Automatically assessing the relevance, quality, and usability of spreadsheets

Thomas Levine
_@thomaslevine.com

ABSTRACT

When you have lots of spreadsheets, it gets hard to look through all of them. In my research, I have been exploring methods for understanding the contents of thousands of spreadsheets at once. I will discuss strategies for automatically assessing the usability of spreadsheets, the quality of data in spreadsheets, and the relevance of specific spreadsheets to particular analysis questions; I will explain both how these methods work and how they can help you you manage and analyze your spreadsheets.

Keywords

data management, spreadsheets, open data

1. INTRODUCTION

Through recent open data initiatives, governments and large organizations have begun releasing much of their internal data as public files on the internet. From just a few different websites, we can easily download 100,000 different spreadsheets [3]. My research began with this question: What becomes possible when we use quantitative methods to look at 100,000 different datasets at once?

The research has since focused more specifically on trying to understand what is going on in data-sharing ecosystems. I've been collecting data about publicly shared open data and looking for patterns in publishing and usage across the datasets. Here are three specific issues that I look at.

1. How can I find datasets that I care to see?
2. What can I do about incomplete metadata?
3. Can we quantify how good a particular dataset is?

I'll briefly discuss some related work by other people and explain how I acquire lots of spreadsheets; then I'll review some of my findings in the above three areas.

2. RELATED WORK

Many other people have used relatively qualitative means to make sense of the release of diverse open data spreadsheets. For example, Open Knowledge Foundation volunteers manually looked through many websites to assemble a census [18] of the availability of key datasets released by different countries.

McKinsey [15] and the Governance Lab [16, 2] have looked at how specific businesses use specific publicly available spreadsheets.

The general approach in these various studies is to look in depth at how a few datasets are used, or how data-related projects are run. My research, on the other hand, tries to get a broad picture across many different spreadsheets.

3. ACQUIRING LOTS OF SPREADSHEETS

Data catalogs make it kind of easy to get a bunch of spreadsheets all together. The basic approach is this.

1. Get all of the dataset identifiers.
2. Download the metadata document about each dataset.
3. Download data files about each dataset.

I've implemented this for the following data catalog softwares.

- Socrata
- CKAN
- Junar (partially)
- OpenDataSoft

This allows me to get all of the data from most of the open data catalogs I know about.

Most of these spreadsheets are represented as tables, where rows correspond to records and columns correspond to variables. [5]

After I've downloaded spreadsheets and their metadata, I assemble them into a spreadsheet about spreadsheets. [4]

In this super-spreadsheet, each record corresponds to a full sub-spreadsheet; you could say that I am collecting features or statistics about each spreadsheet.

4. FINDINGS

Here are some of the things I've found by looking at lots of spreadsheets at once. I think of them as ways of automating some of the early steps in data analysis, but I'm packaging them into the three categories I mentioned above (looking for datasets, dealing with metadata problems, and quantifying quality).

4.1 New ways of looking for datasets

We search for prose by typing prose into a search bar; why don't we search for spreadsheets by typing spreadsheets into a search bar? Spreadsheets are much more structured than arbitrary prose, and we can use this structure to enable new search paradigms. We could search for things like the following.

- Spreadsheets that were produced by the same program as another spreadsheet
- Spreadsheets that I can join to a particular spreadsheet
- Spreadsheets with a particular statistical unit
- Spreadsheets in long format (rather than wide format)

One example is the detection of spreadsheets that can be stacked on top of each other (unioned). My work on this started with AppGen,[13] which was a system to generate random apps based on randomly combined datasets. I combined spreadsheets by matching spreadsheets with the same column headers.

Spreadsheets with the same column headers seemed to be semantically related to each other. For example, there were 23 spreadsheets with these same columns.

- type_of_abuse_of_authority_allegation
- substantiated_number
- sunstantiated_rate
- exonerated_number
- exonerated_rate
- unsubstantiated_number
- unsubstantiated_rate
- unfounded_number
- unfounded_rate
- officer_unidentified_number
- officer_unidentified_rate
- miscellaneous
- miscellaneous_rate

All of these spreadsheets were uploaded by the same person, and they all had titles of the form "${crime} Allegations ${year}", such as "Disposition Of Force Allegations 2006".

When we organize spreadsheets by their column headers, groups of related spreadsheets pop out at us.

4.2 Dealing with incomplete metadata

People complain about how data are bad and metadata are bad. Rather than fixing it on a case-by-case basis, I think we should just come up with ways of dealing with it. As an example, let's talk about unique identifiers.

I don't know of anyone who specifies within a spreadsheet file which columns should be unique. There are methods with external metadata files, such as data packages [19], but hardly anyone uses those either. Rather than trying to get people to write down which columns are unique, we can just figure it out for ourselves.

special_snowflake [9] is a package that does just that. It looks at all combinations of columns within a particular spreadsheet and determines which combinations function as unique identifiers.

