How can we figure out what is inside thousands of spreadsheets?

Thomas Levine
_@thomaslevine.com

ABSTRACT

We have enough data today that we it may not be realistic to understand all of them. In hopes of vaguely understanding these data, I have been developing methods for exploring the contents of large collections of weakly structured spreadsheets. We can get some feel for the contents of these collections by assembling metadata about many spreadsheets and run otherwise typical analyses on the data-about-data; this gives us some understanding patterns in data publishing and a crude understanding of the contents. I have also developed spreadsheet-specific search tools that try to find related spreadsheets based on similarities in implicit schema. By running crude statistics across many disparate datasets, we can learn a lot about unweildy collections of poorly structured data.

Keywords

data management, spreadsheets, open data, search

1. INTRODUCTION

These days, we have more data than we know what to do with. And by "data", we often mean unclean, poorly documented spreadsheets. I started wondering what was in all of these spreadsheets. Addressing my curiosity turned out to be quite difficult, so I've found up developing various approaches to understanding the contents of large collections of weakly structured spreadsheets.

My initial curiosity stemmed from the release of thousands of spreadsheets in government open data initiatives. I wanted to know what they had released so that I may find interesting things in it.

More practically, I often am looking for data from multiple sources that I can connect in relation to a particular topic. For example, in a project I had data about cash flows through the United States treasury and wanted to join them to data about the daily interest rates for United States bonds. In situations like this, I usually need to know the name of the dataset or to ask around until I find the name. I wanted a faster and more systematic approach to this.

2. TYPICAL APPROACHES TO EXPLORING THE CONTENTS OF SPREADSHEETS

Before we discuss my spreadsheet exploration methods, let's discuss some more ordinary methods that I see in common use today.

2.1 Look at every spreadsheet

As a baseline, one approach is to look manually at every cell in many spreadsheets. This takes a long time, but it is feasible in some situations.

2.2 Use standard metaformats

Many groups develop domain-specific metaformats for expressing a very specific sort of data. For example, JSON API is a metaformat for expressing the response of a database query on the web [4], Data Packages is a metaformat for expressing metadata about a dataset [17], and KML is a metaformat for expressing annotations of geographic maps [19].

Agreement on format and metaformat makes it faster and easier to inspect individual files. On the other hand, it does not alleviate the need to acquire lots of different files and to at least glance at them. We spend less time manually inspecting each dataset, but we must still manually inspect lots of dataset.

The same sort of thing happens when data publishers provide graphs of each individual dataset. When we provide some graphs of a dataset rather than simply the standard data file, we are trying to make it easier for people to understand that particular dataset, rather than trying to focus them on a particular subset of datasets.

2.3 Provide good metadata

Data may be easier to find if we catalog our data well and adhere to certain data quality standards. With this reasoning, many "open data" guidelines provide direction as to how a person or organization with lots of datasets might allow other people to use them [16, 1, 18, 13, 15].

At a basic level, these guidelines suggest that data should be available on the internet and under a free license; at the

other end of the spectrum, guidelines suggest that data be in standard formats accompanied with particular metadata.

Datasets can be a joy to work with when these data quality guidelines are followed, but this requires much upfront work by the publishers of the data.

2.4 Asking people

In practice, I find that people learn what's in a spreadsheet through word of mouth, even if the data are already published on the internet in standard formats with good metadata.

Amanda Hickman teaches journalism and keeps a list of data sources for her students [3].

There entire conferences about the contents of newly released datasets, such as the annual meeting of the Association of Public Data Users [14].

The Open Knowledge Foundation [16] and Code for America [2] even conducted data censuses to determine which governments were releasing what data publically on the internet. In each case, volunteers searched the internet and talked to government employees in order to determine whether each dataset was available and to collect certain information about each dataset.

3. ACQUIRING LOTS OF SPREADSHEETS

In order to explore methods for examining thousands of spreadsheets, I needed to find spreadsheets that I could explore.

Many governments and other large organizations publish spreadsheets on data catalog websites. Data catalogs make it kind of easy to get a bunch of spreadsheets all together. The basic approach is this.

1. Download a list of all of the dataset identifiers that are present in the data catalog.
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 Open Data Portal
- Common Knowledge Archive Network (CKAN)
- OpenDataSoft

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

After I've downloaded spreadsheets and their metadata, I often assemble them into a spreadsheet about spreadsheets [6]. 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. CRUDE STATISTICS ABOUT SPREADSHEETS

My first approach was involved running rather crude analyses on this interesting dataset-about-datasets that I had assembled.

4.1 How many datasets

I started out by simply counting how many datasets each catalog website had.

