

Post45 Data Collective

Peer Review Criteria

The Post45 Data Collective peer reviews and houses post-1945 literary data on an open-access website designed, hosted, and maintained by Emory University's Center for Digital Scholarship. Here we provide the criteria by which submitted datasets will be evaluated. Submissions must be accompanied by a 800-1000 word paper that addresses the following questions, as appropriate.

Relevance and Reuse

How is the data relevant to post-1945 scholarship? Who might it be useful for? What could it be used for? Please suggest at least three specific uses.

For what purpose was the dataset created? Was there a gap that needed to be filled? Has the data been used already? Does similar or overlapping data exist publicly? If so, please describe.

Description

What does the data describe? Are all instances included or a selection? If selected, what principles were used to justify inclusions and exclusions?

If your dataset uses categorical variables or other labels or fields that you have created, explain how they were constructed. Should the user be aware of any categories or fields that condense or erase information?

Is any information missing? If so, please provide a description, explaining why this information is missing (e.g. because it was unavailable). Are there any errors, sources of noise, or redundancies? If so, please describe.

What is the file type and size of the data?

Collection and Creation

How was the data acquired or created? What mechanisms or procedures were used to collect it (e.g. hardware apparatus, human curation, software, API)?

If the data was hand-curated, what organizational heuristic was adopted, and why? What aspects of the data are products of the researcher's judgment or interpretation, and which aspects were inherited? What are the implications of these decisions?

Who was involved in the data collection process (e.g. students, crowdworkers, contractors) and how were they compensated? Over what timeframe was the data collected?

Was any cleaning of the data done (e.g. removal of instances, processing of missing values)? Was the "raw" data saved in addition to the cleaned (e.g. to support unanticipated future uses)?

Provide sufficient detail such that readers understand how the dataset was created, and would within reason be able to recreate it.

Ethics

What possible negative impacts or harms might result from the publication of your data?

Does the dataset contain data that might be considered confidential (e.g. data that includes the content of individuals' non-public communications)? If so, please describe.

Does the dataset contain data that might be considered sensitive (e.g. data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please describe.

Were any ethical review processes conducted (e.g. by an institutional review board)? If so, please describe these review processes, including the outcomes, as well as a link or other access point to supporting documentation.

Format

The Collective aims to maximize interoperability. To that end, we have strict requirements for the format of submitted data if it can be merged with extant data. For example, data oriented around book titles must use columns that match those used by HathiTrust. If it is a new category of data, the Collective will work with submissions toward creating exemplary standards.

For further help, consider the following resources:

- [Format your data](#) from the UK Data Service
- [Sustainability of Digital Formats](#) from the Library of Congress

Versioning

Will the data be updated (e.g. to correct errors, add new instances, delete instances)? If so, please describe how often and by whom.

Bibliography

Provide a list of sources consulted or drawn from to produce the dataset.

Licensing

If applicable, the data must be deposited under an open license that permits unrestricted access (e.g. CC0, CC-BY).

The language for these criteria was drawn from Katherine Bode, Jennifer Doty, Lauren F. Klein, Melanie Walsh, Cultural Analytics, Journal of Open Humanities Data, and "Datasheets for Datasets" by Timnit Gebru et. al.