Notes on Operations

Maximizing the Discovery of Data Sets in the Yale University Library Catalog

Rowena Griem, Yukari Sugiyama, and Tachtorn Meier

In response to the desire to include data set holdings in the Yale University Library (YUL) catalog, the Dataset Cataloging Task Force was formed in spring 2019 to assess the existing cataloging practices and current integrated library system environment. This paper describes the process of developing cataloging guidelines in the absence of authoritative resources while implementing best practices for cataloging data sets with the goal of optimizing the discoverability and accessibility of data sets in the online library catalog. The authors recommend the establishment of a national group to discuss, establish, and document national guidelines for cataloging data sets so that these increasingly important resources are treated in a consistent manner in institutional, consortial, and global catalogs.

With the growing importance of digital scholarship in academia, there has been a marked increase in the systematic acquisition of data sets by libraries. A data set is "a collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer."[1] Yale University Library (YUL) holds over 10,000 data sets ranging from statistical and geospatial data, to text and sound corpora, and image data sets. While most of these are remote data sets, some are available in direct access formats such as CD-ROMs and hard drives.

YUL has demonstrated its commitment to digital scholarship with the establishment of dual research centers for data analysis. The StatLab, now housed within the Marx Science & Social Science Library, works with data in the natural and social sciences, technology, engineering, and mathematics (STEM) fields. The Digital Humanities Lab (DHLab) was established in fall 2015 to probe the arts, humanities, and humanistic social sciences through technology. Thanks to support from Barbara and Richard Franke and the Goizueta Foundation, the DHLab has been transformed from a one-person operation to a fully staffed department with cutting-edge computing technology in a renovated space in a prime location in Sterling Memorial Library.

In 2018, Yale University published the *Report of the University Science Strategy Committee* with a plan to invest in the sciences at Yale by making it a top academic priority. The report highlighted data science as one of its top priority investment areas, noting "The confluence of the volume, speed, and availability of data is transforming information and knowledge production."[2] To support that investment, YUL anticipates increased use and, consequently, acquisition of data sets, escalating the accessions to a steady flow. It is essential to ensure that these emerging, complex, and evolving resources are easily discovered, identified, and accessed by members of the Yale community, including students, educators, and researchers, via the library catalog.

The authors were appointed to the newly formed Dataset Cataloging Task Force in April 2019. The group was charged with reviewing the current library

Rowena Griem (rowena.griem@yale.edu) is Catalog Librarian for E-Resources & Serials Management, Technical Services at Yale University Library. **Yukari Sugiyama** (yukari.sugiyama@yale.edu) is Librarian for Discovery & Metadata Assessment, Technical Services, Yale University Library. **Tachtorn Meier** (tachtorn.meier@yale.edu) is Catalog Librarian for Monographic Latin Script Receipt and Cataloging, Technical Services, Yale University Library.

Manuscript submitted June 6, 2021; returned to authors for minor revision July 12, 2021; revised manuscript submitted September 10, 2021; accepted for publication November 9, 2021.

The authors wish to thank their readers, Dominique Bourassa and Jeanette Norris, Yale University Library.

landscape and existing cataloging guidelines for data sets, analyzing the needs to integrate data sets into the general collection instead of creating silos, and developing best practices to ensure, optimize, and improve the discoverability and accessibility of data sets in YUL's discovery interface, Quicksearch. The focus was on commercial and open access data sets acquired and licensed by the library, not on research data generated by Yale affiliates. Since cataloging data sets was never addressed in a comprehensive way at YUL, they were not always readily identifiable or displayed in an effective or consistent way in the catalog. Additionally, some data sets require mediated access due to strict licensing requirements, necessitating a workflow for addressing access to them. While the authors' primary goal was to create documentation and tools for cataloging data sets, additional work was necessary to optimize the effectiveness of the bibliographic records created for data sets, such as proposing new subject and genre/form headings and modifying the Blacklight-based discovery interface. This paper describes the issues that arose, and the solutions, deliverables, and resulting enhanced discoverability of data sets in the YUL catalog.

Literature Review

The history of cataloging data sets, which broadly fall under computer-related materials, dates to the 1970s when microprocessors and microcomputers had yet to be developed. At the time, data was stored on punched cards, magnetic tapes, and other data storage products to be processed by machines. Under the 1978 *Anglo-American Cataloguing Rules* (AACR2), second edition, such data was categorized as "machine-readable data file (MRDF)" with the general material designation (GMD), a medium designator added to the title statement.[3] The term MRDF "embraces both the data stored in machine-readable form and the programs used to process that data."[4] As microcomputers became popular and libraries started adding computer-based media such as computer cartridges, computer cassettes, and computer reels, MRDF was renamed "computer file" in the AACR2 1988 revision.[5] Chapter 9 explained that these files are "encoded for manipulation by computer" and "comprise data and programs," and added, "Computer files may be stored on, or contained in, carriers available for direct access or by remote access."[6] With the advent of the Internet, the GMD term was changed to "electronic resource" in the AACR2 2001 amendments to encompass remote access electronic resources, in addition to direct access electronic resources.[7]

These revisions were accompanied by changes and additions to the cataloging rules. Although the revisions were necessary to keep up with the development of new formats and carriers, they also caused some complications. For example, Weiss argued: "Observation of OCLC record errors and problems suggests that transition periods or periods in which more than one standard is in use are the times when there is the greatest confusion among catalogers and the greatest inconsistency of cataloging for electronic resources."[8] Likewise, using video games as an example, de Groat showed how a physical description field was affected and altered by constant shifts of terminology, making "it difficult to collocate materials or provide a consistent search or limit strategy to find all like materials."[9] In 2013, Resource Description and Access (RDA) was fully adopted by the Library of Congress (LC), the National Library of Medicine, and the National Agricultural Library as the successor of AACR2, leading to significant changes in cataloging electronic resources. The GMD became obsolete. Content type, media type, and carrier type were introduced in its place and recorded in the MARC 336, 337, and 338 fields respectively. For data sets, RDA provides just two content types in section 6.9.1.3: "cartographic dataset" and "computer dataset."[10] Whereas "cartographic dataset" is distinctly designated for geospatial data sets, other types of data sets must be described using the less granular content type "computer dataset." Nonetheless, the RDA vocabulary encoding scheme for content type is one of the first terminologies that includes terms to describe data sets in cataloging. RDA also provides controlled terms for file type such as "audio file," "text file," "image file," and "data files." These terms can be used in the MARC 347 field, which was added to the MARC 21 Standard in 2011 to describe digital file characteristics.[11]

