

Maximizing Discovery of Datasets in the Library Catalog

Rowena Griem, Tachtorn Meier, Yukari Sugiyama
Yale University Library

Yale Library Digital Scholarship Support

Digital Humanities Lab



StatLab

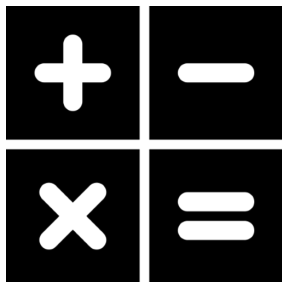


What is a Dataset?



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	State	Total	State Vote	Top Vote	Margin of Victory	Democrat	Republican	Democrat	Republican	Democrat	Republican	Democrat	Republican	Democrat
2	Alabama	2,084,308	1	1	1	1	1	1	1	1	1	1	1	1
3	Alaska	638,301	1	1	1	1	1	1	1	1	1	1	1	1
4	Arizona	6,392,011	1	1	1	1	1	1	1	1	1	1	1	1
5	Arkansas	2,915,324	1	1	1	1	1	1	1	1	1	1	1	1
6	California	35,981,325	1	1	1	1	1	1	1	1	1	1	1	1
7	Colorado	5,773,714	1	1	1	1	1	1	1	1	1	1	1	1
8	Connecticut	3,589,079	1	1	1	1	1	1	1	1	1	1	1	1
9	Delaware	923,854	1	1	1	1	1	1	1	1	1	1	1	1
10	District of Columbia	689,848	1	1	1	1	1	1	1	1	1	1	1	1
11	Florida	19,320,000	1	1	1	1	1	1	1	1	1	1	1	1
12	Georgia	9,010,000	1	1	1	1	1	1	1	1	1	1	1	1
13	Hawaii	1,212,343	1	1	1	1	1	1	1	1	1	1	1	1
14	Idaho	1,567,582	1	1	1	1	1	1	1	1	1	1	1	1
15	Illinois	12,812,508	1	1	1	1	1	1	1	1	1	1	1	1
16	Indiana	6,483,832	1	1	1	1	1	1	1	1	1	1	1	1
17	Iowa	3,190,861	1	1	1	1	1	1	1	1	1	1	1	1
18	Kansas	3,655,381	1	1	1	1	1	1	1	1	1	1	1	1
19	Kentucky	4,468,947	1	1	1	1	1	1	1	1	1	1	1	1
20	Louisiana	4,488,947	1	1	1	1	1	1	1	1	1	1	1	1
21	Maine	1,345,029	1	1	1	1	1	1	1	1	1	1	1	1
22	Maryland	6,045,381	1	1	1	1	1	1	1	1	1	1	1	1
23	Massachusetts	6,892,011	1	1	1	1	1	1	1	1	1	1	1	1
24	Michigan	10,712,343	1	1	1	1	1	1	1	1	1	1	1	1
25	Minnesota	5,773,714	1	1	1	1	1	1	1	1	1	1	1	1
26	Mississippi	2,915,324	1	1	1	1	1	1	1	1	1	1	1	1
27	Missouri	6,392,011	1	1	1	1	1	1	1	1	1	1	1	1
28	Montana	1,080,371	1	1	1	1	1	1	1	1	1	1	1	1
29	Nebraska	1,932,000	1	1	1	1	1	1	1	1	1	1	1	1
30	Nevada	2,915,324	1	1	1	1	1	1	1	1	1	1	1	1
31	New Hampshire	1,345,029	1	1	1	1	1	1	1	1	1	1	1	1
32	New Jersey	9,010,000	1	1	1	1	1	1	1	1	1	1	1	1
33	New Mexico	2,084,308	1	1	1	1	1	1	1	1	1	1	1	1
34	New York	19,320,000	1	1	1	1	1	1	1	1	1	1	1	1
35	North Carolina	10,712,343	1	1	1	1	1	1	1	1	1	1	1	1
36	Ohio	12,812,508	1	1	1	1	1	1	1	1	1	1	1	1
37	Oklahoma	3,655,381	1	1	1	1	1	1	1	1	1	1	1	1
38	Oregon	3,589,079	1	1	1	1	1	1	1	1	1	1	1	1
39	Pennsylvania	12,812,508	1	1	1	1	1	1	1	1	1	1	1	1
40	Rhode Island	1,080,371	1	1	1	1	1	1	1	1	1	1	1	1
41	Tennessee	6,392,011	1	1	1	1	1	1	1	1	1	1	1	1
42	Texas	25,145,560	1	1	1	1	1	1	1	1	1	1	1	1
43	Utah	3,190,861	1	1	1	1	1	1	1	1	1	1	1	1
44	Vermont	638,301	1	1	1	1	1	1	1	1	1	1	1	1
45	Virginia	8,010,000	1	1	1	1	1	1	1	1	1	1	1	1
46	Washington	7,373,714	1	1	1	1	1	1	1	1	1	1	1	1
47	West Virginia	1,812,343	1	1	1	1	1	1	1	1	1	1	1	1
48	Wisconsin	5,773,714	1	1	1	1	1	1	1	1	1	1	1	1
49	Wyoming	567,582	1	1	1	1	1	1	1	1	1	1	1	1
50	Total	311,000,000	1	1	1	1	1	1	1	1	1	1	1	1
51	Special Elections													
52	Mississippi	2,084,308	1	1	1	1	1	1	1	1	1	1	1	1
53	Missouri	6,392,011	1	1	1	1	1	1	1	1	1	1	1	1
54	Total	311,000,000	1	1	1	1	1	1	1	1	1	1	1	1
55	Election Date	11/02/2010												

