OhioLINK

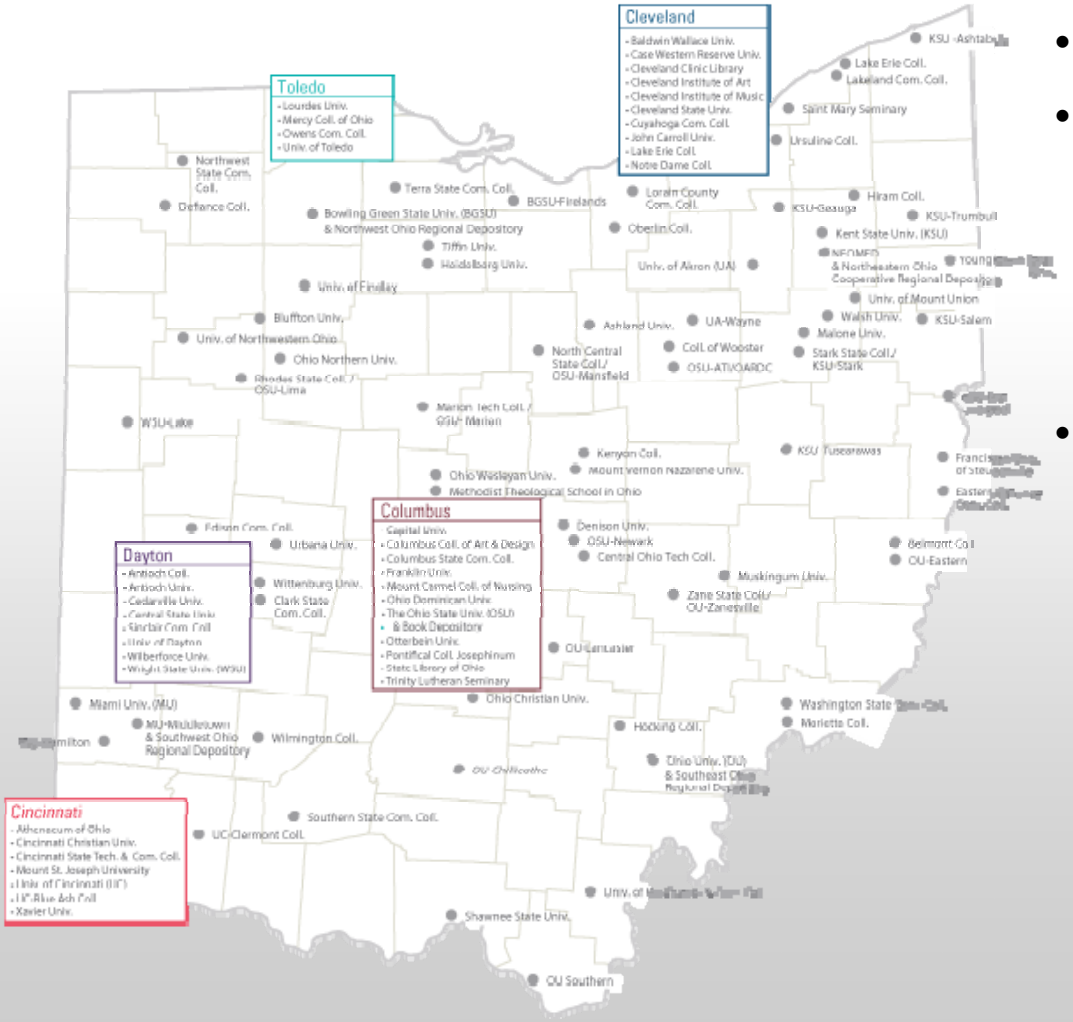
Ohio's Academic Library Consortium

Uniqueness and Duplication in Ohio's Shared Depository System

Joanna Voss
Theda Schwing

Great Lakes Resource Sharing Conference
June 10th, 2016

OhioLINK is Member-Driven



- 121 member libraries
- Diverse academic libraries:
  - Public universities
  - Two-year and technical colleges
  - Private colleges & universities
  - State Library of Ohio
- Services:
  - Shared collections via courier
    - Public library partnership SearchOhio
  - Millennium INNReach central catalog
  - Shared eresources:
    - ebooks, ejournals, Databases
  - Theses & dissertations, discovery layer, link resolver

Ohio Regional Depository System

- Shared regional depositories were built in the 1990s
- Purpose is to efficiently store material but retain the ability to access items quickly

- Regional Depositories Governing Council

- 13 participating libraries
- 5 locations (NW, NE, SW, SE, Central)
  - Each library is associated with one of these locations
  - Each location is managed by one of it's participating libraries

- About 8 million items – nearing capacity

Ohio Regional Depository System

Shared Ohio Depository Catalog (OHDEP)

- Purpose
  - To simplify the processing necessary at the depository locations
  - To make it easier to deduplicate the collection

- What is it?
  - An Innovative catalog, launched in 2010
  - Shared by 8 institutions
  - Covering 3 physical locations
  - Containing about 2.67 million items (or about 928K titles)
  - With about 58% monographs and 42% serial items

Questions Raised with a Shared Collection

- Not everyone is participating – can we get data on the whole system?
- How often do items circulate?
  - Reliable stats only start with when items are loaded into OHDEP
  - Stats are at the item, not title, level
- How to keep MARC records normalized?
  - Have a series of record loaders to help with this, but it's not perfect
- What is the duplication rate?