MARC-based cataloging of data sets is discussed in only a handful of papers, mostly within the context of geospatial data. Although it was written more than twenty years ago and in the AACR2 era, Welch and Williams's 1999 paper is still remarkably relevant and valuable for cataloging geospatial data. As is Larsgaard's "Cataloging Cartographic Materials on CD-ROMs." In both papers, however, the authors pointed out the limitation of existing subject terms to describe the physical carrier aspect of digital cartographic materials. To mitigate this shortfall, LC used uncontrolled subject headings in the MARC 653, such as "Maps–Digital," "Maps–Digital–Raster," "Maps–Digital–Vector."[12] According to Lage, this practice of using local headings was also employed by several academic libraries.[13] Examples of local vocabulary included "Geographic information systems data," "Geodatabases–Electronic resources," and "Digital spatial data."[14] Lage discusses a "critical need" to standardize subject access to Geographic Information System (GIS) data "through the creation of authorized subject, form, and genre headings."[15] In June 2010, LC announced its decision to separate genre/form headings from the Library of Congress Subject Headings (LCSH) and named this thesaurus

The *Library of Congress Genre/Form Terms for Library and Archival Materials* (LCGFT).[16] Today, LCGFT has some geospatial data-related terms such as "Geospatial data," "Raster data," and "Vector data," and a few data-related terms such as "Census data" and "Statistics."

Over the years, many other metadata schemas emerged to describe geospatial data, including Dublin Core, the Federal Geographic Data Committee (FGDC)'s Content Standard for Digital Geospatial Metadata (CSDGM), the International Organization for Standardization's Geospatial Metadata Standard (ISO 19115), XML, METS, and MODS.[17] Among them, the FGDC metadata standard is the most widely used schema as Executive Order 12906 in 1994 mandated that federal agencies use it.[18] Although more GIS data became available in the FGDC metadata standard, Reese reiterates "the traditional need for MARC bibliographic data will still exist within the library into the foreseeable future."[19] Reese also showed how building a crosswalk between FGDC and MARC or other schemas using eXtensible Stylesheet Language Transformations (XSLT) is complicated but possible and cost effective.[20]

Although there are still no established nationwide best practices for cataloging data sets using RDA, metadata elements useful for facilitating access to data sets were cited in the *Library of Congress Recommended Formats Statement, 2020–2021*.[21] It identified the recommended formats, technical characteristics, and associated metadata to ensure the preservation and long-term access of creative works. For metadata elements for data sets, it is recommended that one include title; creator; creation data; place of publication; publisher/producer/distributor; contact information; a list of software used to produce, render or compress the data; and character encoding whenever possible. Other elements such as language of work, other relevant identifiers, subject descriptors, and abstract were suggested if available. While these recommended metadata elements are for preservation purposes, rather than resource discovery, many are descriptive metadata. As more data sets are born digital and will require digital preservation efforts for future accessibility, the recommended metadata elements should be considered for inclusion in catalog records.

Cataloging practices for computer-based materials have been in flux, leading to a lot of confusion among catalogers and inconsistency in legacy records, jeopardizing the discoverability of those materials. Data sets are no exception. Various authors in the map cataloging community have published guides to help catalogers work with GIS data using AACR2. With the development of RDA and more data set-related terms being added to LCGFT, now seems to be a good time to develop new comprehensive cataloging guidelines for data sets.

Descriptive Cataloging for Data Sets

Unlike most other library materials, data sets are not always incorporated into library catalogs. Some institutions use different library platforms for data sets such as A–Z lists, LibGuides and institutional repositories, whereas others use data-specific repositories such as the New York University Data Catalog; the University of Maryland, Baltimore Data Catalog; and Harvard University's Dataverse Repository. In contrast, at YUL, at the request of project stakeholders (data librarians, DHLab staff, and Technical Services directors), the task force was charged with integrating data sets into the library catalog, making it a one stop shop for all library collections.

In the process of establishing best practices for cataloging data sets, the authors discovered that there do not appear to be detailed national guidelines to distinguish data sets from other types of electronic materials or to record data set-specific characteristics in MARC. Judging from an examination of bibliographic records in OCLC Connexion, it seems that catalogers have relied on their own interpretations of existing rules when cataloging data sets. A lack of clear rules leads to inconsistent cataloging within and across institutions, affecting the discoverability of these resources in library catalogs and OCLC WorldCat. Clear, comprehensive, universally accepted guidelines are crucial to ensure the consistent discoverability of data sets in institutional, consortial, and global catalogs.

Data sets are manifested in various content and data structures. The authors identified five broad types of data sets that each required separate cataloging documentation and templates:

1. Sound data sets, including the subset speech data sets: The resource is a corpus of digital sound recordings, including music, ambient sounds, such as nature sounds, or spoken language, such as speeches. Notable formats are FLAC, MP3, MP4, and WAV.
2. Geospatial data sets: The resource consists of data that identify the geographic location of an object in space according to a geographic coordinate system. Many data sets use the ESRI Shapefile format to be processed in GIS software.
3. Image data sets: The resource is a digital collection of still or moving images, such as graphic materials, photographs, illustrations, or video. Significant formats include JPEG, PNG, BMP, and TIFF.
4. Numeric data sets: The resource consists predominantly of statistical data, such as census or election data. Formats may include CSV, Excel, SAS, and SPSS.
5. Text data sets: The resource is a corpus of digital text derived from written sources, both published and unpublished, such as books, newspapers, periodicals,

documents, correspondence, and emails. Formats include, but are not limited to, TXT, DOC, XML, and DjVu.