Complexity and Variability of Datasets



Numeric



Geospatial



Image



Text

Three Main Problems:

1. Lack of clear guidelines to distinguish datasets from other computer files and record dataset characteristics in MARC
2. Limited dataset-related terms in controlled vocabularies
3. Existing dataset records were not cataloged in a consistent way

Creation of Documentation for Cataloging Datasets (Fixed Fields)

Type of record:
m=Computer File

Physical Description:
c=Computer File

Leader	02712cm m a2200589 i 4500	006	
005:	20200113094911.0	007	n
008	190610 m 2002 2019 miu _ _ o d _ _ _ _ _ eng _ d		

Tag	I1	I2	Subfield Data
040			\$a CtY \$b eng \$e rda \$c CtY
043			\$a n-us-ny
050	4		\$a F128.68.Q4
090			\$a yuldset
090			\$a yuldsetmediated
090			\$a yuldsettxt
245	0	0	\$a Newsday dataset, \$f 1940-1990.
246	3		\$a Historical Newsday dataset
246	3		\$a ProQuest Newsday dataset
264		1	\$a [Ann Arbor, Michigan] : \$b [ProQuest LLC], \$c [between 2002 and 2019?]

Type of File:

- a = Numeric
- c = Representational
- d = Document
- e = Bibliographic data
- m = Combination
- | = No attempt to code

Creation of Documentation for Cataloging Datasets (Variable Fields)

300				‡a 1 online resource (approximately 2 million text files)	Physical Description: Number and type of files
336				‡a computer dataset ‡b cod ‡2 rdacontent	Content Type for Dataset + Additional Content Type
336				‡a text ‡b txt ‡2 rdacontent	
337				‡a computer ‡b c ‡2 rdamedia	Digital File Characteristics: <ul style="list-style-type: none">• a [File type]• b [Encoding format]• c [File size]
338				‡a online resource ‡b cr ‡2 rdac	
347				‡a text file ‡2 rdaft	
347				‡b PDF ‡b XML	
347				‡c 508.18 GB	Summary, Etc.
500				‡a Title, variant titles, and title of collection devised by cataloger.	
506				‡a Access restricted by licensing agreement and agreement to terms of use.	
520				‡a Dataset of articles for text data mining (TDM) from Newsday, dating from September 3, 1940-December 31, 1990. The set contains digital reproductions in PDF and XML format, both segmented into issues and articles.	
786	0	8	‡i	Based on (work) ‡s Newsday (Nassau edition). ‡t Newsday. ‡b Nassau edition ‡w (OCoLC)ocm05371847	Linking Field Entries: <ul style="list-style-type: none">• 786: Data Source Entry• 787: Other Relationship Entry
787	0	8	‡i	Related work: ‡t Proquest historical newspapers	