The smaller sites had just a few spreadsheets, and the larger sites had thousands.

4.2 Meaninglessness of the count of datasets

Many organizations report this count of datasets that they publish, and this number turns out to be nearly useless. As illustration of this, let's consider a specific group of spreadsheets. Here are the titles of a few spreadsheets in New York City's open data catalog.

- Math Test Results 2006-2012 - Citywide - Gender
- Math Test Results 2006-2012 - Citywide - Ethnicity
- English Language Arts (ELA) Test Results 2006-2012 - Citywide - SWD
- English Language Arts (ELA) Test Results 2006-2012 - Citywide - ELL
- Math Test Results 2006-2012 - Citywide - SWD
- English Language Arts (ELA) Test Results 2006-2012 - Citywide - All Students
- Math Test Results 2006-2012 - Citywide - ELL
- English Language Arts (ELA) Test Results 2006-2012 - Citywide - Gender
- Math Test Results 2006-2012 - Citywide - All Students
- English Language Arts (ELA) Test Results 2006-2012 - Citywide - Ethnicity

These spreadsheets all had the same column names; they were "grade", "year", "demographic", "number_tested", "mean_scale_score", "num_level_1", "pct_level_1", "num_level_2", "pct_level_2", "num_level_3", "pct_level_3", "num_level_4", "pct_level_4", "num_level_3_and_4", and "pct_level_3_and_4".

These "datasets" can all be thought of as subsets of the same single dataset of test scores.

If I just take different subsets of a single spreadsheet (and optionally pivot/reshape the subsets), I can easily expand one spreadsheet into over 9000. This is why the dataset count figure is near useless.

4.3 Size of the datasets

I can also look at how big they are. It turns out that most of them are pretty small.

- Only 25% of datasets had more than 100 rows.
- Only 12% of datasets had more than 1,000 rows.

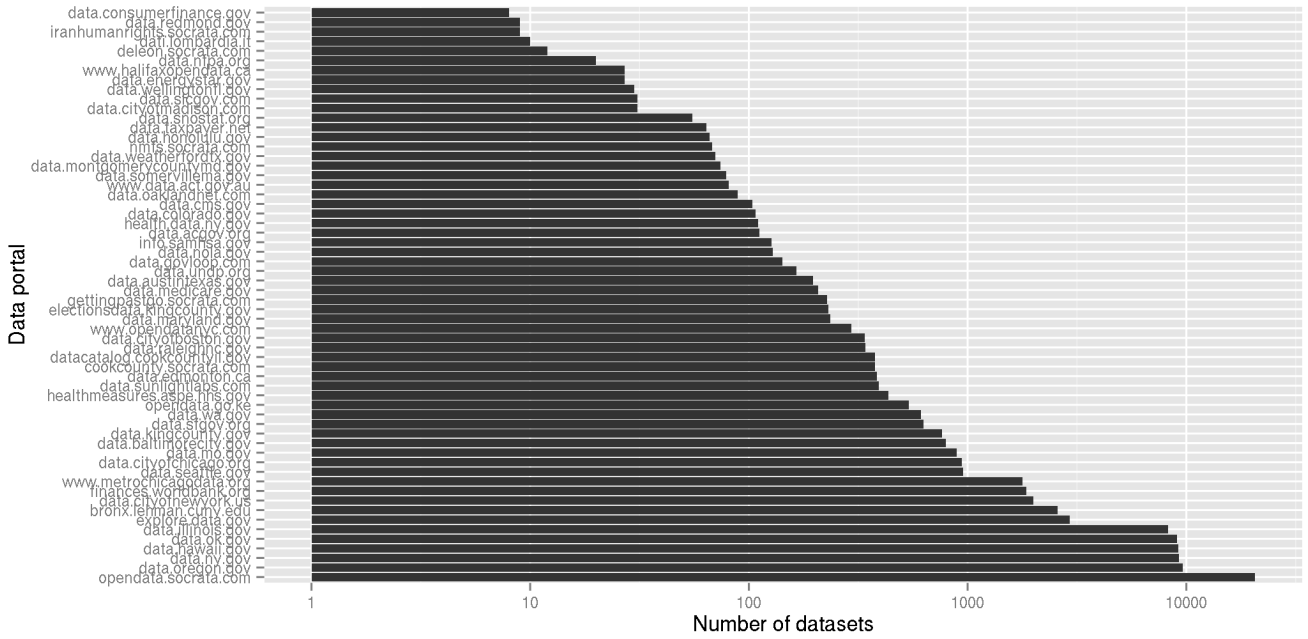

Figure 1: How many datasets (spreadsheets) each data catalog had

- Only 5% of datasets had more than 10,000 rows.

