Harvard Library Open Bibliographic Metadata in MARCXML[1]

v.1.0 (2 February 2023)

MOTIVATION

1. For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

   > The dataset was created to aid in the exploration of bibliographic records from Harvard Library's primary catalog, to further understanding of collections, publishing history, bibliographic descriptive practices, and other areas of research.

2. Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

   > The dataset was generated by Library Technology Services, Harvard University Information Technology on behalf of the Harvard Library, Harvard University, from records in the Library's Alma library information management system.

3. Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

   > Harvard Library Information & Technical Services expands and maintains the bibliographic records that comprise this dataset as part of their ongoing business processes, which are supported financially by Harvard University.

4. Any other comments?

---

[1] *Datasheet template drawn from the Appendix of Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. Commun. ACM 64, 12 (December 2021), 86–92. DOI:https://doi.org/10.1145/3458723.*

None.

COMPOSITION

5. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

   Each instance in the dataset is a record describing an intellectual entity held in the Harvard Library collections. Those entities may be books, journals, scores, databases, sound recordings, films and video, images, maps, archival collections, objects, or other materials. Each instance also contains one or more sections of "holdings" information about the specific items in the collection, such as the library responsible for managing the material, the physical or online location of the material, call numbers, volumes and issues held, and restrictions on access.

6. How many instances are there in total (of each type, if appropriate)?

   This information is not currently available, but will be generated for future versions of this dataset.

7. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

   The dataset includes all active (i.e., not suppressed or deleted) bibliographic records that have one or more active holdings records from the Harvard Library's Alma library management system.

8. What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is a record consisting of UCS/Unicode XML text. The content was converted from the MARC Bibliographic format, the dominant standard for communicating library catalog records and associated information, into its XML expression, MARCXML.

The dataset is split across multiple XML files due to size limitations. Each XML file except the last contains 200,000 <records> within a <collection> wrapper element. The <collection> element itself has no significance; the records from all the files together form a single dataset.

9. Is there a label or target associated with each instance? If so, please provide a description.

   Each instance is identified with a number (MMSID) that is unique within the dataset. While there is no label, each instance has one or more MARCXML tags containing a transcribed or cataloger-supplied title, sometimes combined with other identifying information (in a datafield with the tag attribute value "245"), that can serve as a human-readable label. (See Marc 21 Format for Bibliographic Data for more information about specific tags.)

10. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

    The data elements included vary from record to record, depending on the kind of materials being described and the level of cataloging. All records contain a small subset of required fields:

    - <leader> element: positionally-defined fixed length coded values that define parameters for the processing of the record
    - <controlfield> elements with the following tag attribute values:
        - 001: record control number [identifier]

        - 008: positionally-defined fixed-length coded values characterize the material to facilitate searching, faceting, reporting and display
    - <datafield> elements with the following tag attribute value
        - 245 (title)

11. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

    All instances are self-contained bibliographic records. Most records do not express any relationships to other records. A small proportion of records do express relationships to other bibliographic records, but these are not the norm, and the records they reference may or may not be part of Harvard's corpus of records.

12. Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

    While we can offer no specific points along which to split the data, creating subsets of records based on the format of the intellectual entity being described may be a useful approach. Each record contains a leader field whose 6th and 7th bytes contain coded values indicating the broad category to which the resource belongs (books, maps, visual materials, etc.) and upon which descriptive practices frequently depend.

    Using dates as a basis for data splits may also be helpful. Implementations of changing community standards and local practices are not usually uniform across all records from a given time period, but creating date-based subsets may be helpful nonetheless.

13. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

    The dataset contains a compilation of bibliographic data that has been created by a range of sources, both the cooperative cataloging community of libraries and agents acting on their behalf. The creation dates of the dataset cover a time period of evolving cataloging rules and punctuation conventions. These are among the factors that result in variability in the data that could look like or result in errors in data quality or completeness when querying the data.

    Occasional human error as well as data transformation and enrichment errors will also be found. Redundancies in the dataset may also be found when duplicate records have not been identified and resolved; however, due to the richness of the collections represented, multiple entries for the same or similar works are present and not redundancies, but rather are representatives of different editions, printings or manifestations of works that warrant their own entries.

14. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

    The records in the dataset are self-contained. The records may describe and contain links to digital content, but that digital content cannot be considered to be part of the dataset. Harvard Library has a strong commitment to the digital content that it holds and manages, and those links and content are designed to be as persistent and stable as possible. Records may also describe and

link to digital content that is not managed by Harvard Libraries. In both cases, there is no absolute guarantee that the external resources will exist or remain constant over time.

External resources may be restricted to the Harvard community or a subset thereof. If so, a user following the link will be requested for credentials to obtain access.

15. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' nonpublic communications)? If so, please provide a description.

    No.

16. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

    The dataset potentially contains controlled vocabulary terms and other textual descriptions that may be offensive or insulting. Subject headings are drawn from a global, interconnected dictionary used by librarians to create a consistent experience for researchers, and other information comes from a variety of sources, including the author, publisher, and librarians both at and beyond Harvard. Offensive terminology can be (and has been) reported by library users, and we continue to pursue both systemic changes and local remediation.

17. Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

    No.

18. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

While the instances in the dataset do not represent persons or describe them directly, many records contain one or more relationship to a person who was involved in the creation of the material being described, including authors and other creators, as well as persons who are the subjects of the content. This typically includes a name, frequently supplemented by dates of birth, death, or activity, and other information intended to uniquely identify a natural person and differentiate that person from others with similar names.

19. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race 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 provide a description.

    No. The data was created based on publicly-available information and information volunteered by publishers and authors.

20. Any other comments?

    None.

COLLECTION PROCESS

21. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

    The basic data within each instance will have been created by Harvard Library staff or by staff at other institutions, organizations, or companies, and it will have been augmented, corrected, and/or maintained by Harvard Library staff. The data was formulated

according to national and international cataloging rules and standards, which have changed over time. The values in some data elements such as title, author, and publisher, are typically transcribed from the materials in hand or imputed from other sources, while others, including subject terms and classification numbers, are assigned based on analysis of the content or other characteristics of the material.

22. What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

    Human manual curation and interpretation, recorded primarily through manual data entry in various library software applications, aided and supplemented by additional software, such as user macros and automated scripts as well as batch record loading.

23. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

    Not applicable. The dataset is not a sample from a larger set.

24. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

    The information was collected by library staff including (catalogers, acquisitions staff, and others) as well as graduate and undergraduate student workers, whose compensation varies according to their roles. Many records contain data collected by staff in similar roles at other libraries, as well as staff from commercial vendors.

25. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g.,

recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

> The records in the dataset have been put into machine-readable form over a period of approximately 50 years, but some of the data existed in printed or handwritten form up to centuries prior.

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

> Not applicable.

27. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

> Not applicable.

28. Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

> Not applicable.

29. Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

> Not applicable.

30. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

> Not applicable.

31. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

    No.

32. Any other comments?

    None.

PREPROCESSING/CLEANING/LABELING

33. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

    Instances that represent items that may no longer exist in the library's collection were removed. Internal administrative data for identifying individual catalogers was omitted.

34. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

    No.

35. Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

    No.

36. Any other comments?

USES

37. Has the dataset been used for any tasks already? If so, please provide a description.

No.

38. Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

    No.

39. What (other) tasks could the dataset be used for?

    Records in the dataset could be reused to supply descriptions for the materials (books, journals, films, and so forth) in other catalogs or transformed for use in other applications. Similar datasets have been used, for example, as training sets for automated subject heading assignment.

40. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

    Records in HOLLIS catalogs may contain language that is overtly or implicitly harmful, outdated, or biased. We acknowledge that the act of collecting materials is not neutral, and the work of describing library materials is influenced by inherent personal, institutional, and societal biases. We caution potential users against drawing conclusions from the presence or absence of materials that reach beyond changing collection practices and priorities specific to Harvard Library.

41. Are there tasks for which the dataset should not be used? Ifso, please provide a description.

    Not to our knowledge. This document will be updated if specific issues with reuse come to our attention.

42. Any other comments?

None.

DISTRIBUTION

43. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

    The dataset was initially created for and shared with the POD (Platform for Open Metadata) Aggregator <https://pod.stanford.edu/>.

44. How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

    The dataset will be distributed in a collection on Harvard Dataverse <https://dataverse.harvard.edu/>, a free data repository on the web. The dataset has a DOI: <https://doi.org/10.7910/DVN/I8L0ZZ>

45. When will the dataset be distributed?

    The initial version of the dataset will be distributed in December 2022. New versions of the dataset will be distributed on a recurring schedule, yet to be determined.

46. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

    The dataset will be distributed under a CC0 license. No fees will be associated with these restrictions.

    Although Harvard does not impose any legally binding conditions on access to the Metadata, Harvard requests that you act in accordance with the following Community Norms of the Harvard Library with respect to the Metadata:

First, Harvard requests that the Harvard Library and OCLC Online Computer Library Center, Inc. ("OCLC") be given attribution as a source of the Metadata, to the extent it is technologically feasible to do so.

Second, Harvard requests that you make the Metadata and any improvements thereto freely available on the same terms as Harvard has done, i.e., without claiming any legal right in, or imposing any legally binding conditions on access to, the Metadata or your improvements, and with a request to act in accordance with these Community Norms.

Third, with respect to Metadata consisting of or contained in records Harvard has obtained from the OCLC WorldCat database, Harvard requests that you respect and act in accordance with the community norms set forth in the WorldCat Rights and Responsibilities for the OCLC Cooperative. Use of metadata from the WorldCat database for study and research is consistent with those norms, but if you plan to use such Metadata for other purposes, whether or not you are an OCLC member, we ask that you review and comply with those norms.

47. Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

    No.

48. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

MAINTENANCE

49. Who will be supporting/hosting/maintaining the dataset?

Harvard Library Metadata Management & Project Strategy and Library Technology Services will support the dataset.

50. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

    Questions about the dataset can be directed to Metadata Management & Project Strategy, Information and Technical Services, Harvard Library, Harvard University. Please use the "contact owner" button on the Dataverse collection page.

51. Is there an erratum? If so, please provide a link or other access point.

    No.

52. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom,and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

    The underlying database from which this dataset is derived is a working system for the management of materials in Harvard Library collections. As a result, records can be added, changed, or deleted. The intent is to reissue this dataset in the future, reflecting the state of the data at the point of reissue.

53. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

    Not applicable.

54. Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

    Older versions will be retained in the Dataverse collection.

55. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

    Not applicable.

56. Any other comments?

    None.