Maximizing Discovery of Datasets in the Library Catalog

Rowena Griem, Tachtorn Meier, Yukari Sugiyama
Yale University Library

Yale Library Digital Scholarship Support

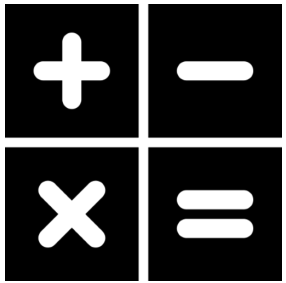
Digital Humanities Lab

StatLab

What is a Dataset?

ReStore @ National Centre for Research Methods
http://www.restore.ac.uk/geo-refer/91010mwors01y00000000.php

Complexity and Variability of Datasets



Numeric

Geospatial

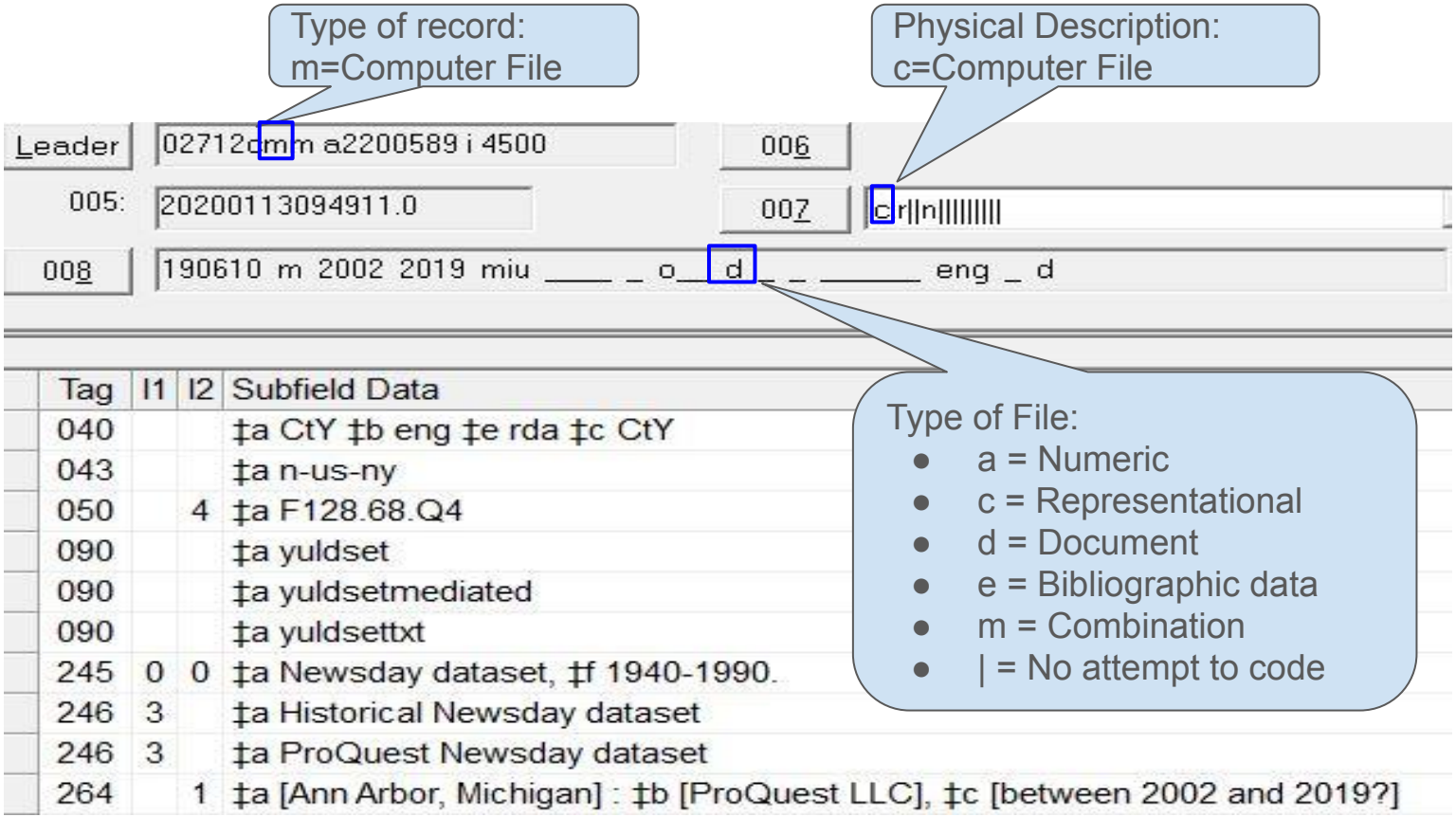
Image

Text

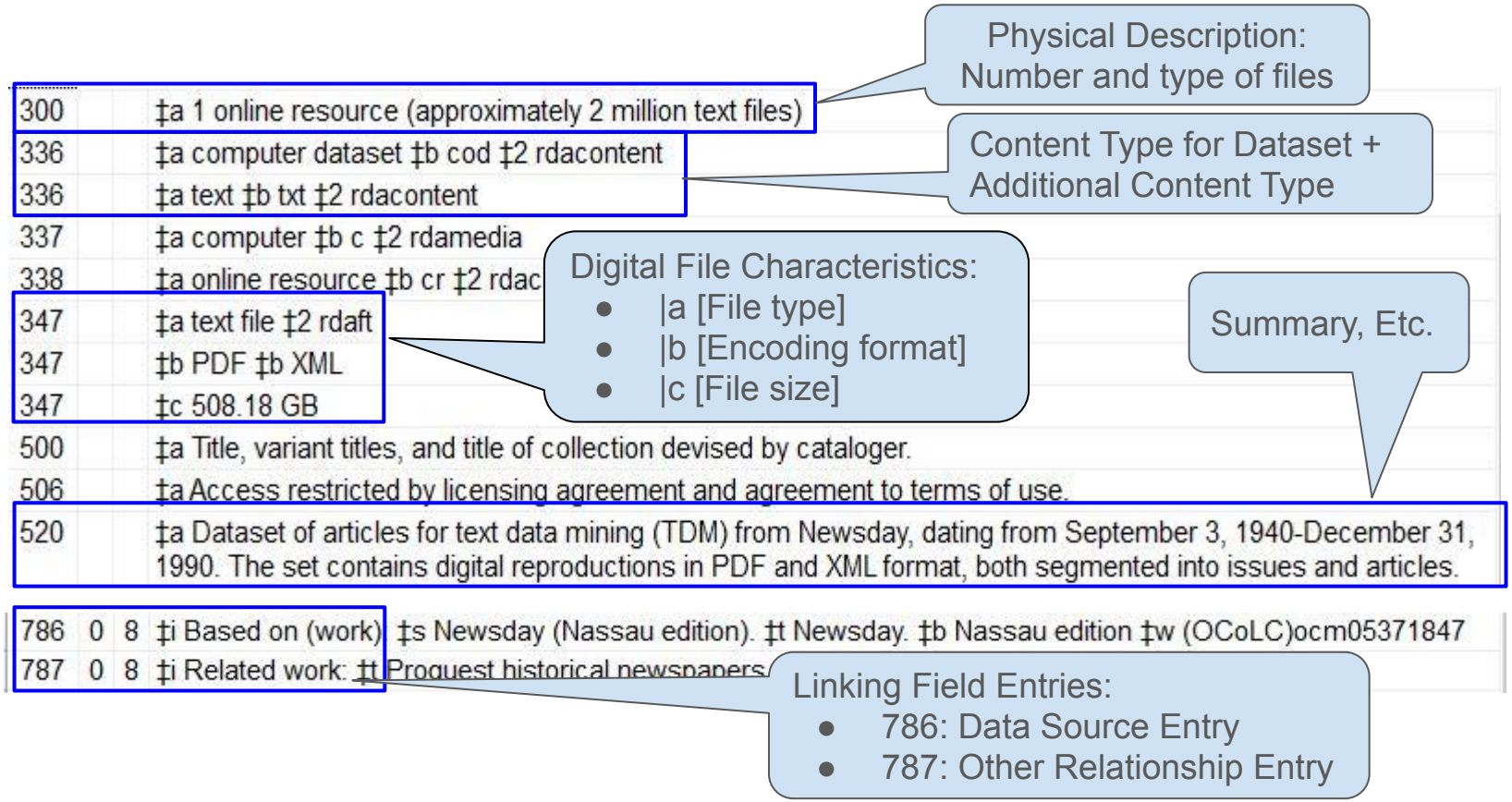
Three Main Problems:

