

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



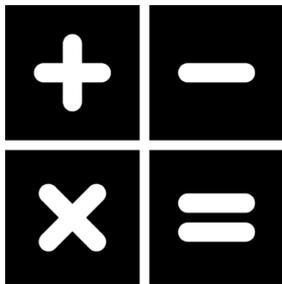
What is a Dataset?



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	State	Total	Seats Won	Pop. Vote	Margin of Victory	Democrat	Democrat	Democrat	Democrat	Democrat	Democrat	Democrat	Democrat	Democrat
2	Alabama	2,984,328	1	2	1	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
3	Alaska	633,334	1	1	1	88.82%	54.36%	1,899,421	54.36%	1,899,421	54.36%			
4	Arizona	2,984,328	1	2	1	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
5	Arkansas	2,984,328	1	2	1	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
6	California	37,751,234	55	55	55	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
7	Colorado	5,312,000	7	7	7	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
8	Connecticut	3,586,444	7	7	7	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
9	Delaware	923,854	3	3	3	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
10	Florida	19,314,551	11	11	11	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
11	Georgia	10,519,654	14	14	14	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
12	Hawaii	1,295,381	1	1	1	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
13	Idaho	1,609,477	3	3	3	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
14	Illinois	12,812,232	13	13	13	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
15	Indiana	6,483,802	11	11	11	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
16	Iowa	3,191,878	7	7	7	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
17	Kansas	3,655,381	6	6	6	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
18	Kentucky	4,469,297	6	6	6	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
19	Louisiana	4,488,924	6	6	6	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
20	Maine	1,345,000	3	3	3	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
21	Maryland	2,984,328	6	6	6	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
22	Massachusetts	6,779,603	11	11	11	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
23	Michigan	10,475,871	11	11	11	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
24	Minnesota	5,597,065	10	10	10	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
25	Mississippi	2,984,328	6	6	6	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
26	Missouri	5,937,923	11	11	11	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
27	Montana	1,043,923	3	3	3	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
28	Nebraska	1,934,457	3	3	3	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
29	Nevada	2,984,328	3	3	3	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
30	New Hampshire	1,323,417	3	3	3	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
31	New Jersey	8,901,901	14	14	14	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
32	New Mexico	2,079,321	3	3	3	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
33	New York	19,453,561	13	13	13	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
34	North Carolina	10,419,328	13	13	13	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
35	Ohio	11,353,281	13	13	13	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
36	Oklahoma	3,755,381	5	5	5	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
37	Oregon	3,755,381	6	6	6	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
38	Pennsylvania	12,465,381	13	13	13	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
39	Rhode Island	1,057,381	3	3	3	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
40	Tennessee	6,345,381	11	11	11	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
41	Texas	28,189,328	32	32	32	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
42	Utah	2,984,328	3	3	3	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
43	Vermont	623,381	3	3	3	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
44	Virginia	8,001,381	13	13	13	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
45	Washington	7,001,381	7	7	7	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
46	West Virginia	1,801,381	3	3	3	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
47	Wisconsin	5,937,923	10	10	10	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
48	Wyoming	523,381	3	3	3	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
49	Total	312,334,328	538	538	538	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
50	Special Districts													
51	Montana	2,984,328	1	1	1	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
52	Mississippi	2,984,328	1	1	1	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
53	Alabama	2,984,328	1	1	1	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
54	Total	11,937,328	3	3	3	55.36%	22.34%	1,191,100	49.96%	1,131,200	47.61%			
55	Effective Date:	11/6/2016						47,438,100		34,693,828				

ReStore @ National Centre for Research Methods
<http://www.restore.ac.uk/qeo-refer/91010mwrws01y0000000.php>

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 (Variable Fields)

300	‡a 1 online resource (approximately 2 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 rdac
347	‡a text file ‡2 rdaft
347	‡b PDF ‡b XML
347	‡c 508.18 GB
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
787 0 8	‡i Related work: ‡t Proquest historical newspapers.

Physical Description:
Number and type of files

Content Type for Dataset +
Additional Content Type

Digital File Characteristics:

- |a [File type]
- |b [Encoding format]
- |c [File size]

Summary, Etc.

Linking Field Entries:

- 786: Data Source Entry
- 787: Other Relationship Entry

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	<p style="text-align: center;">ACCEPTED <i>UNDER CONSIDERATION</i> NOT APPROVED*</p>
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:

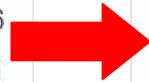
- Well-known collections (Linguistic Data Consortium, ICPSR)
- Keywords (Dataset, Data set)
- Topical subject headings (Corpora (Linguistics), Geographic information systems, etc.)
- Form subdivisions (Statistics, Census, etc.)
- Genre headings (Geospatial data, Census data, Raster data, Vector data, 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	14
Publication Date	+
Author/Creator	+
Library Access	+
Location	-

philadelphia

249 Items

Relevance By Philadelphia Format: Is - Data Sets

1 - 20 of 140 items

View Options Sort by Relevance Detailed View

1 Philadelphia inquirer dataset

Published: [Ann Arbor, Michigan] : ProQuest LLC, between 2000 and 2010

Genre: Full-text (222) Images (2) Audio (0)Location: All Internet Resources Online ResourcesFormat: Data Sets Online

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

Published: Ann Arbor, Mich. : ProQuest LLC, between 2000 and 2010

Genre: Online (14)Location: All Internet Resources Online ResourcesFormat: 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