

Purdue University Purdue e-Pubs

Libraries Faculty and Staff Scholarship and Research

Purdue Libraries

7-9-2015

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

Ilana Stonebraker

Purdue University, stonebraker@purdue.edu

Follow this and additional works at: http://docs.lib.purdue.edu/lib_fsdocs

Recommended Citation

Stonebraker, Ilana, "Good Library Data Made Better With Technology! Using OpenRefine and Google Fusion Tables in Academic Business Libraries Instruction" (2015). *Libraries Faculty and Staff Scholarship and Research*. Paper 118.
http://docs.lib.purdue.edu/lib_fsdocs/118

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

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 cultivated 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-cultivated 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 may seem comparable, but when we give students the tools to slog 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 help students do this deeper dive, using the data tools Google Fusion Tables and OpenRefine. It is my hope to make librarians more aware of the 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 librarians 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 librarians 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.

Merge of data and Alabama - Wikipedia, the free encyclopedia

en.wikipedia.org - imported via Fusion Tables search - Edited on February 26, 2015

binartames@gmail.com

File Edit Tools Help Rows 1 Cards 1 Map of State/Province...

Country = 'United States' AND Jan IN (60.7; (15.9); (48.9; (9.4); (52.8; (11.6); (57.6; (14.2))

Country 1-71 of 71

Jan 9 distinct values

- 35.5; (1.9) 23
- 39.5; (4.2) 14
- 48.9; (9.4) 11
- 52.8; (11.6) 23
- 57.6; (14.2) 23
- 60.7; (15.9) 14

City	Company Name	Ticker	Street Address	State/Province	Postal Code	Country	Telephone Number	Company Type	No. ...	Primary SIC Code	Sales/Revenue	Month	Jan
Birmingham			509 40th St S	Alabama	35222-3300	United States	205-226-4000	PRIVATE - PARENT	18	8231 Libraries	700,000	Average high	52
Birmingham			2009 Little Ridge Cir	Alabama	35242-2813	United States	205-980-6441	PRIVATE - PARENT	2	8231 Libraries	77,000	Average high	52
Birmingham			5237 Meadow Garden Ln	Alabama	35242-3412	United States	205-991-8195		3	8231 Libraries	110,000	Average high	52
Birmingham			800 Lakeshore Dr	Alabama	35229-0001	United States	205-726-2558		3	8231 Libraries	90,000	Average high	52
Birmingham			1018 Oxmoor Rd	Alabama	35209-5318	United States	205-637-0735	PRIVATE - PARENT	3	8231 Libraries	110,000	Average high	52
Birmingham			P O Box 361966	Alabama	35236	United States	803-649-4952	PRIVATE - PARENT	3	8231 Libraries	100,000	Average high	52
Birmingham			800 Lakeshore Dr	Alabama	35229-0001	United States	205-726-2846	PRIVATE - PARENT	3	8231 Libraries	150,000	Average high	52
Birmingham			505 20th St N Ste 225	Alabama	35203-4651	United States	205-443-0670	PRIVATE - PARENT	7	8231 Libraries	475,000	Average high	52
Birmingham			2601 Carson Rd	Alabama	35215-3007	United States	205-856-8524	PRIVATE - PARENT	3	8231 Libraries	95,000	Average high	52
Birmingham	Jefferson County Library Cooperative Inc		2100 Park Pl	Alabama	35203-2744	United States	205-226-3615	PRIVATE - PARENT	37	8231 Libraries	938,974	Average high	52
Birmingham			800 Lakeshore Dr	Alabama	35229-0001	United States	205-726-2714		3	8231 Libraries	110,000	Average high	52
Birmingham			400 Breland Dr	Alabama	35228-2732	United States	205-923-1027	PRIVATE - PARENT	1	8231 Libraries	54,000	Average high	52
Birmingham			33 Olmsted St	Alabama	35242-1826	United States	205-991-1660	PRIVATE - PARENT	3	8231 Libraries	83,000	Average high	52
Birmingham			5521 Cahaba Valley Rd	Alabama	35242-4901	United States	205-439-5500	PRIVATE - PARENT	23	8231 Libraries	1,044,184	Average high	52
Birmingham			5521 Cahaba Valley Rd	Alabama	35242	United States	(205) 439-5500	PRIVATE - PARENT	23	8231 Libraries	1,100,000	Average high	52
Birmingham			1530 3rd Ave S	Alabama	35294-0002	United States	205-934-4475	PRIVATE - PARENT	5	8231 Libraries	220,000	Average high	52
Birmingham			2100 Highland Ave S	Alabama	35205-4083	United States	205-933-8037	PRIVATE - PARENT	13	8231 Libraries	960,000	Average high	52
Birmingham	The Birmingham Public Library Foundation		2100 Park Pl	Alabama	35203-2744	United States	205-226-3610	PRIVATE - PARENT	3	8231 Libraries	1,908	Average high	52
Birmingham			700 19th St S 124	Alabama	35233-1927	United States	205-930-9912		5	8231 Libraries	180,000	Average high	52
Birmingham			1301 Tiff Ave S	Alabama	35294-0001	United States	205-934-3555	PRIVATE - PARENT	3	8231 Libraries	110,000	Average high	52
Birmingham			101 Oxlin Cir	Alabama	35211-5965	United States	205-944-3900		3	8231 Libraries	110,000	Average high	52

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 that opens 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 acronym issues, spelling errors, etc. Several videos about how it works are available here: <http://openrefine.org/> OpenRefine allows input 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, I pulled together a small dataset, combining library locations with the average temperatures of cities in Alabama (see Figure 1). I first showed students the library guide, then Google Fusion Tables, then the data merge I created, and finally showed them how I could use Fusion Tables to clean up the data (for example, removing the degree symbol from temperatures so they could be averaged). I mentioned OpenRefine, but if I were doing the instruction section again, I would show them more in depth so that they could see how they might use this to clean up data a little faster than Google Fusion Tables. I then shared my contact information and calendar if they had questions about the resources I showed or any of the datasets that were on the library guide page.

The short session highlighted two major business information issues likely to plague the students: messiness and collaboration. The consultations that followed were very different from other consultations 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 student 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.