Approach

- Collection development in a shared catalog
  - One part of wider examination of the Ohio Shared Depository System
    - Define current state, chart a path forward
  - Look at items in shared depository catalog (OHDEP)
  - Approach monographs and serials separately

- What can we do with the data and tools available?
  - ILS data
  - OpenRefine
  - Current library staff and equipment

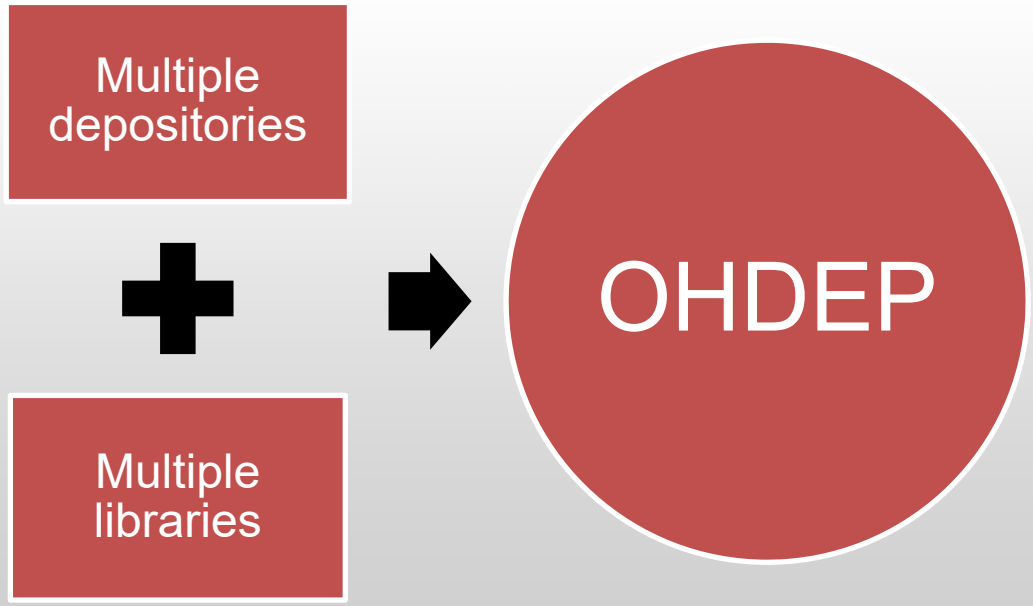
Now What?

- What can we ask to shed light on the current situation and a path forward for the depositories?
  - Preserve access to breadth and depth of content while preparing for the future
    - OHDEP as test case for analysis
    - Establish tools and methods for future analysis

- Questions:
  - How many monograph items are duplicated within those depositories participating in the OHDEP catalog?
  - Can we characterize these items in terms of publisher, publication date, or other parameters?

Depository Catalog Data

- Which catalog data is useful?
  - Bibliographic
    - Title
    - Author
    - Imprint
    - Date
    - Volume
    - Call number
  - Identifiers
    - OCLC #
    - ISBN
  - Administrative
    - Circulation
    - Bib / item record #



How to get the right data out?

- OHDEP catalog is already just depository items
- Create List in Millennium
  - Only monographs (Bib level = m)
  - 001, bib #, 008 Date 1, 090|a, 050|a, 245|a, Volume, 260|b & 264|b

- Item-level export: 1,404,314 records (150 MB)
  - Export in 4 chunks of 400,000
  - Field delimiter: |
  - Repeated field delimiter: ~
  - Text in “

```
"001"|"RECORD #(BIBLIO)"|"008 Date One"|"090|a"|"050|a"|"245|a"|"VOLUME"|"260|b"|"264|b"
"413327"|"b18270281"|"1954"|"DT108.6"|""|"Sudan days and ways"|""|"Macmillan;"~"St. Martin's Press,"|""
```

Data Considerations

- Excluded records with multiple or no OCLC number
- Final analysis on 1,402,400 records

- Cleanup:
  - Volume
    - How to normalize?
  - Title: MARC 245a
  - Imprint: MARC 260b or 264b

