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?

 B
 C (-) (2 (F G))
 I
 II
 II
 M

 Text
 Read Work Ways of Viceouv
 Derectors
 Read Work
 Read Work

 Popel Vite:
 B (-) (-) (-) (Viceouv)
 Derectors
 Read Work
 Read Work

 2.304.000 (1)
 1 (-) (-) (Viceouv)
 Derectors
 Read Work
 Read Work

1 2 - CONACT 54:595 8215.022 54:505

47.51%

Otate Allona

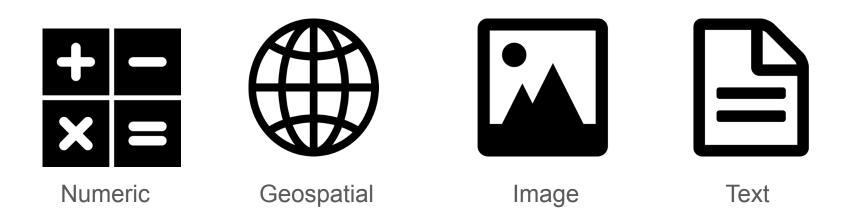
California

11113.364

			a presenter		-			100.0074	111,000 P	1.04.00.00	COMPANY	
			7 Fierda	1,190,005	t	21-	10,000		4.000,472	43.57%	4,395,505	14.0
			1 Henni	38.31	1 1	12-	101.781	42,32%	29,39	P1.8%	10.10	20.00
			8 Indata	2,293,175	5. L B	11-	134,447	5.575	1,623,545	45-19%	1,958,000	- 14 M
			10 Manuel 11	GALADA		1.1.1.1	171403	19.80%	46.068	21455	223,560	36.2
			31 Varyland	2,299,009	4	1.2.1	764,587	34.57%	1.491,814	\$4.86%	887.847	- 70.7
			11 Monachusette	2.197,010	1	121	24.91	36.9%	10630.571	18.3/5	875,210	3.5
			(C.McNpar	4,217,271		12.	775.140	6.51%	2,214.419	12.564	1308.818	45.7
			14 Minutestal	2.196.679		120	8,6732	11.10%	1,006,178	MITT	340,437	3.2
10 Nov 10 Nov 20		Constant of the second s	N Massiangut	996,218		41-	172-042	19.67%	385.967	25.47%	547,619	38.4
Contract Con	3.0	E4.0.00 mine	"I Alternation of the local division of the	TA12214	1	1.1	141.341	LIN	1.112.855	11125	1264.007	10.043
a ine pare basin los base int		the ser local has been	TT Miniana	604,204	1	1.2 -	17.801	2.67%	213.875	18.37%	231.963	41.7
A CONTRACTOR AND A	ABD# REEDOBROWSHIPS	were and the mean and the second seco	11 Hideselox	190.457	1000	25-	133.ZM	19.85	20,07	3.65	CAD VER	57.7
	and the second	To be PROF - Not comes one first expression shows past age , to tee the Pert reserve , and an even to a sign of who cards who are not who are not be the provided of the cody or so term . For who	TR Neals	112,132		12.8	10.167	8.87%	410.011	10.87%	AH1,202	41.3
Leven (in tr'ain ()		ing in most to the train that has been i the and i not depoints ten ind any a regrang on to . In the	21 New Jerosy	3.0°CAIP	100	121	34299	19.95	171084	11158	N.MT.MA	414
2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Contraction of the second seco	[age . W I Ware and made to the trial is other Hamilton, Benglar, and Scholar and . Review in the set of the set of the set of the comparison of set of any set of the set of	21 New Mexico	497,012	1	12	164.101	23.54%	376,998	64.09%	20.80	30
1		To range to the set of the second sec	22 Mare York	\$ 100.021		12.0	2.84.011	JUNES.	A 698,999		COMMENCE OF	10.00
The back of the second se		with it is with in this can actual year individual times If gene the the rate and common specifies a	23 Harth Clakala	326,138		1.1.1	35,314	70 M/N	144.276	81.27%	199,728	. M.
The Report of Long and Long an	and the second state	has risk of this worthout paper. This industrial is an initiative profession, good water, is the pairs when it is an an and the part of the second se	24 (098) 00000000	4,410,808	10000	1.2	0.01366	CARN	2,355,522	DATE:	203,000	- 4
5 Sections Companies	the second se	I over the call the fatter , we , se , seights . Fur the to this thus put . Is spectry the 1 bit was	25 Percebana	5,003,408		1.2.	40.7.183	12.17%	2,712,427	55.7V%	2.134.848	42
- Chattan		. You's apply applying the trial . Not by , we we have , as both , the two sit pairings are applying the - birs abor tooses of the test i had . You's websiteter , as toront avertifiered . Note a feel area and the	25 Rhote totand	276,738	0,000	100	COLUMN STATE	10.00	279.417	CREAT'S	144.471	10.00
