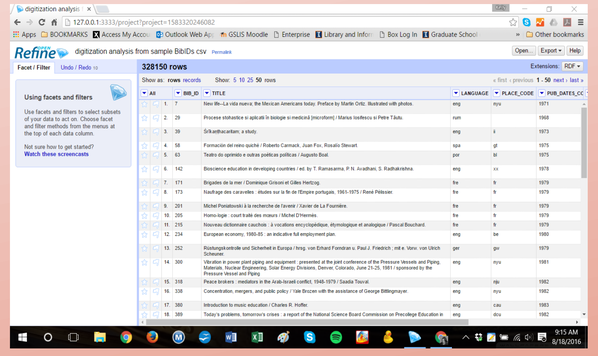
Buried Treasure: How a Deep Data Dive Can Uncover Global Language Gems

Kelly J. Applegate
Illinois School of Information Sciences, University of Illinois at Urbana-Champaign

Introduction

Digitization projects are still a lively topic among academic libraries, although small in-house projects have become overshadowed by large scale partner projects like Google Books and the Internet Archive. By using the filtering power of Open Refine, leaders at the helm of their university libraries can put together collections of unique texts written in lesser used languages to bring to their digitization departments. By doing so, they can make scholars of less-studied languages more aware of the resources in their collections.

Library research projects based on integrated library system reports produce enormous data sets that can make finding specific items difficult. But such data sets can be imported into Open Refine (OR), a web-based, open source desktop application used for cleaning data. One of the powerful tools of OR is it's faceting and filtering capabilities. By using this software on bibliographic based data sets, a greater level of granularity can be achieved.



Example Data Set from the University of Illinois Library

ILS REPORT
- Started with a Voyager report of 328,150 records generated from BibIDs

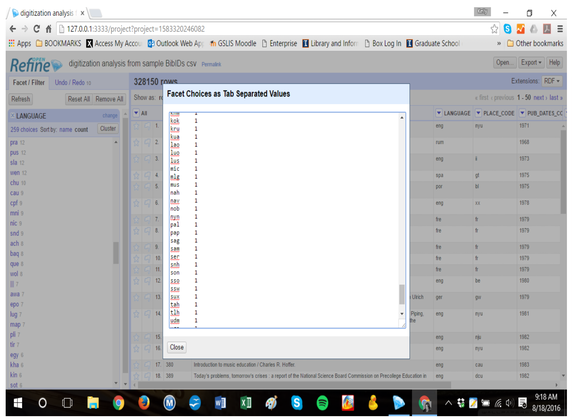
DISTILLED CSV SPREADSHEET
- Repurposed that data into a CSV of titles, dates, and MARC 008 language codes

OPEN REFINE
- And used Open Refine to drill down

Findings

328,150 records were associated with 250 languages in total. 53 of those languages were associated with only ONE item. These are the gems in the stacks!

This data paints a picture that can and should inspire more digitization projects among academic language libraries.



Discovered Languages Perfect for Digitizing

Libraries with rich language collections haven't been in the spotlight of renowned digitization projects, but with a data dive into their own integrated library system reports, and filtering with the power of Open Refine, they could be!

What's So Interesting About These Languages?

Virtually all languages have a rich and complex history. Here are some of the highlights from the languages discovered in the sample project data deep dive analysis.

Akkadian

was a semitic language spoken in Mesopotamia (modern Iraq and Syria) between about 2,800 BC and 500 AD.

Bokmål

Literally "book tongue" is an official written standard for the Norwegian language, alongside Nynorsk. Bokmål is the preferred written standard of Norwegian for 85–90% of the population in Norway.

Oirat

In Mongolia, there are seven historical Oirat dialects, each corresponding to a different tribe. Scholars differ as to whether they regard Oirat as a distinct language or a major dialect.

Pahlavi

Pahlavi denotes a particular and exclusively written form of various Middle Iranian languages. The earliest attested use of Pahlavi dates to the reign of Arsaces I of Parthia (250 BC) in early Parthian coins with Pahlavi scripts.

Mapuche

is spoken by between 240,000 and 700,000 people in parts of southern Chile and western Argentina. It is a language isolate unrelated to any other language.

Chechen

Before the Russian conquest, most writing in Chechnya consisted of Islamic texts and clan histories, written usually in Arabic but sometimes also in Chechen using Arabic script. Those texts were largely destroyed by Soviet authorities in 1944.

Micmac

Like many Native American languages, Micmac uses a classifying system of animate versus inanimate words. However, while the animacy system in general is common, the specifics of Micmac's system differ from even closely related Algic languages: for instance, in Wampanoag, the word for "sun", *cone*, is inanimate, while the word for "earth", *ahkee*, is animate, a fact used by some scholars to claim that the Wampanoag people were aware of the earth's rotation around an unmoving sun.

Who Can Benefit from Digitizing These Language Treasures?

- ❑ **Scholars who have limited access to source analog material**
- ❑ **Languages, linguistics, translation, and philology students**
- ❑ **University students who attend a school that doesn't have a languages library**
- ❑ **Anyone in addition to students looking for books in hard to find languages**
- ❑ **People researching their own heritage's language and literature**
- ❑ **Those interested in languages as a hobby**

Conclusions

Finding language gems in your library is just *one* of many potentially useful things library leaders can do with Open Refine. This poster was meant to be one of any number of possibilities. Use OR to filter dates, authors, countries of publication, subject headings, and more.

With filtering and faceting power that goes far beyond programs like Excel and Open Office, OR is a new skill that can be taught with relative ease to library staff who work with everyday library cataloging based reports. Make Open Refine an essential tool at your library!

Acknowledgments

The author wishes to acknowledge Jennifer Teper, Head of Preservation Services at the University of Illinois, who allowed me to utilize Voyager reports and their corresponding Open Refine data from another research project.

School of Information Sciences
The iSchool at Illinois