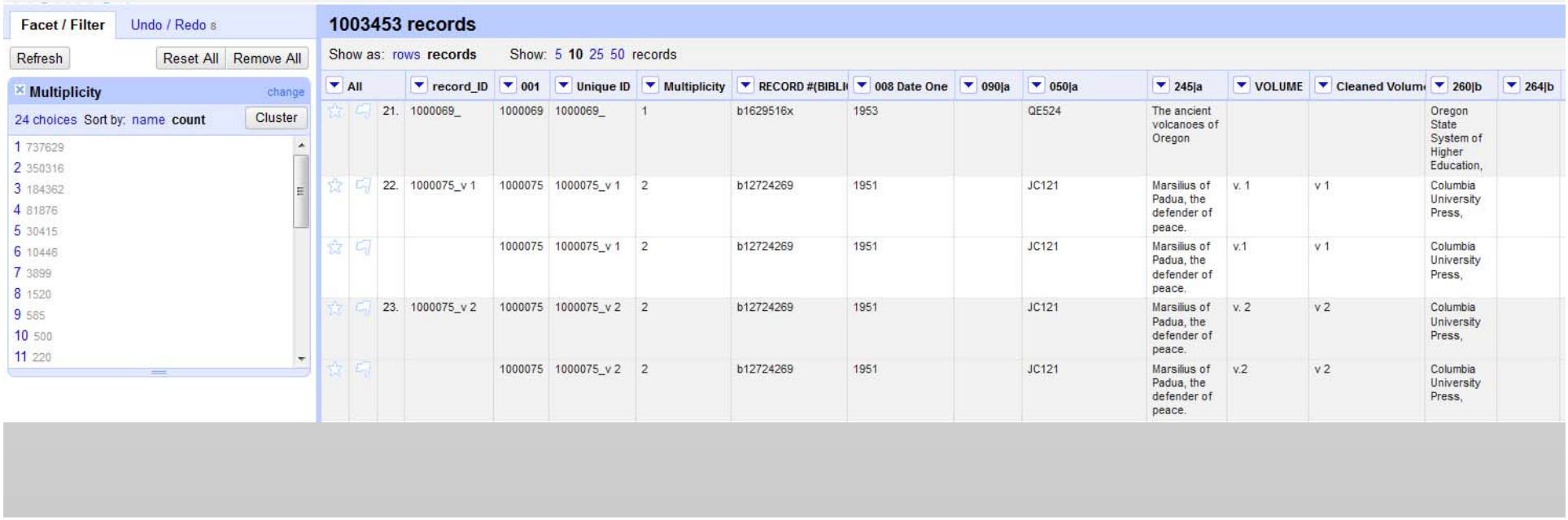
Methodology

- OpenRefine (http://openrefine.org/)
  - Open source data cleanup & analysis tool
    - Formerly Google Refine, Freebase Gridworks
    - Current stable release 2.5 (2011)
  - Built-in text cleanup features
  - Google Refine Expression Language (GREL)

- Data Cleanup & Analysis
  - Append text files together
  - Validate data
  - Test different duplication algorithms
  - Characterize monographs by publisher, year of publication, etc.
  - Can we look at circulation?

OpenRefine

- Create unique identifier based on varying criteria
- “Duplication” = proportion of redundant copies
- “Multiplicity” column to identify total # of copies

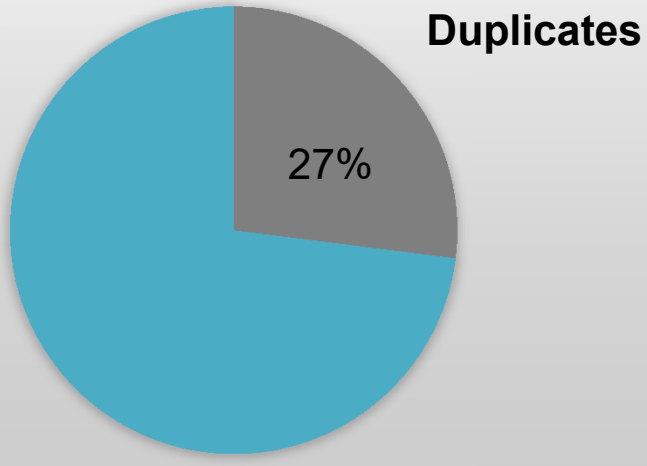


All		record_ID	001	Unique ID	Multiplicity	RECORD #(BIBLI	008 Date One	090|a	050|a	245|a	VOLUME	Cleaned Volum	260|b	264|b
21.	1000069_	1000069	1000069_	1	b1629516x	1953		QE524	The ancient volcanoes of Oregon			Oregon State System of Higher Education,	
22.	1000075_v 1	1000075	1000075_v 1	2	b12724269	1951		JC121	Marsilius of Padua, the defender of peace.	v. 1	v 1	Columbia University Press,	
		1000075	1000075_v 1	2	b12724269	1951		JC121	Marsilius of Padua, the defender of peace.	v.1	v 1	Columbia University Press,	
23.	1000075_v 2	1000075	1000075_v 2	2	b12724269	1951		JC121	Marsilius of Padua, the defender of peace.	v. 2	v 2	Columbia University Press,	
		1000075	1000075_v 2	2	b12724269	1951		JC121	Marsilius of Padua, the defender of peace.	v.2	v 2	Columbia University Press,	

Deduplication Method 1

- OCLC Number + Volume (as written)
  - OCLC number as identifier
    - Does not account for different formats or changes over time
  - Excluded blank or otherwise invalid identifiers in 001

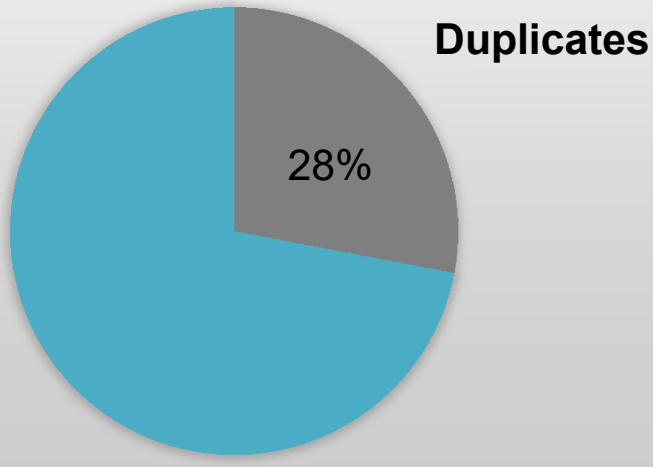
- Results
  - **27% redundant items**
  - Lowest bound of what may be duplicated
  - False uniques due to volume
    - “v.2” “2” “vol.2” etc.