Regardless of the format of these datasets, you can think of them as spreadsheets without code, where columns are variables and rows are records.

5. MEASURING HOW WELL DIFFERENT SPREADSHEETS FOLLOW DATA PUBLISHING GUIDELINES

Having gotten some feel for the contents of these various data catalogs, I started running some less arbitrary statistics. As discussed in section 2.3, many groups have written guidelines as to how data should be published [16, 1, 18, 13, 15]. I started coming up with measures of adherence to these guidelines and running them across all of these datasets.

5.1 File format

File format of datasets can tell us quite a lot about the data. I looked at the MIME types of the full data files for each dataset on catalogs running Socrata software and compared them between data catalogs [11].

If datasets are represented as tables inside the Socrata software, they are available in many formats. If they are uploaded in formats not recognized by Socrata, they are only available in their original format.

I looked at a few data catalogs for which many datasets were presented in their original format. In some cases, the file formats can point out when a group of related files is added at once. For example, the results indicate that the city of San Francisco in 2012 added a bunch of shapefile format datasets to its open data catalog from another San Francicso government website. As another example, most of the datasets in the catalog of the state of Missouri are traffic surveys, saved as PDF files [5].

5.2 Licensing

Many of the data publishing guidelines indicate that datasets should be freely licensed. All of the data catalog websites that I looked at include a metadata field for the license of the dataset, and I looked at the contents of that field. I found that most datasets had no license [10], and this is thought to be detrimental to their ability to be shared and reused [16, 1, 18, 13, 15].

5.3 Liveliness of links

One common guideline is that data be available on the internet. If a dataset shows up in one of these catalogs, you might think that it is on the internet. It turns out that the links to these datasets often do not work.

I tried downloading the full data file for each dataset refenced in any of these catalogs and recorded any errors I received [9, 12]. I found most links to be working and noticed some common reasons why links didn't work.

- Many link URLs were in fact local file paths or links within an intranet.
- Many link "URLs" were badly formed or were not URLs at all.
- Some servers did not have SSL configured properly.
- Some servers took a very long time to respond.

I also discovered that one of the sites with very alive links,

`https://data.gov.uk`, had a "Broken links" tool for identifying these broken links.

6. SEARCHING FOR SPREADSHEETS

While assessing the adherence to various data publishing guidelines, I kept noticing that it's very hard to find spreadsheets that are relevant to a particular analysis unless you already know that the spreadsheet exists.

Major search engines focus on HTML format web pages, and spreadsheet files are often not indexed at all. The various data catalog software programs discussed in section 3 include a search feature, but this feature only works within the particular website. For example, I have to go to the Dutch government's data catalog website in order to search for Dutch data.

To summarize my thoughts about the common means of searching through spreadsheets, I see two main issues. The first issue is that the search is localized to datasets that are published or otherwise managed by a particular entity; it's hard to search for spreadsheets without first identifying a specific publisher or repository. The second issue is that the search method is quite naive; these websites are usually running crude keyword searches.

Having articulated these difficulties in searching for spreadsheets, I started trying to address them.

6.1 Searching across publishers

When I'm looking for spreadsheets, the publishing organization is unlikely to be my main concern. For example, if I'm interested in data about the composition of different pesticides, but I don't really care whether the data were collected by this city government or by that country government.

To address this issue, I made a a disgustingly simple site that forwards your search query to 100 other websites and returns the results to you in a single page [7]. Lots of people use it, and this says something about the inconvenience of having separate search bars for separate websites.

6.2 Spreadsheets-specific search algorithms

The other issue is that our search algorithms don't take advantage of all of the structure that is encoded in a spreadsheet. I started to address this issue by pulling schema-related features out of the spreadsheets (section 4.2).

6.3 Spreadsheets as input to a search

Taking this further, I've been thinking about what it would mean to have a search engine for spreadsheets.

When we search for ordinary written documents, we send words into a search engine and get pages of words back.

What if we could search for spreadsheets by sending spreadsheets into a search engine and getting spreadsheets back? The order of the results would be determined by various specialized statistics; just as we use PageRank to find relevant hypertext documents, we can develop other statistics that help us find relevant spreadsheets.

