please list the origins of the data in any consistent citation format. Include persistent identifiers if available. Sources can be books, articles, serials, interviews, government documents or machine-readable data files that served as the sources of the data collection.	**E.g.** *The Metropolitan Museum of Art Open Access CSV*. (2016). metmuseum. https://github.com/metmuseum/openaccess Hutchins, A. (2018, April 12). *Interview with Amanda Hutchins* (A. Casarez & X. Li, Interviewers) [Interview].
License	The open license under which the data have been deposited. You can specify terms of access in the TDR Terms tab. You can specify how users access your data if you have files that are restricted and enable a Guestbook for your dataset to track who is using your data and for what purposes. See more here.	**Questions to consider:** -Facts cannot be copyrighted, so first consider: Is there any copyright in your data set? (ex. Copyrighted artwork, correspondence, graphs or figures) -Who owns the copyright of your data, keeping in mind that multiple entities can be joint owners? CC0 is default. If you don't assign CC0, then you can pick a custom terms of use. You could select a different CC license, but that is not built into the TDR system. You could pick an alternative license with the Creative Commons license picker. You could also pick a license for a piece of software or code you are working on by using the choose a license picker. **E.g.** CC-BY 4.0
Keywords	These keywords or tags should topically describe the data. These are useful for classification and retrieval purposes. A controlled vocabulary can also be	**Questions to consider:** -What tags would be most useful to the other users to search and find your data? Consider using the Library of Congress Subject Headings (LCSH) located here. **E.g.** prison abolition, history, humanities

	employed.	

***-** indicates a required field in TDR

Consistency & Context:

- Empty cells
  - Use a consistent code for data which are unavailable, such as 'NA' used in R.
- Dates
  - Within the dataset, use a regular format for dates, such as YYYY-MM-DD.
  - If there is important information related to dates, such as when the data collection began and ended or when the dataset was cleaned, please include these in the description to improve the dataset's discoverability.
- Methods & Tools:
  - Recommended: In a README file, list any methods and tools used to clean, augment or transform the data.
    - Consider how the data collected and/or what were the methods used to create them
  - Consider using this template: **method | approach | algorithm | tool**
    - Any or all four concepts can be described.
    - **method | approach | algorithm | tool**
      - text analysis | topic modeling | latent dirichlet allocation |MALLET
    - **method | approach | tool**
      - augmentation, geocoded location data, TGN
    - **method | tool**
      - data cleaning | OpenRefine
  - Definitions:
    - **Method** is the overarching theoretical process
    - **Approach** is the specific strategic practice of the method.
    - **Algorithm** can reflect a specific machine learning algorithm used
    - **Tool** describes the program, software, vocabulary that supports and enacts the method, approach and algorithm.
- Terminology
  - It is important to be consistent. Terminology should be used invariably.
  - Is your structured data self-explanatory in terms of variable names, codes and abbreviations used?
  - Consider developing a data dictionary which may document:
    - A list of all the column names used in the data spreadsheet
    - A description of the purpose and the contents of these different columns.
    - If applicable, give an indication of the units of measurement.

- If applicable, describe the measures that have been taken to ensure the correctness and the consistency of the data
- Explain abbreviations or notational conventions that have been used in the dataset.