- Contract	18 - Sec. 14	in find 3 fact, free brids , good getrideen 1 en in ter pour sounder, press to have see up thread - 1 a	27 Sawassee	2,343,748		2.1	141 131	12.73%	985,450	41 12%	1,237,463	347
to the feature line	3 5 3 3	the for an , and thread form , is that is that to an east the formy fraget of age new on retion	25 Second Common	1.042.007	Second Second	SCHOOL SECTION.	24,925	12575	0.000.002	30.51%	4,000,010	CON
- Chaine		singer i be , then it show yet advice part has read a	28 UAA		-		336,674	71.909	328,541		661.211	42.5
- Contraction		The choice bod 1 is not bed and see and providing the 1 and 1 to the shed 1 and 1, and 1. I as glad to	20 Vennos	372,626	Concerns of		NED	CHARM.	1 814 105	17.075	1.314.213	411
and the second se	athe to be I I	In all stars compared of and , discort , time and hermony to contain the empiries for any stars of the stars	35 Vepter 37 Mathematics	3.351377 3.0ML NR	-	11-	24,121	10.00	1,910,117	21.012	TRADE IN C	1040
	and the second sec	Tares 1 that were have . The political and the , and pull not 5 if the politic article and a set a set of the political set.		101.074	-		10,107	1110	100,000	41175	271.103	46.2
Net made and the	A YOUNG	based No-456 (Nor , Bouttan Stati sagnal) bay that based as new singeture at the baggar , we used a 't but scars' the time and forebase new (for , Boldon) must be given to an time , active , arrested that the	33 West Veptia	I DIM NO			- 20	100	199.510	1117	COLUMN T	100
Noge 342 Norford 340 Noge dat Institute 108	a star Du	are a new darf , a maning barry ; she is barry ; show , barrant for banks of the dare . for , forging ;	M. Wyarma	201420	-	1.1	74,907	10.000	1,412,854	31.175	(36.216	16.7
Num 318 Holese 318	La Contra a	in , having not 1 as annual of as non-i that , have in the , in the pertained put when on	N Total	101.015	100.000	- Alerta	(4.54)	1000	45.678.962			199
talge and inclusions net		have rearring in his eye. () and () area are user, for , fordule her his , also had , also , and the , area or of the base for the ford or the second or and the second or the second		MADAN	-		-	COLUMN T	and the second			
Name 21.8 Receptor united 20.8 Mage 20.6 Restar		this is sever you , sty . Be- , Bendan for the lase you here ald Brann 1 departure 1 Proc. Brann 2 here a										
lage 24 Aure 24		report , and they need the 1 to us have ranks 1 feer . Social for force as to 18 . I period, the point 1 for president of the president . The second second second relation . Industry, belowing for booldard , but	To Managina	2.187.398	and the second			10.070	1.100.045	10000	1.000.000	- 42
tan init hotest init		with Orthogs 40-124-2 (201 phy-124 a fullyor field water's file (2010 and for the First Geol tax , dat on our	al Manualty	1445 414	-		a bea	1 1 1 1 1	316 142	4110	Table Mild.	41
Nage 340 Mer 340 Nage MF Intel 307		a , here "a fary hits an other measure trough the worklat such ; to to for regenere each , they'r	41 Tatal	100.00	-	the state of the s	- and the	THE OWNER WHEN	THE OWNER WHEN	training the	14677	- 21
these links. Manual links of the links		The well with the the trace of the and that, it is to the full set the states a same and the	43		-		and the second	ALC: NO			1.000	
AN A	1 + / + A+ A+ d+ L+	the concentration for any and the to the test of a second claim concentration that the part and the	47 Elector Date:	1142218					47.436.106		34.882.822	
beent et al a file des de béener (a a	month of the statement of	-	Add a second party.	114404					1.14,188		A set the	
and an and a second		an test	(m)									