Most sets held by YUL are remote data sets, although there are some, chiefly older titles, available via direct access formats such as CD-ROMs, DVD-ROMs, external hard drives, and USB flash drives. At YUL, if the licensing agreement allows, some of these direct formats are converted to locally hosted remote versions to make them more accessible. To address the variability of data sets, the authors identified the following key MARC fields that are unique to data sets in the bibliographic record.

Fixed Fields

There is no uniform Leader/06 (Type of record) code for data sets. Prior to the 1997 revision of the definition for code "m" (Computer file) in Leader/06, all data sets were coded as "m" (Computer file), since anything electronic was defined as a computer file.[22] Following that major revision, the definition of computer file is as follows (the emphasis is the authors'):

> m - Computer file
> Used for the following classes of electronic resources: computer software (including programs, games, fonts), numeric data, computer-oriented multimedia, online systems or services. For these classes of materials, if there is a significant aspect that causes it to fall into another Leader/06 category, the code for that significant aspect is used instead of code m (e.g., vector data that is cartographic is not coded as numeric but as cartographic). Other classes of electronic resources are coded for their most significant aspect (e.g. language material, graphic, cartographic material, sound, music, moving image). In case of doubt or if the most significant aspect cannot be determined, consider the item a computer file.[23]

By this definition, only numeric data should be coded "m" in Leader/06. Other types of data sets are coded based on "the significant aspects of their content, as opposed to their carrier."[24] Combination data sets, for example those including geospatial and numeric data, are coded according to the primary characteristic. Since the Leader and 008 fields are not repeatable, an 006 field for "Computer File" is added to reflect additional material characteristics for data sets that are not coded "m" in the Leader/06, ensuring that the resource is identified as an electronic resource in the catalog and OCLC.[25] The other key elements for data sets are the "Form of Item" (008/23 or 006/06), which is coded either "o: online" or "q: direct electronic," depending on the carrier of the data set, and "Type of File" (008/26 or 006/09) to bring out other characteristics of each type of data set. For example, "a: numeric data" for numeric data sets; "c: representational" for both still and moving image data sets, as well as geospatial data sets; "d: document" or "e: bibliographic data" for text data sets; and "h: sound" for sound data sets. The 007 field (Physical Description fixed field) is mandatory for anything electronic, so when the main item described in a record is a data set, the record must have a 007 field identifying the resource as electronic.[26]

Transcribed Variable Fields (2XX Fields)

Data sets often include little or no identifying information, such as title or publishing information. Consequently, describing data sets in the bibliographic record can be challenging. RDA instructions "1.7 Transcription" and "2.2 Sources of Information" provide guidance for MARC field elements that require transcribed information. In the bibliographic record, information regarding the title proper, edition statement, and publication statement are required transcription elements in RDA. The carrier of the content plays an important role in determining the source of information for electronic resources, including data sets. Data sets can be available via physical carrier (direct access) or over-the-network (remote access). According to RDA,

Table 1. Leader/06 for Types of Data Sets

Type of Data Set	Fixed Fields
Geospatial	Leader/06=e (Cartographic material) 006/00=m (Computer file) + 006/09=c (Representational) 007/00=c (Electronic resource)
Image	Leader/06=g (Projected medium) or k (Two-dimensional nonprojectable graphic) 006/00=m (Computer file) + 006/09=c (Representational) 007/00=c (Electronic resource)
Numeric	Leader/06=m (Computer file) 008/26=a (Numeric data) 007/00=c (Electronic resource)
Sound	Leader/06=i (Nonmusical sound recording) or j (Musical sound recording) 006/00=m (Computer file) + 006/09=h (Sound) 007/00=c (Electronic resource)
Text	Leader/06=a (Language) 006/00=m (Computer file) + 006/09=e (Bibliographic data) or d (Document) 007/00=c (Electronic resource)

the chief source of information for electronic resources, whether tangible or online, is the resource itself, namely "a textual source on the manifestation itself (e.g., a slide) or a label that is permanently printed on or affixed to the manifestation, excluding accompanying textual material or a container (e.g., a label on an audio CD or a model)."[27] That said, this source of information may not be available for data sets. For tangible resources, the title screen is the second choice for the preferred source of information, followed by the labels as the last preferred source of information. If the information cannot be ascertained from any of the preferred sources for tangible or online resources, "[give] preference to sources in which the information is formally presented."[28] This information can be found from the accompanying materials or on the publisher's website.

For data sets derived from previously published resources, for example databases or newspapers, the title and publication information can be based on the original publication with the word "dataset" appended to the title. When the title, or part of the title, is devised, a MARC 500 field should be added noting: "Title supplied by cataloger."

Physical Description Field, 33x Fields, and Digital File Characteristics (3XX Fields)

The physical description, whether it is direct access or remote, is included in the MARC field 300. The authors decided not to follow RDA's main instruction 3.3.1.3 to use the term "computer chip cartridge" from the list of carrier types to record tangible carriers such as USB flash drives or external hard drives.[29] Instead, the alternative instructions 3.4.1.3 were followed to "use a term in common usage (including a trade name, if applicable) to indicate the type of unit."[30] If desired, the number of files can be included in a parenthetical statement in the $a. Accompanying materials such as codebooks, manuals, maps, or CD-ROMs can be recorded in the MARC 300 field, subfield $e. If accompanying materials are available online, access to the accompanying material can be provided in a MARC 856 field with the second indicator "2" to indicate that it is a related resource, using the following format:

856 42 $3 Documentation $u [URL to accompanying material]

The MARC 33X fields are used to describe Content, Media, and Carrier Types. The 336 Content Type field is used in conjunction with Type of Record in the Leader/06 and reflects the form of the content of the resource. It is a core element in RDA. The Term and Code List for RDA Content Types lists two applicable codes: "cartographic dataset," which is expressly used for geospatial data sets, and "computer dataset," which can be used for all varieties of data sets.[31] The latter should be coupled with a second 336 field to reflect the specific content type, for example "still image," "audio," or "text," which allows image, sound, and text data sets to be mapped both to the data set format and the content type of the source material. These combinations of content types allow for expanded discoverability of data sets in the catalog, whether looking for all data sets or a specific type of data set. Interestingly, the new RDA, which is scheduled to be implemented by American libraries sometime after October 2022, allows for the extension of RDA content categories to

