Collecting, Cleaning and Using Bibliographic Data to Perform a Large-Scale Assessment Project on University Library Collections

Joshua Hutchinson, UCI Libraries, University of California, Irvine
jchutchi@uci.edu

Abstract

This poster describes the process by which librarians at the University of California, Irvine have collected, cleaned and used bibliographic data (which normally lives in the library's catalog) in order to perform a large-scale assessment project on their collections.

Introduction

This project was conceived of as a means to demonstrate to academics outside of the library profession that library bibliographic data can be useful in performing scholarly analysis of library collections.
The project goals are to:

- Identify areas where bibliographic data can effectively serve as an assessment tool;
- Demonstrate what skills are necessary to make use of this data;
- Learn those skills and tools;
- Perform an assessment project on the UCI Libraries collection.

Team members

- Joshua Hutchinson – Cataloging and Metadata Services
- Sarah Wallbank – Cataloging and Metadata Services
- Danielle Kane – Digital Scholarship Services
- Madelynn Dickerson – Digital Scholarship Services

Tools and skills

The members of this team made use of a variety of tools and skills for this project. These include:

- Alma LSP
- C# Marc Editor
- OpenRefine
- Microsoft Excel
- Voyant Tools
- GREL, Excel formulas and basic knowledge of linked data have been the most well-used skills.

Methodology

This project has been undertaken using the following methodology:
Using the Alma LSP, bibliographic records for all printed books in the LC classification areas for history were exported and converted to spreadsheet format. The data was cleaned using Excel and OpenRefine, while initial research was conducted using Voyant Tools. This helped to demonstrate that it was worth continuing with this research– it is possible to draw conclusions from this corpus of data.
Once the database cleanup was complete, each member of the team was assigned a selection of the records (divided chronologically) in order to be able to work more closely with the data and draw conclusions, as well as to improve our individual skills with OpenRefine. Having familiarized themselves with the data, the team was able to refine the scope of this project, ensuring that the project goals were aligned with the level of data available.

The initial focus of this assessment was to compare fields from the catalog records over time in order to determine whether UCI's current collections change based on their year of publication– or whether they didn't.

One example of this is examining whether UCI has diversified our collections over time, or whether our collections have maintained stability in terms of place of publication, gender of the author, and in other areas: whether we're predominantly purchasing from university presses; whether we consistently collect about American and European history; and whether we are collecting books written by people with male and anglophone names. In all three of these examples, the word clouds at right show that bibliographic metadata is a useful tool for assessing the UCI Libraries' History collection.

Data from selected MARC fields

Recurrence of words – Publisher field

- 1970s:

- 1990s:

Recurrence of words – Title field

- 1970s

- 1990s

Recurrence of words – Author field

- 1970s

- 1990s

Note that this is imperfect. It uses the Statement of responsibility rather than 100/700 fields, the stopwords that I used were just from me glancing at the results.

Further Projects

The team is currently working further with names, using Open Refine to identify gender in the authors' names. We are working through a variety of different methods to achieve this, including different formulas in OpenRefine, adding names to a master list of baby names, and determining which part of the bibliographic record is best to use for this part of the project.

Looking at one of the sets of results, for a group of 26,621 records, the team was able to get results for 21,810 (82%) by comparing the list of authors with a list of gendered given names, in order to make a rough approximation of the gender of the authors of these books. We recognize that making assumptions about gender from names says very little; our focus for this project is to show that collection assessment research using bibliographic data is possible.

Findings

Interim findings were published in the CILIP Cataloging & Index journal dated December 2019. These findings include the conclusion that these data are appropriate to use and yield clean and valid results.