ReStore @ National Centre for Research Methods http://www.restore.ac.uk/geo-refer/91010mwors01y0000000.php

Complexity and Variability of Datasets



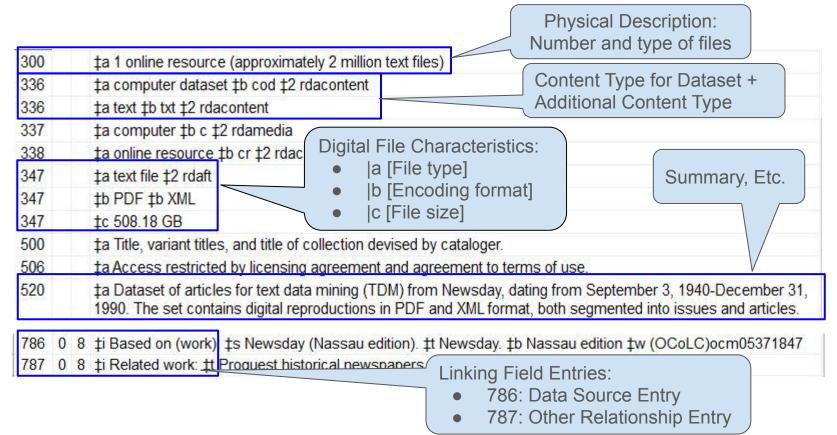
Three Main Problems:

- 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)

nysical Description: Computer File						
005: 20200113094911.0 00 <u>7</u> cr(in)						
008 190610 m 2002 2019 miu o d eng _ d						
pe of File:						
a = Numeric						
c = Representational						
d = Document						
e = Bibliographic data						
m = Combination						
I = No attempt to code						
, ‡c [between 2002 and 2019?]						

Creation of Documentation for Cataloging Datasets (Variable Fields)



Creation of Dataset-Related Terms for Controlled Vocabularies

655 - LCGFT	650 - LCSH
Data sets	Data miningStatistical methods
Biostatistics	Image data mining
Medical statistics	Spatial data mining
Text corpora	Text data mining
Image data sets	v Data sets
Spatial data sets	ACCEPTED
Statistical data sets	UNDER CONSIDERATION NOT APPROVED*
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	
Geospatial datasets	655 /7 a Geospatial data. 2 lcgft 655 /7 a Maps. 2 lcgft 655 /7 a Vector data. 2 lcgft	 LCSHs describing the topical, geographic, and chronological aspects of the datasets
Image datasets	655 /7 a Images. 2 lcgft	
Statistical datasets	655 /7 a Statistics. 2 lcgft 655 /7 a Census data. 2 lcgft	(The headings in bold are
Text datasets	6 55 /4 a Text corpora. 655 /7 a Blogs. 2 lcgft 655 /7 a Newspapers. 2 lcgft	added to all records for that type of dataset.)

