LIBRARY OF CONGRESS COLLECTIONS POLICY STATEMENTS SUPPLEMENTARY GUIDELINES

Digital Datasets

Contents

Preface

Datasets are widely recognized as a critical component of research that needs to be preserved and shared alongside other forms of research output. For example, a 2013 Executive Office of the President memo, "Increasing Access to the Results of Federally Funded Scientific Research," directed agencies with an annual research and development budget over $100 million to develop a public access plan for disseminating the results of their research. Several other funding entities now provide grants to researchers with the stipulation that research reporting and associated data be made openly available. This has led to the creation of significant technological infrastructure and policy for preserving and providing access to datasets and a rapid increase in computationally-intensive research involving both quantitative and qualitative data across all domains. Indeed, several types of digital repository programs have been established to serve the needs of varied research communities, including university institutional repositories that store and provide access to affiliated research and domain-specific repositories that are geared for specialized user bases across multiple institutions. The Library is dedicated to supporting emerging styles of research by focusing on the selective acquisition of datasets that are 1) within scope under relevant Collections Policy Statements and Supplementary Guidelines, and 2) at risk of loss in not being under the stewardship of an institution dedicated to long-term data management.

I. Scope

For the purpose of this policy, the Library defines a dataset as collections of similar or related data that are usually assembled as a matter of record or research.[1] This document further defines datasets as digital content that consists of records stored in digital form, including text, numbers, images, video, audio, software, algorithms and models. The following guidelines are broadly focused on datasets acquired as part of the Library's digital collections, wherein the Library bears responsibility for storage and access to content. Datasets may be acquired on external media or through network transfer. This

[1] Definition from the Subject Record for “Data Sets” in Library of Congress Subject Headings.

document is intended to support the acquisition of selected datasets and the development of practices and policies that facilitate computationally-driven research of Library collections.

Any data resources published in print or as an e-book, such as the *Statistical Abstract of the United States*, unpublished print lists, or any other collection of organized data in analog form are out of scope for this document. The *Open Digital Content* Supplementary Guidelines should also be consulted when considering open e-books or e-serials, or for higher-level parameters for what the Library considers open content. The *Web Archiving* Supplementary Guidelines should be consulted when considering collecting websites related to datasets. Additionally, this document does not concern databases or collected datasets stored and made accessible exclusively through online vendor platforms, unless that platform is managed solely by the Library. Such resources are considered databases and covered by the *Electronic Resources* Supplementary Guidelines. Data visualizations may be collected as supplementary materials for a dataset acquisition or, in certain cases, as unique collection items; however, data visualizations are not a suitable replacement for a dataset and are not the focus of this document.

Datasets produced by the Library are not automatically considered as collection items and as such do not fall under this guidance; instead, see the *Library of Congress Publications and Other Content* Supplementary Guidelines.

II. Background

Historically, the Library collected datasets on a highly selective basis for some subject areas and not necessarily with the goal of universal research level coverage. Subject areas where the Library has traditionally collected datasets are demography, geography, business, economics and science. Much of the data produced in these areas have special collecting considerations and as such they are mentioned in other Collections Policy Statements. For example, the Collections Policy Statement for *Digital Geospatial Materials* provides information on the collecting of geographic information systems (GIS) data; *Local History* and *Genealogy* reference collecting local census information; *Earth Sciences, Environmental Sciences,* and *Science -- General* reference research data collecting; *Government Publications -- United States* mentions collecting of government statistical publications; and *Business and Economics* mentions the Economic Census and business statistics as important data collections. It should be noted that many of the Library's previously acquired datasets have been on external media (e.g., CD-ROM) and/or have been harvested and archived as part of the web archives.

Data production and computationally-intensive research can involve both quantitative and qualitative data, and are not restricted to any single domain. In scientific fields such as astronomy, meteorology, and genomics, data may originate from information gathered by networks of sensors and instruments, or through various forms of experimentation. Disciplines within the social sciences may generate and work with more qualitative types of data, including questionnaire responses, interviews, and texts. Researchers in the humanities also show a growing interest in large-scale cultural analysis of datasets that uses digital humanities methodologies, such as data mining and non-consumptive reading.

III. Diverse and Inclusive Collecting Statement

As the nation's de facto national library, the Library of Congress strives to build an expansive, yet selective, collection that records the creativity of the United States and is reflective of the nation's diversity and complexity. The Library's mandate is to have collections that are inclusive and representative of a diversity of creators and ideas. A priority includes acquiring material of underrepresented perspectives and voices in the Library's collections to ensure diverse authorship, points of view, cultural identities, and other historical or cultural factors. The Library also seeks to build a research collection that comprises a globally representative sample of international materials that are diverse in voice and perspective, relative to their places of origin, further supporting the Library's mission to sustain and preserve a universal collection of knowledge and creativity for Congress and future generations.