Duplication Method 2

- OCLC Number + Volume (normalized)
  - Reconcile data entry messiness in volume
  - Tried different text normalization functions
    - Issues with word order and number handling
  - Use GREL and regular expressions to balance clean data & specificity of information

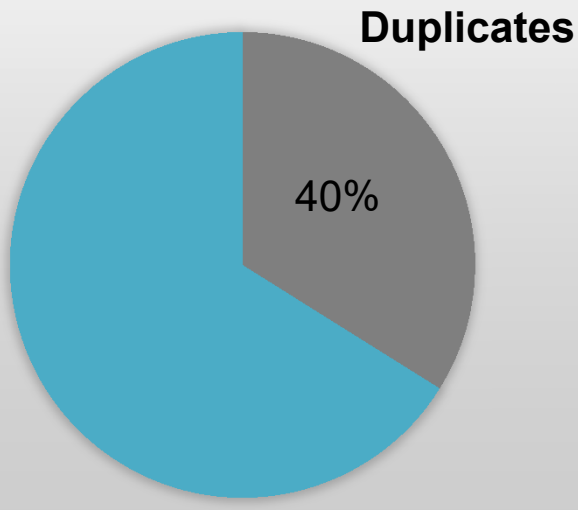
- Results
  - **28% redundant items**
  - Not much higher than 27% for non-normalized volume in Method 1



Duplication Method 3

- Title (normalized) + Volume (normalized)
  - Title used to get upper bound of possible duplication
  - Used fingerprint text normalization function on title
  - Used same volume normalization approach as Method 2

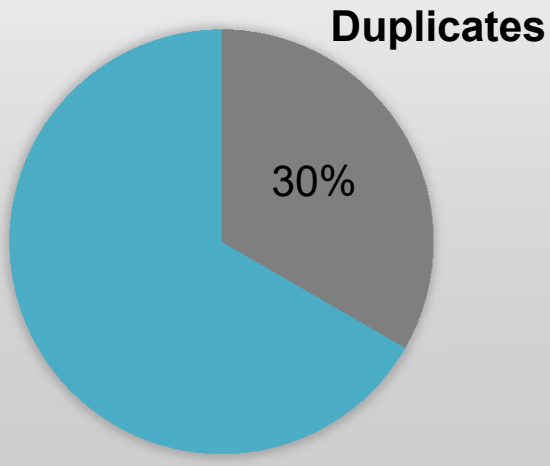
- Results
  - **40% redundant items**
  - Significantly higher than OCLC number estimates
  - Main title is insufficiently specific for a reliable estimate
    - Ignores edition and subtitle
    - Does not distinguish “Proceedings” etc.



Duplication Method 4

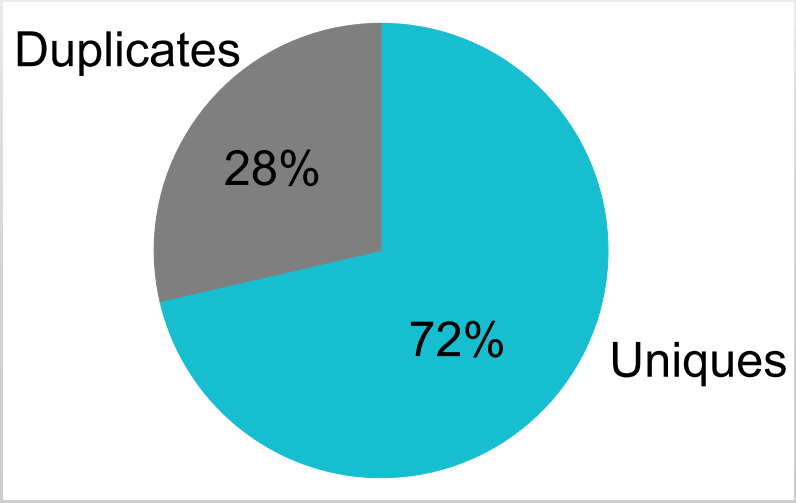
- Title (normalized) + Date + Volume (normalized)
  - Same title normalization as Method 3
  - Use 008 Date One
  - Same volume normalization as Methods 2 & 3
  - Do varying OCLC numbers skew Methods 1 & 2?

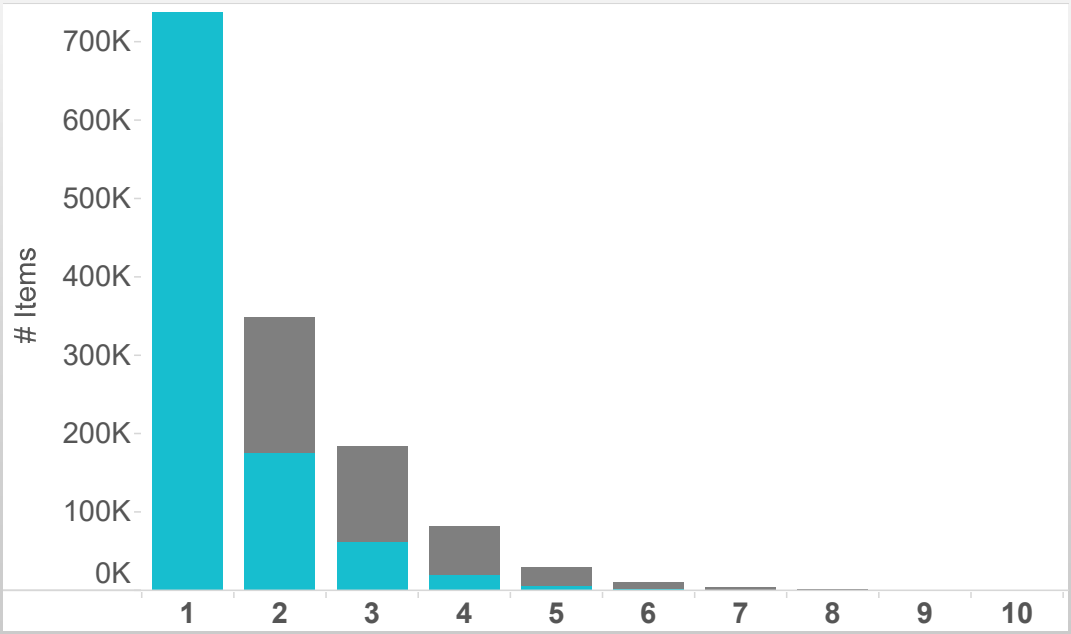
- Results
  - **30% redundant items**
  - Same issues as Method 3 with main title
  - Indicates variance in OCLC numbers is small factor
  - Indicates true duplication rate is in the 27%-30% range



