Bibliographic Records as Data: Making research use of our shared collections

UC Libraries Forum
October 27th-29th, 2021

Madelynn Dickerson
Head, Digital Scholarship Services, UCI Libraries

Danielle Kane
Computational Research Librarian, UCI Libraries
About the Project

The project started as a/an:

- Creative exploration!
- Demonstration to ourselves and others the ways that library bibliographic data can be useful in performing scholarly analysis of library collections
- Opportunity for collaboration between Digital Scholarship Services and Cataloging and Metadata Services departments
Collaborators

Madelynn Dickerson - Research Librarian for Digital Humanities and History
Joshua Hutchinson - Cataloging and Metadata Librarian
Danielle Kane - Computational Research Librarian
Sarah Wallbank - Electronic Resources and Serials Cataloging Librarian
Project Goals

- Identify areas where bibliographic data might effectively answer scholarly research questions
- Demonstrate what skills are necessary to make use of this data
- Learn those skills and tools
- Perform a sample analysis on the UCI Libraries history monograph collection
Sample Research Questions

We imagined a digital humanities researcher could have questions like these:

- Of all the history monographs in our catalog (books with the call numbers C-F), how many were written by women?
- What topics are women historians writing about?
Research Question Implications

Heading into the sample research question, we wanted to explore the following:

- What processes and tools would be required to make this analysis?
- What are the challenges and pitfalls?
- Is it even possible to accurately and ethically identify an author as a woman based on their name alone? How would (and should?) one go about doing this for the purpose of scholarly analysis?
- What other data is available in the bibliographic record that could be of scholarly interest (or of value for collection analysis?)
Tools Used

- **C# MARC Editor**: open source editor for Library of Congress MARC21 and MARCXML bibliography records.
- **OpenRefine**: an open-source desktop application for data cleanup and transformation to other formats.
  - **GREL functions**: (General Refine Expression Language) is designed to resemble Javascript. Formulas use variables and depend on data types to do things like string manipulation or mathematical calculations
- **Voyant Tools**: a web-based reading and analysis environment for digital texts.
Methods

1. Exported MARC records from Alma
2. Popped raw data into Voyant Tools to see what came out
3. Strategized how to clean and organize the data
4. Prioritized fields we were interested in
5. Divided data by decade, focusing on 1970s, 1980s, 1990s, 2000s
6. Cleaned and parsed data
7. Assembled a preliminary list of baby names by gender
Data Collected

History monographs (call numbers C-F)

- 184,103 items
- File size: 292 MB
- Author and title
- OCLC number (marc_35)
- Publisher, year, place (marc_260a-c)
- Any fields with subject terms
Casual Findings

<table>
<thead>
<tr>
<th>Rank</th>
<th>Term</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>john</td>
<td>5,312</td>
</tr>
<tr>
<td>2</td>
<td>robert</td>
<td>3,546</td>
</tr>
<tr>
<td>3</td>
<td>david</td>
<td>3,423</td>
</tr>
<tr>
<td>4</td>
<td>william</td>
<td>3,228</td>
</tr>
<tr>
<td>5</td>
<td>james</td>
<td>2,886</td>
</tr>
</tbody>
</table>
Casual Findings
## Casual Findings

<table>
<thead>
<tr>
<th>Term</th>
<th>Count</th>
<th>Length</th>
<th>Trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>world war</td>
<td>1743</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>world the</td>
<td>433</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>world of</td>
<td>274</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>world history</td>
<td>157</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>world a</td>
<td>142</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>world order</td>
<td>121</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>world politics</td>
<td>121</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>world in</td>
<td>106</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>world and</td>
<td>84</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
OpenRefine (cleaning data)