Creation of Dataset-Related Terms for Controlled Vocabularies

655 - LCGFT	650 - LCSH
Data sets	Data mining--Statistical methods
Biostatistics	Image data mining
Medical statistics	Spatial data mining
<i>Text corpora</i>	Text data mining
Image data sets	v Data sets
Spatial data sets	ACCEPTED UNDER CONSIDERATION NOT APPROVED*
Statistical data sets	
Text data sets	
* For details, see Summary of Decisions, Editorial Meetings for March & June 2019	

Implementation of Dataset-Related Terms for Controlled Vocabularies

	LCGFT	LCSH
All datasets	655 /7 a Data sets. 2 lcgft	+ LCSHs describing the topical, geographic, and chronological aspects of the datasets (The headings in bold are added to all records for that type of dataset.)
Geospatial datasets	655 /7 a Geospatial data. 2 lcgft 655 /7 a Maps. 2 lcgft 655 /7 a Vector data. 2 lcgft	
Image datasets	655 /7 a Images. 2 lcgft	
Statistical datasets	655 /7 a Statistics. 2 lcgft 655 /7 a Census data. 2 lcgft	
Text datasets	655 /4 a Text corpora. 655 /7 a Blogs. 2 lcgft 655 /7 a Newspapers. 2 lcgft	

Remediation of existing dataset records

Identify potential dataset records by:

- Well-known collections (Linguistic Data Consortium, ICPSR)
- Keywords (Dataset, Data set)
- Topical subject headings (Corpora (Linguistics), Geographic information systems, etc.)
- Form subdivisions (Statistics, Census, etc.)
- Genre headings (Geospatial data, Census data, Raster data, Vector data, etc.)

False positive examples:

- PDF document of voting data on CD-ROM
- Statistics on computer reel

Recommendation for discovery layer (Format Facet)

Format	
Archives or Manuscripts	42,149
Audio	165,987
Books	10,371,510
Databases	1,060
Dissertations & Theses	224,095
Images	31,133
Journals & Newspapers	518,054
Maps & GIS	78,216
Microforms	573,241
Notated Music	153,006
Online	2,549,708
Other	17,809
Software & Datasets	18,639
Video	161,147



Format	
Archives or Manuscripts	42,149
Audio	165,987
Books	10,371,510
Databases	1,060
Datasets	10,743
Dissertations & Theses	224,095
Images	31,133
Journals & Newspapers	518,054
Maps & GIS	78,216
Microforms	573,241
Notated Music	153,006
Online	2,549,708
Other	17,809
Software & Electronic Media	7,896
Video	161,147

We mapped all records that have 336\$a = “computer dataset” to Datasets

Recommendations for discovery layer

You searched for:

Form/Genre: data sets ✕

« Previous | 1 - 50 of 10,890 | Next » 

You searched for:

Form/Genre: datasets ✕

No results found for your search.

Quicksearch

Demo:
<https://youtu.be/KEWDUDquEvk>

Quickstart

Books+

Databases

Articles+

Digital Collections

More...

Limit your search:

Format	-
Data Sets	140
Databases	1
Course	140
Publication Date	+
Author/Creator	+
Recently Added	+
Location	-

philadelphia

All Items



140 Items Found

2 days limit

Philadelphia

Philadelphia X

Format: Is - Data Sets X

+ Detailed | 1 - 20 of 140 | Next >

View Options

Sort by Relevance

Detailed Format

14 Philadelphia inquirer dataset

Published: James A. Smith, Michigan | ProQuest LLC | Between 1850 and 1890

Online: Full-text access (online) from multiple off-campus

Location: 140 Internet Resource(s) > Online Resource

Format: Data Sets > Online

6 Philadelphia Social History Project Grid Data, 1850, 1860, 1870, 1880

Published: Ann Arbor: Univ. Microfilms Int. | Copyright for individual use of researchers (2000) only

Online: Online only

Location: 140 Internet Resource(s) > Online Resource

Format: Data Sets > Online

Thank you

Yale dataset MARC cataloging documentation:

<https://web.library.yale.edu/cataloging/datasets>

Our promotional poster for the Yale community:

<https://elischolar.library.yale.edu/dayofdata/2019/posters/7/>

Rowena Griem, Catalog Librarian for E-Resources & Serials Management rowena.griem@yale.edu

Tachtorn Meier, Catalog Librarian tachtorn.meier@yale.edu

Yukari Sugiyama, Metadata Librarian for Discovery and Assessment yukari.sugiyama@yale.edu