Table 2. Sample 3XX Fields

Type of Data Set	3XX Fields
Geospatial	300 \\ $a 1 USB flash drive
	336 \\ $a cartographic dataset $b crd $2 rdacontent
	336 \\ $a computer dataset $b cod $2 rdacontent
	337 \\ $a computer $b c $2 rdamedia
	338 \\ $a computer chip cartridge $b cb $2 rdaccarrier
	347 \\ $a data file $2 rdaft
	347 \\ $b shapefile
Image	300 \\ $a 1 online resource + $e documentation
	336 \\ $a computer dataset $b cod $2 rdacontent
	336 \\ $a still image $b sti $2 rdacontent
	337 \\ $a computer $b c $2 rdamedia
	338 \\ $a online resource $b cr $2 rdaccarrier
	347 \\ $a image file $2 rdaft
	347 \\ $b GIF
Numeric	300 \\ $a 1 computer disc ; $c 4 3/4 in.
	336 \\ $a computer dataset $b cod $2 rdacontent
	337 \\ $a computer $b c $2 rdamedia
	338 \\ $a other $b cd $2 rdaccarrier
	347 \\ $a data file $2 rdaft
	347 \\ $b CSV
Sound	300 \\ $a 1 external hard drive
	336 \\ $a computer dataset $b cod $2 rdacontent
	336 \\ $a sounds $b snd $2 rdacontent
	337 \\ $a computer $b c $2 rdamedia
	338 \\ $a other $b cz $2 rdaccarrier
	347 \\ $a audio file $2 rdaft
	347 \\ $b MP3
Text	300 \\ $a 1 online resource (approximately 6 million text files)
	336 \\ $a computer dataset $b cod $2 rdacontent
	336 \\ $a text $b txt $2 rdacontent
	337 \\ $a computer $b c $2 rdamedia
	338 \\ $a online resource $b cr $2 rdaccarrier
	347 \\ $a text file $2 rdaft
	347 \\ $b XML
	347 \\ $3 Compressed $c 62.60 GB
	347 \\ $3 Uncompressed $c 75.68 GB

accommodate the following attributes: form/genre, subject, purpose, or capture method. To add new terms to the list of RDA content types, such as image data sets, a formal proposal would need to be presented to the RDA Steering Committee. Values could also be defined locally as subvalues of "dataset." The authors elected not to introduce terminology locally, as doing so effectively would require ensuring all YUL catalogers consistently utilize the same terms, and could potentially create variations across the library system.

Additional 33X fields include the Media Type, which is recorded in the MARC 337 field, and reflects the type of device required to access the resource content. The media type "computer" is used for all types of electronic resources, including data sets. The MARC 338 field is used to record the carrier type, the format of the storage medium in which the content is recorded. Computer carrier terms such as "computer disc" and "online resources" are commonly used for electronic resources.

Digital file characteristics are not RDA core. Per LC/Program for Cooperative Cataloging Policy Statements, it is only a core element for cartographic materials. However, this information is valuable to fully describe the content of data sets and allows users to easily identify types of files and determine compatibility with their computer environment. Digital file characteristics such as the file type (audio, data, image, or text files), encoding format, and file size are recorded in the MARC 347 field, while other physical details, such as the number and arrangement of files are recorded in the MARC 300 Physical Description field and note field respectively. The size of compressed and/or uncompressed files (347 $c) has proven useful for data management of locally hosted data files. The authors followed OCLC's guidelines to "prefer use of a separate field for each unique term" to record the file type and encoding format information.[32]

Formatted Contents Note (505 Field)

If the data set contains data from discrete titles, for example newspapers or periodicals, an enhanced content note can be added to maximize discoverability. Cataloger's judgment may be applied to determine whether this is advisable by weighing the number of titles involved and the availability of the information versus the value added. The term "dataset" is added after each title in the contents note to ensure that the nature of the title is evident to patrons. For example, a data set collection featuring New York newspapers would be greatly enhanced with the following 505 field:

> 505 00 $t New York times dataset -- $t New York post dataset -- $t Wall Street journal dataset . . .

Restrictions on Access Note (506 Field)

The presence of a MARC 506 field informs the user when the data set has restrictions and/or requires some level of permission to access. If the resource requires mediated access, it is noted here and paired with a link in the 856 field to request assistance to access the restricted data sets.

Summary Note (520 Field)

A summary note is not an RDA core element, but this information is extremely useful for cataloging data sets. Information about the nature and scope of a resource can help users determine whether a data set is relevant to their research. It is advantageous to record crucial information in one place, using terminology that the patron can easily understand, even though some of this information may be found in a structured format elsewhere in the record. Ancillary information can also be included here, such as granularity (the size into which data fields are subdivided), the organization of the files, etc. Because of the potential usefulness of such details to researchers, it is important to remind selectors to provide catalogers with all available information about the data set (coverage dates, required software, and granularity, for example), so that the resources can be described effectively. It has proved invaluable to informally survey stakeholders working with Yale's data collections to uncover what data they find helpful. Including useful terms is crucial to take advantage of keyword searching, without trying to anticipate or predict how a researcher might use the data.

System Details Notes (538 Field)

The digital file type, encoding format, and file size can present significant challenges for cataloging data sets. Certain encoding formats may require special software or applications to access, manipulate, visualize, or analyze the data associated with the resource. For example, GIS mapping software can be used with GIS file formats such as Shapefile, while statistical analysis and visualization tools such as SPSS, R, or JMP, can be used with data file formats such as CSV or Excel. The authors chose to only record specialized methods of data set access or usage in the MARC 538 field, not common computer standards and peripherals, such as Adobe Acrobat, Excel, Internet Explorer, or the World Wide Web.

