Academic BRASS

Published by the
BRASS Business Reference in Academic Libraries Committee

Vol 9(2), Fall 2014

Ilana Stonebraker
Business Information Specialist and Assistant Professor of Library Science
Parrish Library of Management and Economics
Purdue University

Good Library Data Made Better With Technology! Using OpenRefine and Google Fusion Tables in Academic Business Libraries Instruction

Introduction

Big data just seems to get bigger all the time, but that doesn't mean it gets any less messy. Even large, carefully curated government datasets suffer from irregularities like acronyms, open response items, and misused categories. Steadfast librarians have the patience for such inaccuracies, but undergraduate students are often unprepared for the realities of the big data they crave. Teaching data cleaning and collaboration can help students better understand and use large datasets but also illustrate the importance of library-curated data, as it often has fewer of these problems than datasets found on the open web. At a high level, library data and open datasets maybe seem comparable, but when we give students the tools to look through the data on their own, the small things start to add up.

This short article will discuss shifting the focus of a one-shot instruction away from datasets to helping students do this deeper dive, using the data tools Google Fusion Tables and OpenRefine. It is my hope to make librarians more aware of these two data tools and how they can be integrated in business instruction. I first learned about Google Fusion Tables and OpenRefine from attending the Ann Arbor Data Dive in 2012. Data Dives are nonprofit weekend intensive data events. Other tools used by Data Dives are available at http://opendata-tools.org/en/. As business librarians continue to tackle issues of student data collection and reuse, approaches and tools like this one can help illustrate the importance of good sources and methods.

Setting

During the spring 2015 semester, the library was approached by a professor of an Electronic Commerce and Information Strategies course. The professor, who had worked with the libraries in the past, developed a final assignment where groups of students found and statistically analyzed a dataset to solve a business problem. The students, all of whom were in the upper division of the management program, were given a large amount of freedom to choose any business problem that interested them. The professor invited the libraries to give a short presentation (20 minutes) to the students about library resources.

To prepare for the course, my colleague and I made a library guide of potential sources of data. Given the breadth of the project and the collaborative nature of the course, the time in class seemed best spent focusing on tools likely useful in all cases and not on the large amount of data sources the students could potentially use. The libraries short session focused on Google Fusion Tables and OpenRefine.

Google Fusion Tables

Google Fusion Tables is a web-hosted data management product (see Figure 1). Fusion was first introduced in a scientific paper in 2010 (Gonzalez et al., 2010). Google Fusion tables can be accessed through Google Drive, Google Sheets, or from https://www.google.com/fusiontables/. Like Google Docs, Google Fusion Tables allows for seamless multiple-user editing of spreadsheets, but also includes many basic visualization tools as well. A critical additional feature of Google Fusion Tables is the ability to "fuse" or stitch two tables together if they share a common business information column such as zip-code, city, or age.

Fusion Tables was attractive to me for several reasons. First, I knew some of the students were looking to collect their own data, and I knew they would probably use Google Docs for that. Second, I knew many datasets had common attributes that would allow for seamless merging. Third, the interface is very simple and easy to learn. Students already have Google accounts and so do not need to sign up for any service.

Figure 1 Google Fusion Tables example data. I pulled the data from Wikipedia and LexisNexis and limited by temperature and country.

OpenRefine

OpenRefine is an open source desktop application for data cleanup and transformation in-browser. It is similar to many aspects of Microsoft Excel, however, it functions more like a database. For those working with large datasets, OpenRefine can help the student clean up small issues such as acronyms, spelling errors, etc. Several videos about how it works are available here: http://openrefine.org/ OpenRefine allows imports from Excel and CSV files as well as Google Fusion Tables.

OpenRefine is attractive in an instruction scenario because it helps students visualize the messy characteristics of large datasets. The state of Indiana may be listed as "IN" in some places and "INDIANA, State of" in others, but both will show up in OpenRefine's list.

Putting it all Together in a One-Shot Session

To prepare for the short (under 20 minute) session, [illegible]

The short session highlighted two major business information issues likely to plague the students: messiness and collaboration. The conversations that followed were very different from other conversations I have had in the past. Students had questions about the things I had shown, but were also more aware of the messiness of the data. When I had showed them high level versions of the many datasets the library suggested, they had not seemed much different to the students compared to what they could find searching around. But now that the conversation was more about messiness versus access, the library datasets were a lot more attractive.

Conclusion

This short article has covered two tools, Google Fusion Tables and OpenRefine, and their use in business undergraduate library instruction. Google Fusion Table offers collaborative and data merging features in a familiar environment to students. OpenRefine is a powerful yet simple data cleaning option. These tools highlight ways libraries can provide value to their datasets for novice big-data analysts in ways that amplify the traditional role of access and preservation.

Works Cited:

Gonzalez, H., Halevy, A., Jensen, C. S., Langen, A., Madhavan, J., Shapley, R., & Shen, W. (2010). Google fusion tables: data management, integration and collaboration in the cloud. In *Proceedings of the 1st ACM symposium on Cloud computing* (pp. 175– 180). ACM.