Figure 2: The search engine for words takes words as input and emits words as output

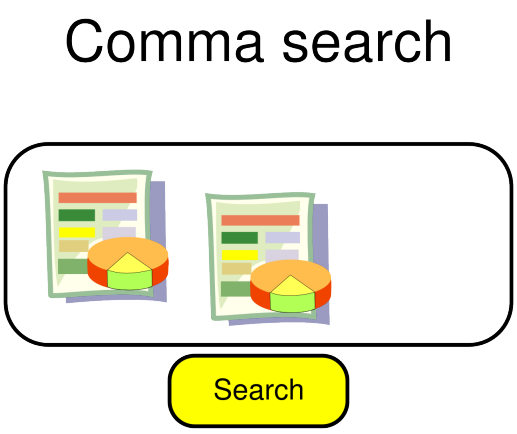

Figure 3: The search engine for spreadsheets takes spreadsheets as input and emits spreadsheets as output

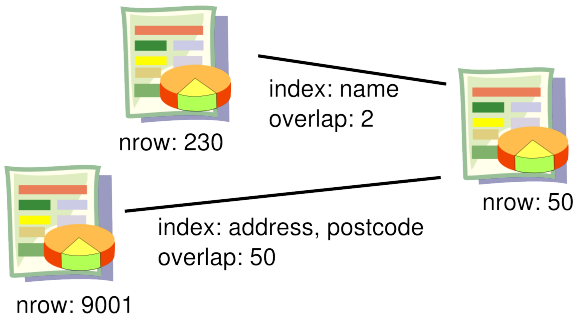

Figure 4: Commasearch infers some schema information about each spreadsheet and looks for other spreadsheets with similar schemas.

6.3.1 *Schema-based searches*

I think a lot about rows and columns. When we define tables in relational databases, we can say reasonably well what each column means, based on names and types, and what a row means, based on unique indices. In spreadsheets, we still have column names, but we don't get everything else.

The unique indices tell us quite a lot; they give us an idea about the observational unit of the table and what other tables we can nicely join or union with that table.

Commasearch [8] is the present state of my spreadsheet search tools. To use comma search, you first index a lot of spreadsheets. Once you have the index, you may search by providing a single spreadsheet as input.

In the indexing phase, spreadsheets are examined do find all combinations of columns that act as unique indices, that is, all combinations of fields whose values are not duplicated within the spreadsheet. In the search phase, comma search finds all combinations of columns in the input spreadsheet and then looks for spreadsheets that are uniquely indexed by these columns. The results are ordered by how much overlap there is between the values of the two spreadsheets.

To say this more colloquially, comma search looks for many-to-one join relationships between disparate datasets.

7. REVIEW

I've been downloading lots of spreadsheets and doing crude, silly things with them. I started out by looking at very simple things like how big they are. I also tried to quantify other people's ideas of how good datasets are, like whether they are freely licensed. In doing this, I have noticed that it's pretty hard to search for spreadsheets; I've been developing approaches for rough detection of implicit schemas and for relating spreadsheets based on these schemas.

8. 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.

Statistics about adherence to data publishing guidelines can be helpful to those who are tasked with cataloging and maintaining a diverse array of datasets. Data quality statistics can 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.

New strategies for searching spreadsheets can help us find data that are relevant to a topic within the context of analysis.

9. REFERENCES

[1] T. Berners-Lee. Linked data. `http://www.w3.org/DesignIssues/LinkedData.html`, 2006.
[2] Code for America. *U.S. City Open Data Census*, 2014.
[3] A. Hickman. *Where to Find Data*, 2014.
[4] S. Klabnik and Y. Katz. Json api: A standard for building apis in json. `http://jsonapi.org/`.
[5] T. Levine. *License-free data in Missouri's data portal*, 2013.
[6] T. Levine. Open data had better be data-driven. `http://thomaslevine.com/!/dataset-as-datapoint`, 2013.
[7] T. Levine. *OpenPrism*, 2013.
[8] T. Levine. *commasearch*, 2014.
[9] T. Levine. Dead links on data catalogs. `http://thomaslevine.com/!/data-catalog-dead-links/`, 2014.
[10] T. Levine. Open data licensing. `http://thomaslevine.com/!/open-data-licensing/`, 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] 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.
[14] A. of Public Data Users. *Association of Public Data Users Annual Conference*, 2013.
[15] Open Data Institute. *Certificates*, 2013.
[16] Open Knowledge Foundation. *Open Data Census*, 2013.
[17] R. Pollock, M. Brett, and M. Keegan. Data packages. `http://dataprotocols.org/data-packages/`, 2013.
[18] Sunlight Foundation. *Open Data Policy Guidelines*, 2014.
[19] T. Wilson. OgcÂő kml. Technical Report OGC 07-147r2, Open Geospatial Consortium Inc., 2008. `http://portal.opengeospatial.org/files/?artifact_id=27810`.