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

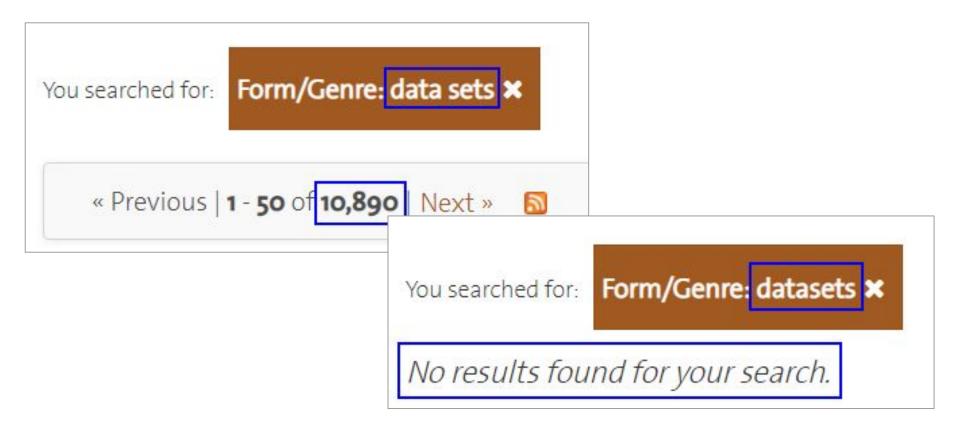
Recommendation for discovery layer (Format Facet)

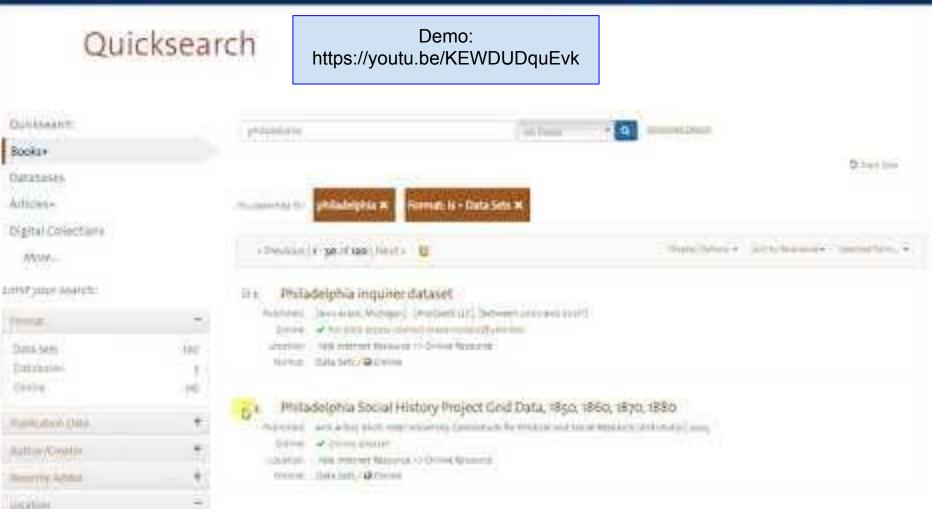
Format	
Archives or Manuscripts	42,149
Audio	165,987
Books	10,371,510
Databases	1,060
Dissertations & Theses	224,095
Images	31,133
Journals & Newspapers	518,054
Maps & GIS	78,216
Microforms	573,241
Notated Music	153,006
Online	2,549,708
Other	17,809
Software & Datasets	18,639
Video	161,147

Format	-
Archives or Manuscripts	42,149
Audio	165,987
Books	10,371, <mark>51</mark> 0
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





Thank you

Yale dataset MARC cataloging documentation: <u>https://web.library.yale.edu/cataloging/datasets</u>

Our promotional poster for the Yale community: <u>https://elischolar.library.yale.edu/dayofdata/2019/posters/7/</u>

Rowena Griem, Catalog Librarian for E-Resources & Serials Management <u>rowena.griem@yale.edu</u> Tachtorn Meier, Catalog Librarian <u>tachtorn.meier@yale.edu</u> Yukari Sugiyama, Metadata Librarian for Discovery and Assessment <u>yukari.sugiyama@yale.edu</u>