Diverse collecting is mentioned within many of the Library's Collections Policy Statements. In addition, the Library has adopted several specific collection policies in an effort to ensure it is building an inclusive and representative collection. For more information, see the Library's Collections Policy Statements on *Ethnic Materials*, *LGBTQIA+ Studies*, *Women's and Gender Studies*, *Independently Published and Self-Published Textual Materials*, *Materials Relating to Indigenous Peoples of the United States, Canada, and Mexico*, and *Countries and Regions with Acquisitions Challenges*.

IV. Current Guidelines

The Library of Congress acquires datasets and the associated content necessary to facilitate the interpretation of the content, such as documentation, scripts, and visualizations. Recommending Officers are encouraged to consult both existing Collections Policy Statements for subject-specific collecting guidance and these supplementary guidelines for format-specific considerations.

Note that some materials already held by the Library may later be made available as datasets, such as a text corpus consisting of several books. These datasets should be considered as new acquisitions with research functionality that goes beyond duplicating the original publications.

The following guidelines are broadly focused on datasets acquired as part of the Library's permanent collection. If a dataset is available only on a database platform, the suitability of the platform should be evaluated following the *Electronic Resources Supplementary Guidelines*. Additionally, the dataset itself should be reviewed following the guidelines established in this document with special consideration given to the stability and scope of the dataset and the ability to export the data. Purchased (or perpetual access) datasets are preferred. If a dataset is open digital content as defined by the *Open Digital Content* Supplementary Guidelines Scope section, those guidelines should be consulted alongside this document. For acquiring supplemental materials directly related to datasets, such as data dictionaries or readme files, direct download or network transfer are the appropriate methods of acquisition rather than web archiving. For acquiring related websites, web archiving is the preferable method, and the *Web Archiving* Supplementary Guidelines should be consulted.
In general, selective collecting of datasets (rather than comprehensive), in keeping with the Library's

Digital Collections Strategy and other loc.gov collections of material such as e-books and e-serials, is encouraged. Special consideration should be given to the factors listed below when recommending a dataset for acquisition. Staff are encouraged to consult with the Digital Collections Management & Services Division and the Collection Development Office when questions on these factors arise.

1. Subject - Does the subject of the dataset fall within the Library's collecting scope as indicated in the relevant Collections Policy Statement(s)?
2. Value - Does the dataset have enduring high value to Congress and the American people, thus making it worthy of long-term preservation and access? Is the content unique and/or scholarly? Is it at risk of loss?
3. Format - Is the data in a format that meets the specifications for Preferred or Acceptable in the *Recommended Formats Statement*?
4. Extent - Does the estimated number of files and size of the content fit within limitations of what the Library can present on loc.gov? Are there previous versions to acquire?
5. Access Rights - Can the Library provide meaningful access and usability to users, both onsite and off? Is the dataset open or rights-restricted?
6. Data Confidentiality - Does the dataset have any potential disclosure risk (the degree of risk that a data record from a study could be linked to a specific person or organization[2])?
7. License - Is the data being offered directly by the creator? Is the creator of the data and/or the entity authorized to grant permissions? Is the content openly available?
8. Frequency - Is this a one-time acquisition? If not, how often should the Library acquire new versions of the dataset to maintain currency of its information?
9. Documentation and Supplementary Materials - Is the dataset accompanied by any files that will help a user interpret the content (e.g., data dictionaries, codebooks, README files, code, websites, visualizations, related publications)?

V. Collecting Policy

The Library of Congress will selectively acquire, preserve, and make accessible datasets and their associated content for use by the U.S. Congress, staff, researchers, and the general public. The Library selects datasets for its permanent collection which rank high on the following list of criteria: usefulness in serving the current or future informational needs of Congress and researchers, unique information provided, scholarly content, currency of the information, and risk of loss (due to not being under the stewardship of an institution dedicated to long-term data management, for example, U.S. Geological Survey or the Census Bureau). The Library will define the attributes for selection, preserve the content in its digital repository, and provide appropriate access to the acquired content. The recommendation of works for the Library's permanent digital collections should be based on the subject and extent of the dataset.

Datasets produced by the Library are not automatically considered as collection items and as such do not fall under this guidance. Examples include the MARC Distribution Service data, derivative datasets

[2] https://www.icpsr.umich.edu/web/pages/datamanagement/confidentiality/

created for in-house analysis and/or crowdsourcing initiatives, datasets compiled by the Congressional Research Service, and scientific research data generated by the Preservation Research & Testing Division. For guidance about datasets produced by the Library, see the *Library of Congress Publications and Other Content* Supplementary Guidelines.

The growth and maturation of data-driven research across all domains will naturally require the Library to periodically re-evaluate the best methods for selecting works to archive.

Revised September 2024.