With tools like special_snowflake, we don't have to rely as much on other people for the creation of accurate metadata; we can lazily figure out the metadata ourselves.

4.3 Quantifying data quality

A bunch of people [18, 1, 20, 14, 17] have come up with relatively qualitative guidelines for publishing data.

I've been exploring ways of assigning numbers to represent data quality. Using only the simple metadata files from open data catalogs, I've come up with automated approaches for describing the updating, [10] licensing, [8] size [6], file format [11] and availability [7, 12] of groups of datasets.

By quantifying these guidelines about data quality, we can more quickly and precisely assess data quality.

5. APPLICATIONS

A couple of people can share a few spreadsheets without any special means, but it gets hard when there are more than a couple people sharing more than a few spreadsheets. In my research, I'm coming up with approaches for assisting this sharing.

Software for the publishing and analysis of data can integrate new search paradigms to assist people in finding relevant datasets or enriching existing datasets.

The ubiquitous and persistent complaint of bad metadata can be tempered through the use of tools that infer metadata or provide alternative search strategies; by becoming more robust to bad metadata, we make available a broader range of datasets.

Quantification of the quality of data can be helpful to those who are tasked with cataloging and maintaining a diverse array of datasets. Data quality statistics provide a quick and timely summary of the issues with different datasets and allow for a more targeted approach in the maintenance of a data catalog.

6. FUTURE

Many people say that releasing open data will create transparency in government, engage citizens in their governments, and stimulate the economy. This all seems reasonable, but I haven't seen much empirical research that suggests that the sharing of data affects any of these outcomes.

As I come up with more ways of computationally describing datasets, I am becoming more able to apply various quantitative analyses to the dataset of datasets; this is starting to enable weaker versions of the aforementioned measure of outcomes. It would be nice to know whether releasing your organization's data on the internet will get people to use them.

7. REFERENCES

[1] T. Berners-Lee. Linked data. http://www.w3.org/DesignIssues/LinkedData.html, 2006.

[2] J. Gurin. Open data now. http://www.opendatanow.com/, 2014.

[3] T. Levine. 100,000 open data across 100 portal. http://thomaslevine.com/!/data-about-open-data-talk-december-2-2013/, 2013.

[4] T. Levine. Open data had better be data-driven. http://thomaslevine.com/!/dataset-as-datapoint, 2013.

[5] T. Levine. What's in a table? http://www.datakind.org/blog/whats-in-a-table/, 2013.

[6] T. Levine. Analyze all the datasets. http://thomaslevine.com/!/socrata-summary/, 2014.

[7] T. Levine. Dead links on data catalogs. http://thomaslevine.com/!/data-catalog-dead-links/, 2014.

[8] T. Levine. Open data licensing. http://thomaslevine.com/!/open-data-licensing/, 2014.

[9] T. Levine. special_snowflake. https://pypi.python.org/pypi/special_snowflake, 2014.

[10] T. Levine. Updating of data catalogs. http://thomaslevine.com/!/data-updatedness/, 2014.

[11] T. Levine. What file formats are on the data portals? http://thomaslevine.com/!/socrata-formats/, 2014.

[12] T. Levine. Zombie links on data catalogs. http://thomaslevine.com/!/zombie-links/, 2014.

[13] T. Levine and A. Williams. Appgen. http://appgen.me/, 2013.

[14] C. Malamud, T. O'Reilly, G. Elin, M. Sifry, A. Holovaty, D. X. O'Neil, M. Migurski, S. Allen, J. Tauberer, L. Lessig, D. Newman, J. Geraci, E. Bender, T. Steinberg, D. Moore, D. Shaw, J. Needham, J. Hardi, E. Zuckerman, G. Palmer, J. Taylor, B. Horowitz, Z. Exley, K. Fogel, M. Dale, J. L. Hall, M. Hofmann, D. Orban, W. Fitzpatrick, and A. Swartz. 8 principles of open government data. http://www.opengovdata.org/home/8principles, 2007. Open Government Working Group.

[15] J. Manyika, M. Chui, D. Farrell, S. V. Kuiken, P. Groves, and E. A. Doshi. Open data: Unlocking innovation and performance with liquid information. http://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information, 2013.

[16] B. S. Noveck. From faith-based to evidence-based: The open data 500 and understanding how open data helps the american economy. http://www.forbes.com/sites/bethsimonenoveck/2014/01/08/from-faith-based-to-evidence-based-the-open-data-500-a 2014.

[17] Open Data Institute. *Certificates*, 2013.

[18] Open Knowledge Foundation. *Open Data Census*, 2013.

[19] R. Pollock, M. Brett, and M. Keegan. Data packages. http://dataprotocols.org/data-packages/, 2014.

[20] Sunlight Foundation. *Open Data Policy Guidelines*, 2014.