Action Note (583 Field)

For materials digitized or hosted locally, a formatted MARC 583 field is added to record details of the action, including the action taken, the date, the acting agent, the code used, and the institution. This field is added to mediated data sets

added to Storage at Yale, an institutional central storage service, prior to being moved to Preservica, an archiving and digital preservation platform. At YUL, this field is added to the holdings, rather than the bibliographic record.

Source of Description Note (588 Field)

The source of title is required for electronic resources, even if taken from the chief source of information. RDA 2.17.13.5 also calls for the creation of a note indicating the date the resource was viewed for remote resources, although this is not applicable when a cataloger needs to supply a title. Below are examples of Source of Description Notes for data sets.[33]

588 \\ $a Description based on print record.
588 \\ $a Description based on source database record.
588 \\ $a Title from homepage (viewed [date]).
588 \\ $a Title from file header (viewed [date]).
588 \\ $a Title from readme file (viewed [date]).

Data Set-Related Subject and Genre Headings

One particularly thorny issue that needed to be addressed was how to provide intellectual access to the materials using the LCGFT and LCSH controlled vocabularies. How should the cataloger effectively describe the resource and what it is about? Despite the complexities involved in this process, the authors believe that assigning detailed headings greatly enriches the catalog, ensuring discoverability of the data sets and linking them to related materials via linked data.

In the planning stages of the project, in late 2018, neither LCSH nor LCGFT included the umbrella term "Data sets" or its variant spelling "Datasets," so the authors began with those headings. Work began to propose them in the one-word form for three reasons: a Google search suggested that the single word form was significantly more common than the two-word form; it was consistent with the existing LCSH and LCGFT headings for "Databases"; and the single word form appears in the MARC 336 field as "Computer dataset "and "Cartographic dataset." However, the proposal for the LCGFT was accepted with the preferred form "Data sets," with the single-word form given as a cross reference. At the same time, LC created an LCSH with the two-word form as the preferred form. A proposal for the free-floating form subheading "$v Data sets" was rejected due to the complexities of linked data. The authors were advised instead to pair the newly established LCGFT for "Data sets" with additional the appropriate subject headings fields to provide satisfactory intellectual access to the resource.[34]

At the start of the project, the authors ran a report in YUL's integrated library system (ILS), extracting a sample of data set records to examine the headings assigned to them. Geospatial data sets were often assigned the LCSH "Geographic information systems "and the LCGFTs "Geodatabases," "Geospatial data," "Raster data," or "Vector data." Numeric data sets were typically assigned some combination of the subject format subdivisions "Census"; "Census, [date]"; "Statistics"; "Statistics, Medical"; "Statistics, Vital"; and the genre/form terms "Census data," "Demographic surveys," "Judicial statistics," "Statistics," or "Vital statistics." The vast majority of text data sets present in Yale's library catalog at the start of the project were published by the Linguistic Data Consortium and generally had bibliographic records available in OCLC. Over 60 percent of these use "$x Data processing" in a 6XX field, despite it being a topical subdivision, and the resources being cataloged not being about data processing, but rather being used for data processing.

An assessment of existing LCSHs identified potentially useful subdivisions: "$x Language" ("use as a topical subdivision under names of individual persons and corporate bodies, individual works entered under title, and under classes of persons and disciplines, individual wars, and types of newspapers") for text or speech data sets), "$x Sounds" ("use as a topical subdivision under individual organs and regions of the body and wars" for sound data sets), and "$v Maps" ("use as a form subdivision under names of countries, cities, etc., and individual corporate bodies, and under topical headings for individual maps or collections of maps on those subjects" for geospatial data sets).[35]

An analysis was conducted, comparing existing LCSHs with LCGFTs to determine whether relevant equivalent terms existed. Several topics of interest to Yale's collection were identified and proposed as new genre terms. For example, while "Corpora (Linguistics)," "Medical statistics," and "Biometry" existed in LCSH, there were no equivalent LCGFTs, so the authors successfully proposed the related genre/form terms: "Text corpora," "Speech corpora," "Medical statistics," and "Biostatistics." Proposals for the genre/form terms "Image data sets," "Spatial data sets," "Statistical data sets," and "Text data sets" were all declined in favor of pairing the LCGFT for "Data sets" with another LCGFT(s) for the type(s) of data.

While subject headings already existed for the generic "Data mining" and more specific headings (such as: "Association rule mining," "Contrast data mining," "Multimedia data mining," "Sequential pattern mining," and "Web usage mining"), the authors successfully proposed genre/form terms to describe types of data sets plus subject headings for additional types of data mining, useful when cataloging

materials about data mining, such as those in the DHLab's reference collection.

The following headings were created for the project, significantly enriching the controlled vocabularies:

- LCSHs (650 field):
  - Audio data mining
  - Data mining—Statistical methods
  - Image data mining
  - Spatial data mining
  - Text data mining

- LCGFTs (655 field):
  - Biostatistics
  - Data sets
  - Medical statistics
  - Sound corpora
  - Text corpora

The authors' next step was to provide guidelines on assigning 6XX fields so that the resources are treated consistently. The authors first analyzed Yale's collections and devised a blueprint:

- All data sets are assigned an LCGFT for *Data sets*, which allows patrons to simultaneously retrieve all data sets with a single search;
- Additional LCGFTs are then assigned to identify each of five broad categories (two of which have subsets): "Maps" for geospatial data sets, "Pictures" for still image data sets (or "Video recordings" for moving image data sets), "Statistics" for numeric data sets, "Sound corpora" for sound data sets (or "Speech corpora" for speech data sets), and "Text corpora" for text data sets, allowing patrons to readily retrieve all of a specific type of data sets;
- To achieve greater granularity, additional LCGFTs may be assigned to describe the original form of the data, for example "World maps" in addition to "Maps," "Aerial photographs" in addition to "Pictures" (or "Industrial films" in addition to "Video recordings"), "Death registers" in addition to "Statistics," "Radio programs" in addition to "Sound corpora" (or "Spoken word poetry" in addition to "Speech corpora)," and "Messages (Official communications)" in addition to "Text corpora";
- Finally, subject headings are added to describe the resource's topic without trying to predict what kind of patterns the researchers might plan to study using any given data set.