Duplication Conclusions

- How many redundant monographs are in the OHDEP catalog?
  - Use Method 2: OCLC # + Volume (normalized)
  - 28% or approximately 399,000 items
  - Range of uncertainty 27% - 30%





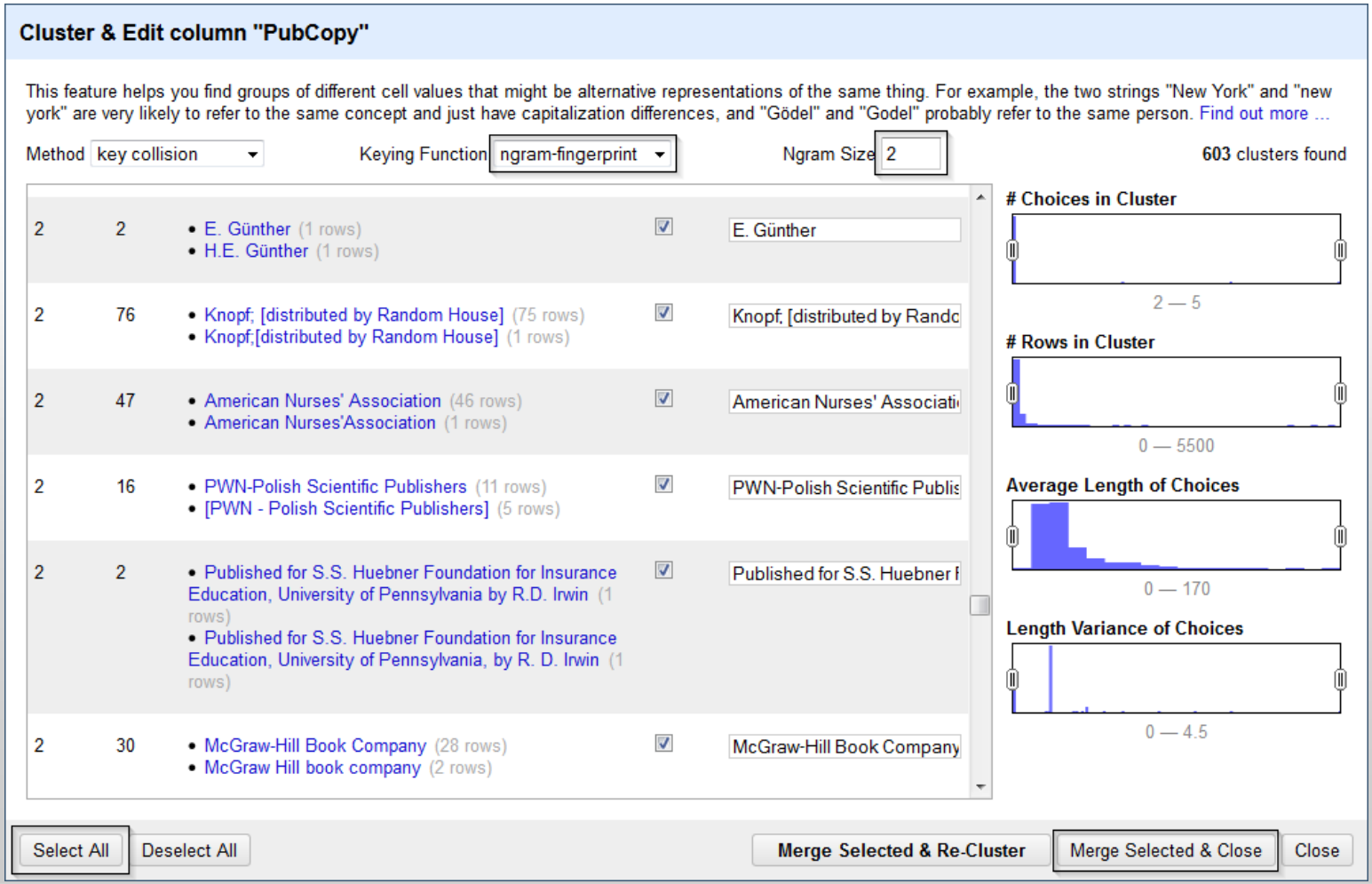
Publisher Analysis Method

- Question: Is this scholarly content?