- Isolating publication dates
- Removing a’s and b’s from start of lines
- Removing terminal punctuation marks and extra spaces
- Separating out information from cells - such as language and country of publication from the 008 field
- Splitting names - sending first and last names into separate cells
- Removing diacritics
- Using GREL to extract information
- Pulling and matching data from other datasets - combining datasets
<table>
<thead>
<tr>
<th>RecordID</th>
<th>Author</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>Hu Hua deng zhuh</td>
<td>Zhou Enlai de si xiang ji li lun gong xian</td>
</tr>
<tr>
<td>9</td>
<td>Guarducci, Margherita</td>
<td>La cosiddetta Fibula Prenestina : elementi nuovi</td>
</tr>
<tr>
<td>10</td>
<td>Moscati, Sabatino</td>
<td>I gioielli di Tharros : origini, caratteri, confronti</td>
</tr>
<tr>
<td>11</td>
<td>[XII Settimana di studi aquileiesi, 30 aprile-5 maggio 1981]</td>
<td>Aquileia nei IV secolo</td>
</tr>
<tr>
<td>12</td>
<td>Eiseman, Cynthia Jones</td>
<td>The Porticello shipwreck : a Mediterranean merchant vessel of 415-365 B.C.</td>
</tr>
<tr>
<td>13</td>
<td>Baldi, Agnello</td>
<td>L'anatema e la croce : Ebrei e Cristiani in Pompei antica</td>
</tr>
<tr>
<td>14</td>
<td>Guarducci, Margherita</td>
<td>La cosiddetta Fibula Prenestina : antiquari, eruditi e falsari nella Roma dell'Ottocento</td>
</tr>
</tbody>
</table>
# Building a Name / Gender List

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Names (#)</td>
<td>3,196</td>
<td>4,818</td>
<td>7,502</td>
<td>7,502</td>
</tr>
<tr>
<td>Female (F)</td>
<td>1,342</td>
<td>1,639</td>
<td>2,794</td>
<td>2,290</td>
</tr>
<tr>
<td>Male (M)</td>
<td>11,939</td>
<td>13,379</td>
<td>17,997</td>
<td>12,022</td>
</tr>
<tr>
<td>Unisex (U)</td>
<td>81</td>
<td>346</td>
<td>676</td>
<td>583</td>
</tr>
<tr>
<td>Initial (I)</td>
<td>585</td>
<td>585</td>
<td>585</td>
<td>1,244</td>
</tr>
<tr>
<td>Blank</td>
<td>12,674</td>
<td>10,672</td>
<td>4397*</td>
<td>6497*</td>
</tr>
<tr>
<td>Not in List (NIL)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>2,420</td>
</tr>
<tr>
<td>Title</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>1</td>
</tr>
<tr>
<td>Unidentified</td>
<td>--</td>
<td>--</td>
<td>172</td>
<td>33</td>
</tr>
</tbody>
</table>
Practical Takeaways

- Lots of bibliographic fields contain interesting information
- Bibliographic data is nicely structured but requires significant cleaning and massaging
- Collaborative data cleaning is hard!
- Not enough data in our dataset to make any broad conclusions about our sample research question
- No clear path forward on determining author gender unless authors self-identify
Future Potential

- Thinking in a “collections as data” context, what would it mean to offer “library catalogs as data”?  
  - Ex: Harvard Library API’s and Datasets  
    https://library.harvard.edu/services-tools/harvard-library-apis-datasets

- Potential to bring together more comprehensive data through UC’s combined catalog

- Opportunities to incorporate more nuanced search and analysis tools into our discovery systems, especially incorporating text analysis and data mining functionality
Examples of Bibliographic Data Analysis Platforms

● Harvard Library API’s and Datasets
  https://library.harvard.edu/services-tools/harvard-library-apis-datasets

● HathiTrust Research Center Analytics
  https://analytics.hathitrust.org/

● JSTOR’s Constellate Platform
  https://constellate.org/

● MIT Libraries’ List of Scholarly Publishing APIs
  https://libraries.mit.edu/scholarly/publishing/apis-for-scholarly-resources/
Resources

- Summary of transformations to decades
- First name and gender list
- LC countries list

Publications


Contact

Madelynn Dickerson
mrosed@uci.edu

Danielle Kane
kaned@uci.edu