Table 3. Sample 6XX Fields

Type of Data Set	LCSH	LCGFT
All data sets		"Data sets"
Geospatial data sets	[Corporate body, Geographic location, or Topical heading] $v Maps	"Geospatial data" + Type(s) of GIS data, for example: Raster data, Vector data; Maps; and specific type(s) of map(s), such as Geological maps, etc.
Image data sets (fixed images)	Subject heading for subject of images	"Pictures" + Type(s) of images, for example: Cartoons (Humor), Illustrated works, Postcards, etc.
Image data sets (moving images)	Subject heading for subject of moving images	"Video recordings" + Type(s) of video, for example: Film clips, Motion pictures, etc.
Numeric data sets	[Class of person, Corporate body, Ethnic group, Geographic location, or Topical heading] $v Statistics	"Statistics" + Type(s) of statistics, for example: Biostatistics, Census data, Judicial statistics, Medical statistics, etc.
Sound data sets	[Animated films, Motion pictures, Radio broadcasting, Television broadcasting, Theaters, or Video games] $x Sound effects	"Sound corpora" + Type(s) of sound, for example: City sounds, Human sounds, Nature sounds, Sound effects recordings, etc.
Speech data sets	[Language] $x *Spoken* [Language] $z [Geographic location]; [Individual person, corporate body, or war; class of person or discipline; type of newspaper] $x Language	"Speech corpora" + Type(s) of speech, for example: Interviews, Oral histories, Speeches, etc.
Text data sets	[Language] $x Written [Language] $z [Geographic location]; [Individual person, corporate body, or war; class of person or discipline; type of newspaper] $x Language	"Text corpora" + Type(s) of text, for example Business correspondence, Newspapers, Periodicals, Records (Documents), etc.

A sample list of subject and genre/form headings for each type of data set appears in table 3, although the headings are neither exhaustive, nor required.

Creation of the Independent Data Sets Facet Value

A crucial task was to remediate existing data set records according to the newly established cataloging guidelines. Identifying existing data set records and examining each data set was an extremely time-consuming step of the project. Although two major data set collections, the Linguistic Data Consortium collection of text data sets and the Inter-university Consortium for Political and Social Research (ICPSR) collection of numeric data sets, were known to make up the majority of YUL's data set collection, to identify others, the task force searched for potential data set records based on:

- Data set-related keywords: Dataset, Data set, Datasets, Data sets
- Subject headings: Corpora (Linguistics), Geographic information systems, Biometry
- Form subdivisions: Statistics, Census
- Genre terms: Geospatial data, Raster data, Vector data, Census data, Statistics, Judicial statistics, Vital statistics, Demographic surveys

These searches, however, introduced tens of thousands of false positives, such as geological surveys in print books, voting data in scanned PDF documents, and statistics on computer reels, resulting in the authors spending a significant amount of time reviewing records to evaluate whether they met the basic criteria for data sets, namely data that can be downloaded, manipulated, and analyzed. This process was largely accomplished by importing the bibliographic records into MarcEdit to identify and eliminate false hits by using the "Select Records for Edit" function. For example, records describing computer reels or physical books in the MARC 300 field without supplemental CD-ROMs or DVD-ROMs were eliminated, as the data cannot be downloaded or manipulated. This lengthy review process further verified how inadequately bibliographic records previously described data sets and, consequently, how difficult it has been for users to discover them in the library software catalog. In the end, the task force identified and remediated over 11,000 data set records in bulk processes, including 10,547 records for numeric, 447 for text, 107 for sound, 24 for geospatial, and 9 for image data sets. While some titles surely remain incorrectly identified, the records will be converted as they are encountered in the future.

Whereas users can now find data sets as part of regular searches using the "Data sets" genre/form heading, it was also deemed crucial to improve Quicksearch's public interface to take advantage of the enhanced records to conduct more effective searches. Quicksearch is built on Blacklight, an open source discovery layer that uses Apache Solr for indexing and searching records.[36] Using Solr allows Blacklight to create and customize facets in a library catalog. With faceted searching, users can see the precise options they have available at any time. For example, a user may limit a keyword search to a specific field such as "Title," and narrow results by adding or removing terms from facets such as "Subject," "Location," and "Language." The user may also browse the facets without a keyword search, for example to display all records for resources with the format Video and in the French language.

Prior to this project, all records with "m" (computer file) in Leader/06, with the exception of database records, were broadly mapped to the format facet "Software & Datasets" in Quicksearch. As a result, the "Software and Datasets" format facet contained 18,639 titles, including not just data sets, but also other types of computer files, such as computer programs, games, fonts, computer-oriented multimedia, and online systems or services, making it difficult to isolate data sets. Moreover, this MARC format mapping was not entirely accurate. As described in Table 1, not all records use "m" in Leader/06 for data sets. As the mapping was neither precise nor sufficient to identify all types of data sets, the authors recommended that Library IT to create an independent "Data Sets" format to separate data sets from other computer files and to collocate all types of data sets. All records containing "dataset" in the core MARC 336 field $a, such as "computer dataset" and "cartographic dataset," were mapped to the new "Data Sets" format. A stand-alone format was also practical from a user experience perspective. Users inconsistently spell the word "data sets," as one word or two words. In Quicksearch, searching "data sets" as a form/genre as two words will return all matches, whereas searching "datasets" as one word returns no matches. To mitigate this inconsistent search behavior, it was deemed practical to explicitly display the "Data Sets" format upfront, with this format now adding up to 10,743 titles. The facet for other computer files, now totaling 7,896 resources, was renamed from "Software & Datasets" to "Software & Electronic Media."

Local Workflow at Yale University Libraries (YUL)

Several local policies and practices were implemented or established for efficiently managing the YUL data set

collection. A local workflow was created in response to the task force charge. It addresses local needs and data specialists and other stakeholders' requests, for example that bibliographic records for mediated data sets not be sent to OCLC due to concerns about strict licensing agreements.