- Limited by data in 260 or 264 fields
  - Data combined & normalized
  - Did not reconcile imprints to parent publishers
- Able to use text clustering features
  - Ngram-fingerprint clustering
  - Remove non-distinguishing words: “co”, “inc”, “ltd”
- Some items too messy to analyze
  - 2% had no publisher information
  - 17% had unusable or indistinguishable information

- 1,131,499 items in publisher results set

Cluster & Edit

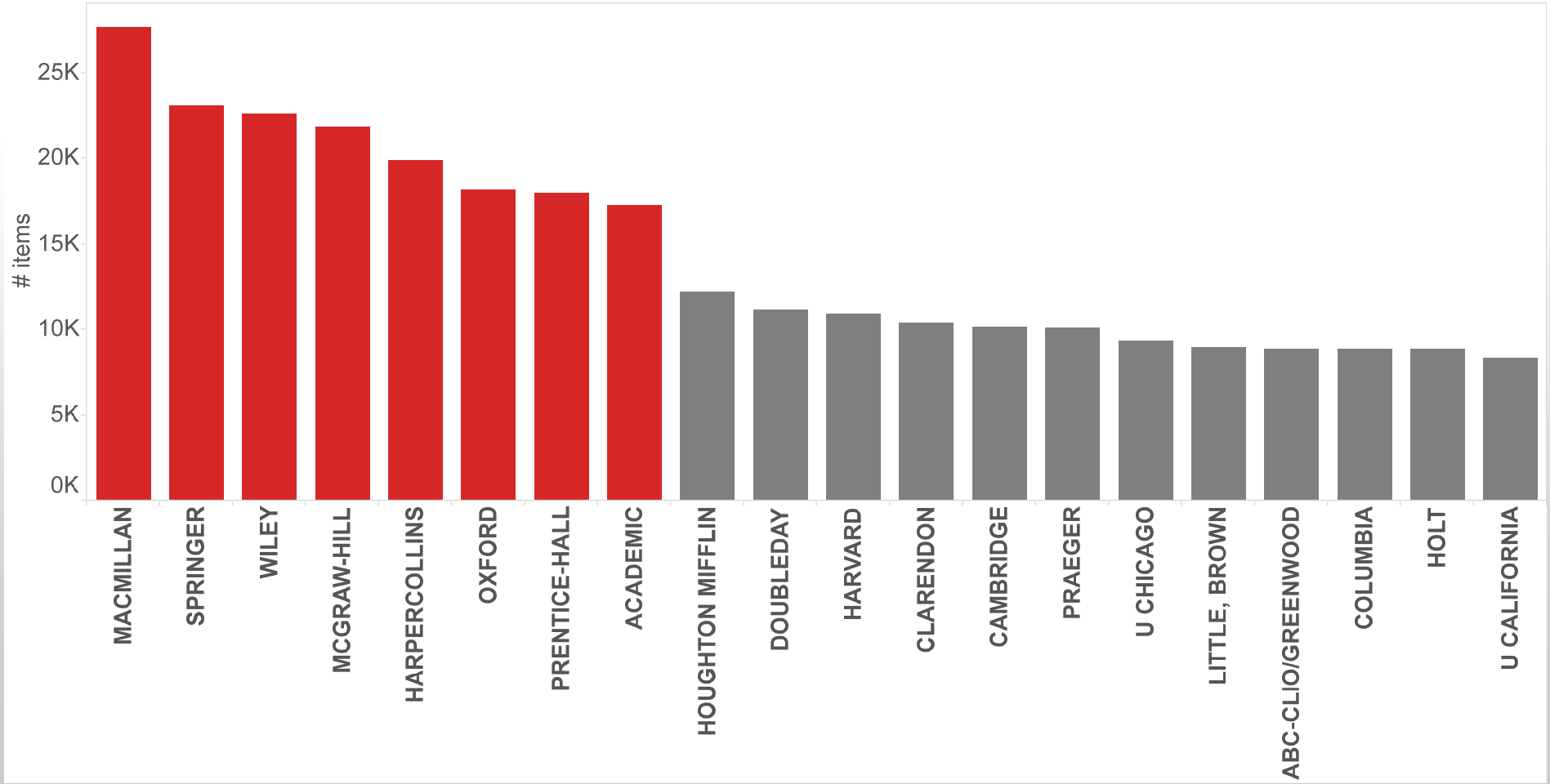


Publisher Analysis Results

- Top 8 publishers in academic market
  - Macmillan, Springer, Wiley, McGraw-Hill, HarperCollins, Oxford, Prentice-Hall, Academic
  - Stable across examination of variants
    - E.g. “Springer” vs. “Springer-Verlag”
  - Represent only 14% of items in publisher analysis (12% of all)