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 (Fixed Fields)



Creation of Documentation for Cataloging Datasets (Variable Fields)



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**
Text corpora	**Text data mining**
Image data sets	|v Data sets
Spatial data sets	**ACCEPTED** *UNDER CONSIDERATION* NOT APPROVED* * For details, see Summary of Decisions, Editorial Meetings for March & June 2019
Statistical data sets	
Text data sets	

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.)

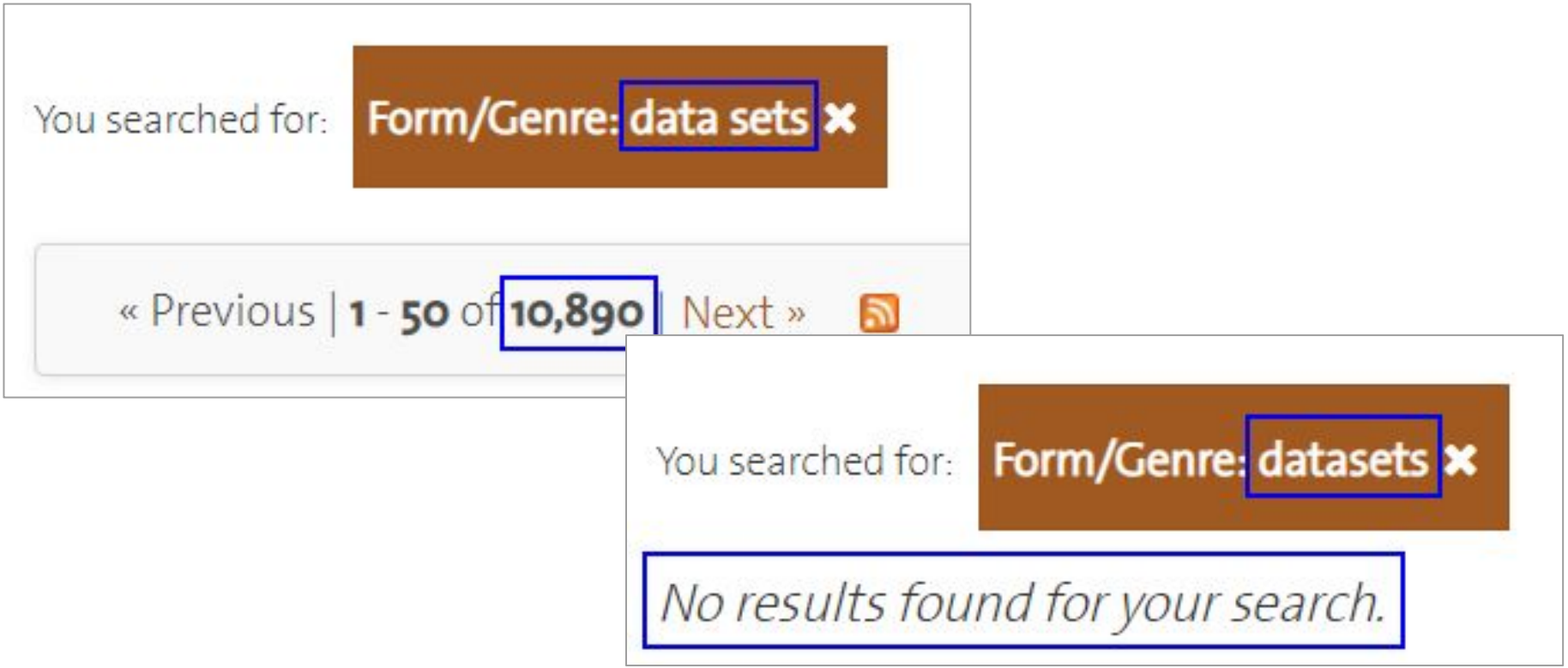
False positive examples:

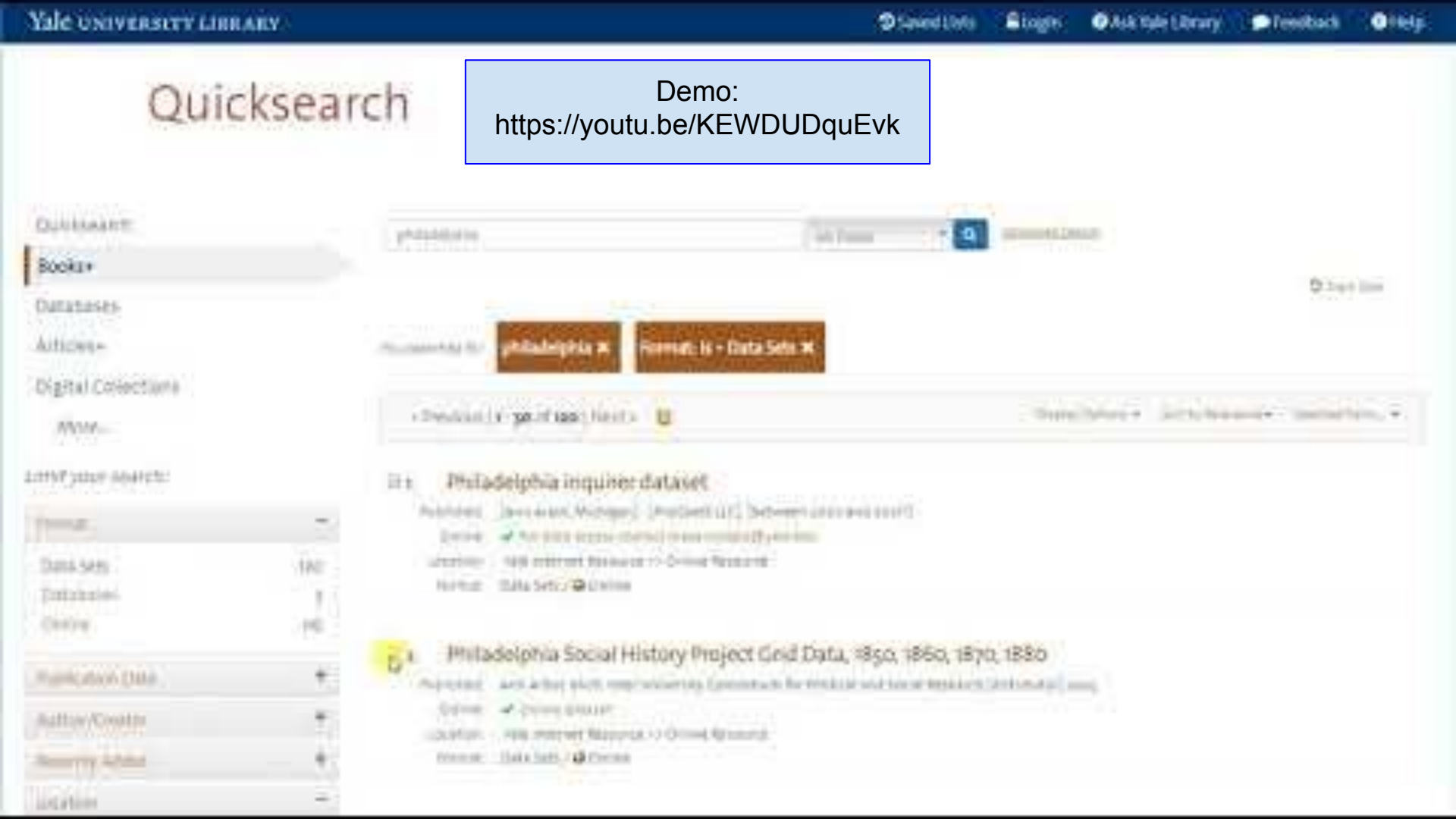
- PDF document of voting data on CD-ROM
- Statistics on computer reel

Recommendation for discovery layer (Format Facet)

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

Recommendations for discovery layer




Yale University Library
Quicksearch
Demo:
https://youtu.be/KEWDUDquEvk
Books+
Databases
Articles+
Digital Collections
philadelphia
Philadelphia Inquirer dataset
Philadelphia Social History Project Grid Data, 1850, 1860, 1870, 1880

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