Simplifying Discoverability with Hooks

In response to stakeholders' request for easy discoverability of all data sets and specific types of data sets in the library catalog and Quicksearch, the authors created convenient searching shortcuts for YUL staff. These hooks were designed to effortlessly identify specific varieties of data sets with keyword searches. These 090 fields are exclusively added to records in the local catalog. Multiple codes can be added to a single title if applicable. They include: yuldset (for all data sets), yuldsetgis (for geospatial data sets), yuldsetimg (for image datasets), yuldsetmediated (for mediated data sets), yuldsetnum (for numeric data sets), yuldsetsnd (for sound data sets), and yuldsettxt (for text data sets).

Providing Access to Mediated Data Sets

Access to data sets licensed by the library is restricted to members of the Yale community. Most resources are available through a direct link or via an intermediary page, which redirects users from accessing the resource directly by diverting them to a secondary page with particulars, such as instructions, information on digital tools and training, and a link to the remote resource.

Some data sets require staff mediation because access is limited to a certain number of simultaneous users, the data is too large for the researcher to store and manipulate on their own computers, or stringent licensing agreements. At the beginning of the project, many of these titles were not represented in the ILS, and the process to provide access to data sets that require staff mediation varied across YUL departments, leading to confusion for staff and users. The authors discussed several possible solutions with our stakeholders, including an online form, local website, a LibGuide, and Customer Relationship Management (CRM) technology, but ultimately settled on employing a mailto: link in the 856 field with the message: "For data access contact researchdata@yale.edu." This generates an email to a small group of YUL data specialists who then facilitate access. This is a straightforward process with little chance of error, as it allows experts to negotiate any issues that may arise.

Outreach to Library Staff

A "Dataset Review Request Form" was created to facilitate requests to review existing bibliographic records in the catalog for potential enhancements to the record. Additionally, requesters are encouraged to provide any special or specific information about a data set that may be helpful for the cataloger and patron, such as system requirements, digital file characteristics, data granularity, etc., so that the resources can be described effectively. This information was disseminated to selectors and other library staff via a mass email and a special edition of the library's *Electronic Resources Troubleshooting Newsletter*.

Work Products: Cataloging Documentation, MarcEdit Templates and Tasks

Documentation was created to address each type of data set to ensure that data sets are described consistently. To facilitate and ensure the accuracy of cataloging records, a variety of templates and a MarcEdit task list were created. A template is useful for cataloging new titles, particularly when a set of resources shares the same type, format, and/or collection. It allows static information to be pre-recorded, such as creators, issuing bodies, publication information, notes and local notes, access information, or subjects and genres. Since MarcEdit task lists enable batch updates of new or existing bibliographic records, this option proved useful for data sets based on previously published resources, e.g. databases, newspapers, and periodicals. The MarcEdit task and templates and all documentation is freely accessible to the greater cataloging community via the Cataloging at Yale website.[37]

Conclusion

YUL has embraced the growing importance of digital scholarship in academia with a strategic response for acquiring an increasing number and variety of data sets and enabling their discoverability. Integrating data sets into the library catalog is an acknowledgment of their standing as a standard research tool, but mainstreaming the collection necessitates precise metadata to ensure that they can be easily identified and retrieved in the discovery interface using facet, subject, and keyword searches.

This project was extremely challenging due to the lack of authoritative cataloging guidelines and the complex and evolving nature of the resources themselves. The authors employed existing best practices and standards, including MARC 21, RDA, LCSH, and LCGFT, resulting in bibliographic records that can be shared with other libraries, while responding to the needs of the YUL community and its local catalog and discovery interface. The project resulted in extensive documentation and tools that are regularly evaluated and updated. These cataloging guidelines enable YUL librarians to catalog both a backlog of data

sets and newly acquired titles in a uniform and systematic way, enhancing the discoverability of data sets in the public interface. The remediation of a large number of existing data set records to make them consistent with the new guidelines and add data set-related terms further improved discoverability and increased visibility and access to the data sets collection. Ongoing updates to the discovery interface ensure that resource discovery will become increasingly agile, while work continues on peripheral issues, such as ensuring that metadata clearly distinguishes electronic files that are not data sets. Clear workflows were implemented to assure that data sets are acquired and cataloged systematically.

The authors note that the project was more complex than anticipated because satisfying the objectives of the project required expanding the tasks from those originally outlined in the task force charge. For example, when the authors identified a lack of appropriate terms in the controlled vocabularies, they enriched them by successfully proposing numerous LCGFTs and LCSHs. These vocabularies, when consistently applied, assure that data sets (and materials about data sets and data mining) are easily retrievable with subject or genre/form searches. The project has greatly exceeded the three-month time frame originally predicted, and is expected to continue, as cataloging guidelines will require ongoing revisions to respond to the linked data environment, the inevitable changes in bibliographic description standards, and to address new issues and types of data sets as they develop or are acquired by the library.

While the authors developed a viable solution for identifying and cataloging data sets in their institution's catalog, they strongly recommend that the issues raised in this paper be addressed on a larger scale, preferably by a national group composed of representatives from various types of institutions. This group could discuss, establish, and document national guidelines for cataloging data sets so that these increasingly important resources are uniformly handled in institutional, consortial, and global catalogs, as the current patchwork of approaches makes for problematic discoverability and reinforces the inconsistent treatment of these resources by catalogers.