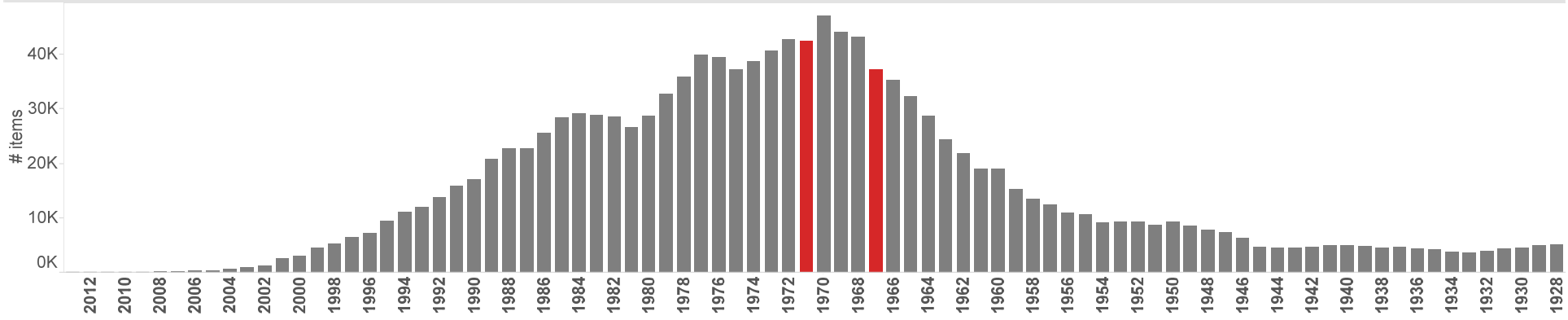
- Long tail of publishers with relatively few items:
  - Those with a few thousand items (37% of total)
  - Those with a few hundred items (23% of total)
  - Might be due to specificity of MARC data

Top Publishers



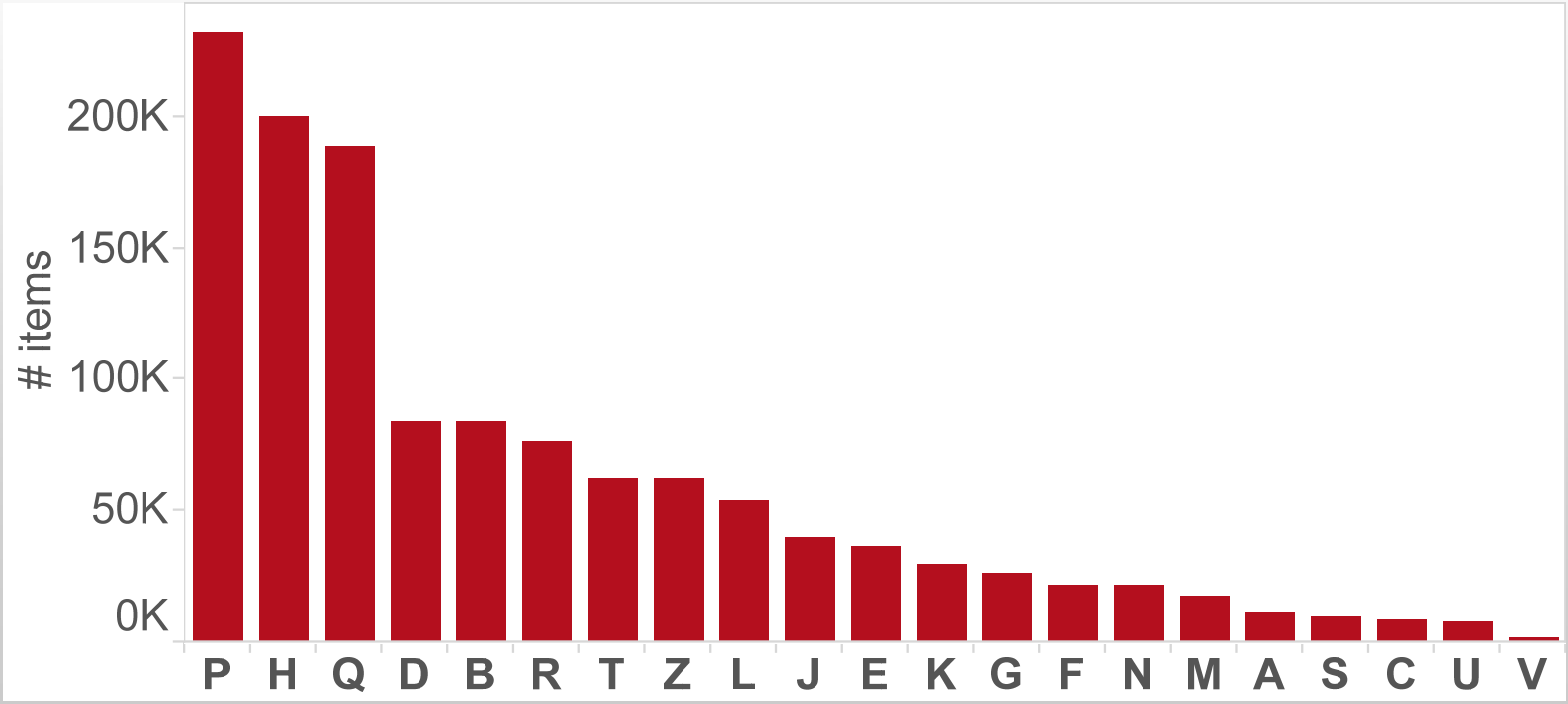
Year of Publication Analysis

- Exclude null or unusable 008 Date One fields
- 1,398,228 items included in analysis
- Most items mid-late 20th century
  - Average year: 1967
  - Median year: 1971
  - Long tail out to 1500



