Data Sets – Interim Guidance

Contents

Preface

This document provides interim guidance to Library of Congress staff regarding the acquisition of data sets. It is anticipated that the Library's approach to data sets collecting will evolve and become more fixed over the coming few years. This document will need to be updated on a frequent basis as the data sets collecting program matures.

I. Definitions

"Big data" refers to extremely large collections of often heterogeneous data that pose significant management and manipulation challenges, particularly in real time search of the data.

"Database" refers to an organized collection of structured records under the control of a software management system.

"Data mining" is the process of using specialized software and other computer tools to generate new information (typically patterns, trends and relationships) from within large collections of data.

"Data set" is set is a collection of records, presented in a digital or non-digital format. For the purpose of this document, a data set is considered to be a type of electronic resource that consists of machine-readable data.

"**Electronic resource**" is defined as any work encoded and made available for access through the use of a computer. It includes data available by (1) remote access and (2) direct access (fixed media).

II. Background

In its simplest form, a data set is simply a collection of records, presented in a digital or non-digital format. The scope of this document is to consider data sets as a type of electronic resource that consists of machine-readable data. Individual records may be stored in any digital form, including text, numbers,

images, video, audio, software, algorithms and models. Data sets are often the product of scientific methods which produce massive quantitative research results. They may originate, for example, from information gathered by networks of sensors and instruments or through various forms of experimentation. Data sets may also be created outside of the realm of scientific research, for instance statistical sets that provide demographic or consumer information collected by various methods. Thus, data sets can exist for any field of inquiry or study.

Data sets play a huge role in the digital information environment. They are the foundation of "big data," which refers to extremely large and usually heterogeneous collections of data that require processing solutions more powerful than those available using standard computer processing applications. Data sets are also integral to databases, which organize data and provide real time query access via database management systems. Data sets are the object of data mining and data warehousing, which develop new information from within collections of data through the use of specialized software.

Data sets are commonly created during research and support and/or accompany published work. Some funding entities, including the United States Government, provide grants to researchers with the stipulation that the research reporting and associated data be made openly available. Thus, research libraries have created processes and infrastructure for preserving and making data sets available. The Library of Congress does not directly support such research and has not yet developed a standard process for identifying, acquiring, processing, and making available data sets. This statement is an initial step in that direction.

III. Current guidelines

Existing Library of Congress subject Collections Policy Statements provide collecting guidance across formats and should be followed for data sets, too, when technically possible. However, the collecting of data sets is currently limited to a standard workflow for acquiring, processing and making available geospatial data sets, which are being acquired by the Library to support the Geospatial Hosting Environment (in development). Collecting guidance for that content can be found in the Digital Geospatial Materials Collections Policy Statement. Standard workflows for other categories of data sets have not yet been developed.

The Library also currently provides access to data sets via its purchased and subscription Electronic Resources program. These resources, though, are generally not pure data sets but are actually databases. If a data set is available only on a database platform, the suitability of the platform should be evaluated following the Electronic Resources Supplementary Guidelines. Additionally, the data set itself should be reviewed following the guidelines established in this document with special consideration given to the stability and scope of the data set and the ability to export data. Purchased (or perpetual access) data sets are preferred.

It is possible to acquire data sets on tangible media, such as thumb drives or hard drives. Any Recommending Officer considering such an item should consult with the Collection Development Office before submitting a recommendation.

As described above, the current collecting of data sets is limited by technical factors. In cases where an acquisition is technically possible, the following factors should be considered.

1) Subject – Does the subject of the data set fall within the Library's collecting scope?

2) Value – Does the data set have enduring high value from a scientific or historical perspective, thus making it worthy of long-term preservation and access?

3) Documentation and metadata – Is the data set accompanied by adequate technical and descriptive information?

4) Format – The data set must be in a format that the Library supports. The Library's Recommended Formats Statement provides specific technical information about data sets.

5) Access – Does the Library have a way to provide access to the data set upon receipt? Is immediate access to the data set necessary? Has the Library defined a path to future access for this type of data set?

Any questions about the appropriateness of a specific data set acquisition being considered should be forwarded to the Collection Development Office.

IV. Future guidance

This document will be updated as the Library's data sets collecting program is further defined and related infrastructure and workflows are developed.

August 2017