References and Notes

1. *Oxford Dictionary of English*, 3rd ed., s.v. "data set (noun)," accessed June 2, 2021, https://doi.org/10.1093/acref/9780199571123.001.0001.
2. Yale University Science Strategy Committee, Report of the University Science Strategy Committee, June 8, 2018, https://research.yale.edu/sites/default/files/ussc_report_may_2018.pdf.
3. Michael Gorman and Paul W. Winkler, eds., *Anglo-American Cataloguing Rules*, 2nd ed. (Chicago: American Library Association, 1978), 201.
4. Gorman and Winkler, *Anglo-American Cataloguing Rules*, 2nd ed., 203.
5. Michael Gorman and Paul W. Winkler, eds., *Anglo-American Cataloguing Rules*, 2nd ed., 1988 revision, (Chicago: American Library Association, 1988), 221.
6. Gorman and Winkler, *Anglo-American Cataloguing Rules*, 2nd ed., 1988 revision, 221.
7. Library of Congress Cataloging Policy and Support Office, "Library of Congress Implementation of Amendments 2001 to AACR2," accessed June 2, 2021, https://www.loc.gov/catdir/cpso/amen2001.html.
8. Amy K. Weiss, "Proliferating Guidelines: A History and Analysis of the Cataloging of Electronic Resources," *Library Resources & Technical Services* 47, no. 4 (2003): 180, https://doi.org/10.5860/lrts.47n4.171.
9. Greta de Groat, "A History of Video Game Cataloging in U.S. Libraries," *Cataloging & Classification Quarterly* 53, no. 2 (2015): 148, https://doi.org/10.1080/01639374.2014.954297.
10. RDA Steering Committee, Section 6.9.1.3, *RDA Toolkit*, accessed June 2, 2021, https://original.rdatoolkit.org/rdachp6_rda6-3421.html.
11. Library of Congress Network Development and MARC Standards Office, "347—Digital File Characteristics," *MARC 21 Format for Bibliographic Data*, accessed June 2, 2021, https://www.loc.gov/marc/bibliographic/bd347.html.
12. Grace D. Welch and Frank Williams, "Cataloguing Digital Cartographic Material," *Cataloging & Classification Quarterly* 27, no. 3–4 (1999): 360, https://doi.org/10.1300/J104v27n03_06.
13. Kathryn Lage, "Cataloging Digital Geospatial Data," *Journal of Map and Geography Libraries* 3, no. 1 (2007): 50, https://doi.org/10.1300/J230v03n01_04.
14. Lage, 51.
15. Lage, 52.
16. Library of Congress, "Library of Congress to Formally Separate LC Genre/Form Thesaurus from LCSH," accessed June 2, 2021, https://www.loc.gov/catdir/cpso/genreformthesaurus.html.
17. Mary L. Larsgaard, "Cataloging Cartographic Materials on CD-ROMs," *Cataloging & Classification Quarterly* 27, no. 3–4 (2005): 231, https://doi.org/10.1300/J104v27n03_07.
18. Data.gov; "Ocean: Data Quality and Documentation," accessed June 2, 2021, https://www.data.gov/ocean/data-quality-and-documentation-subpage.
19. Terry Reese, "Bibliographic Freedom and the Future Direction of Map Cataloging," *Journal of Map and Geography Libraries* 2, no. 1 (2006): 68, https://www.tandfonline.com/doi/abs/10.1300/J230v02n01_04.

20. Reese, 77.
21. Library of Congress, *Library of Congress Recommended Formats Statement, 2020–2021*, 2020, https://www.loc.gov/preservation/resources/rfs/RFS%202020–2021.pdf.
22. Library of Congress, "Proposal No. 97-3R," accessed June 2, 2021, https://www.loc.gov/marc/marbi/1997/97-03R.html.
23. Library of Congress Network Development and MARC Standards Office, "Leader," *MARC 21 Format for Bibliographic Data*, accessed June 2, 2021, https://www.loc.gov/marc/bibliographic/bdleader.html.
24. Library of Congress Network Development and MARC Standards Office, "Guidelines for Coding Electronic Resources in Leader/06," accessed June 2, 2021, https://www.loc.gov/marc/ldr06guide.html.
25. Library of Congress Network Development and MARC Standards Office, "Relationship of Fields 006, 007, and 008," *MARC 21 Bibliographic Format*, accessed June 2, 2021, https://www.loc.gov/marc/formatintegration.html.
26. Library of Congress Network Development and MARC Standards Office, "Guidelines for Coding Electronic Resources in Leader/06," accessed June 2, 2021, https://www.loc.gov/marc/ldr06guide.html.
27. RDA Steering Committee, Section 2.2.2.4, *RDA Toolkit*, accessed June 2, 2021, http://original.rdatoolkit.org/rdachp2_rda2-2904.html.
28. RDA Steering Committee, Sections 2.2.2.4.1 and 2.2.2.4.2, *RDA Toolkit*, accessed June 2, 2021, http://original.rdatoolkit.org/rdachp2_rda2-2905.html.
29. RDA Steering Committee, Section 3.3.1.3, *RDA Toolkit*, accessed June 2, 2021, http://original.rdatoolkit.org/rdachp3_rda3-2058.html.
30. RDA Steering Committee, LC-PCC PS for Section 3.4.1.3, *RDA Toolkit*, accessed June 2, 2021, http://original.rdatoolkit.org/lcpschp3_lcps3-3034.html.
31. Library of Congress Network Development and MARC Standards Office, "Term and Code List for RDA Content Types," accessed June 2, 2021, https://www.loc.gov/standards/valuelist/rdacontent.html.
32. Library of Congress Network Development and MARC Standards Office, "347—Digital File Characteristics," *MARC 21 Format for Bibliographic Data*, accessed June 2, 2021, https://www.loc.gov/marc/bibliographic/bd347.html.
33. RDA Steering Committee, Section 2.17.13.5, *RDA Toolkit*, accessed June 2, 2021, http://original.rdatoolkit.org/rdachp2_rda2-9509.html.
34. Library of Congress Subject Authority Cooperative Program (SACO), "Summary of Decisions, Editorial Meeting Number 1903," accessed June 2, 2021, https://www.loc.gov/aba/pcc/saco/cpsoed/psd-190315.html.
35. Library of Congress Cataloging Distribution Service, Classification Web, accessed June 2, 2021, https://classweb.org/.
36. GitHub, "Project Blacklight/Blacklight," accessed June 2, 2021, https://github.com/projectblacklight/blacklight.
37. Yale University Library, "Datasets (Documentation & Tools)," *Cataloging @ Yale*, https://web.library.yale.edu/cataloging/datasets.