Subject Analysis

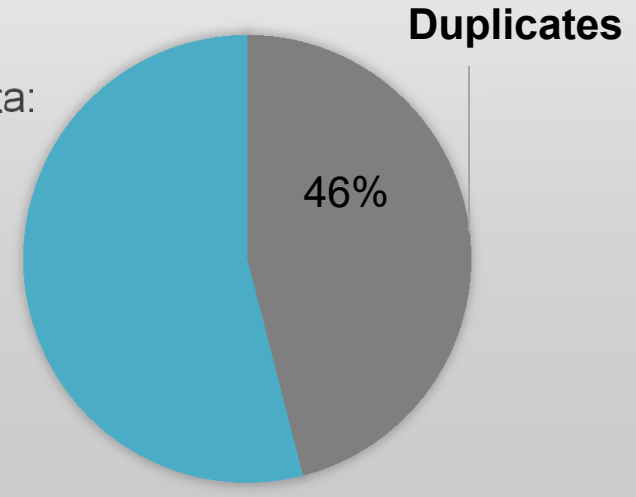
- Subject analysis based on call number in OHDEP
- 1,271,786 have an LC call number (91%)
- Most in Language & Literature, Social Science, Science



Results Checking

- OHDEP catalog circulation data is shallow
- Compare to 2007-2008 OCLC analysis data:
  - OhioLINK Collection and Circulation Analysis
    - http://www.oclc.org/research/themes/systemwide-library/ohiolink.html

- Look at 4,148,533 monograph items in Ohio depositories
  - More complete than OHDEP catalog data
  - Lacking item-specific information in circulation data:
    - Title, call number, volume
  - 46% duplication
    - High bound: volume, scale
  - 26% of items in depositories had circulated



Limitations & Further Work

- Limitations:
  - Relies on matching across the entire data set
    - Size limit (a few million) based on desktop hardware
    - Require specific deselection criteria to use for weeding
  - Serials excluded from this project
  - What is actionable at this scale?

- Follow-up Projects:
  - Can we use the central catalog data to examine duplication of depository items against the consortium collection?
  - Could shareable ebook backfiles be used to replace chunks of content in depositories?

Further Work

- Using Central Catalog data to look at duplication between depository items and the rest of the consortium collection

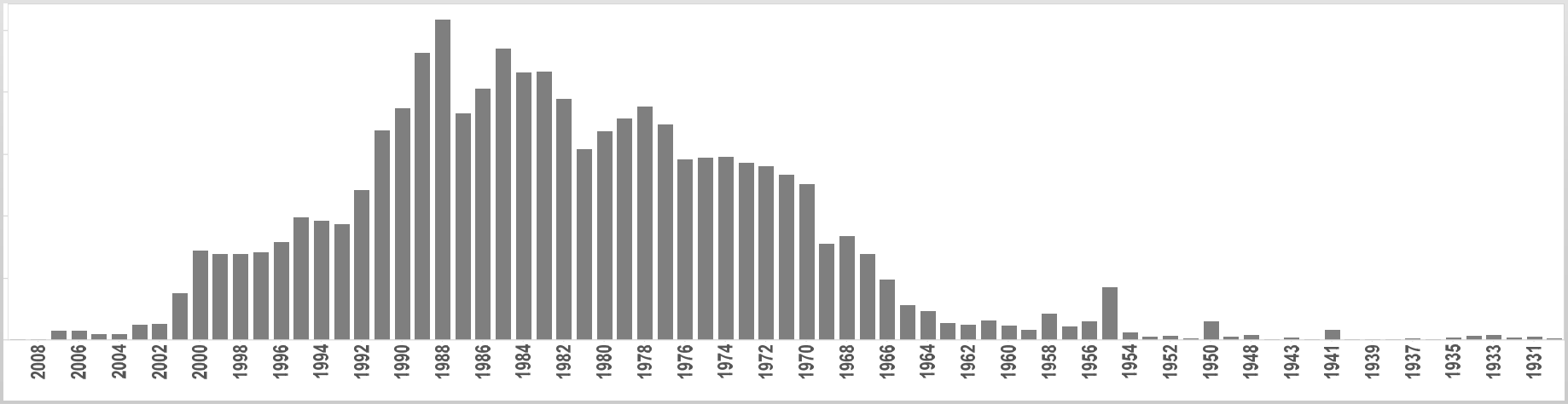
- Advantages:
  - Centralized, broad data
  - Matching algorithm (does the matching for us)
  - Possible using OpenRefine & techniques developed

- Disadvantages:
  - HUGE data set
    - Central has 11.7 million bib records and more than 46 million item records
  - No circulation information
  - What is actionable on this scale?

Follow Up with Central Catalog

- Follow-up work still in progress!
- Think about replacing print with ebooks
  - Use data from the Central Catalog
  - Check against OHDEP results

Springer books in OHDEP by publication year:



Action in the Depositories

- Serials
  - The depositories are already working on serial deduplication

- Monographs
  - This analysis provided reliable numbers to use in discussions of what is or isn't possible on the scale of the depository system
  - It provided a stepping stone to a larger analysis of all depository holdings
  - It also highlighted areas for further questioning (i.e. what level of duplication is acceptable?)

Conclusions

- OpenRefine expands what we can do with “small data”
  - Smaller scale projects can lead to larger scale projects
  - It can handle millions of lines of any text-based data
  - GUI allows interactive data manipulation

- Clearly this has helped OhioLINK with a collection development question that would have been too big to even approach without this tool
  - There is significant duplication in monographs in the Ohio Depository System
  - Is academic content, mid/late 20th century, and broad subject coverage
  - Can expand analysis outside the OHDEP catalog to include other locations
  - Developed methods and tools for similar analysis in the future

- Think about shared collection strategies & policies moving forward
- Still left with the challenges of coordinating action, staffing, …

Questions?

Joanna Voss
Collections Analyst
jvoss@ohiolink.edu

Theda Schwing
Coordinator, OhioLINK Catalogs
tschwing@ohiolink.edu