Towards Visual Overviews for Open Government Data

Alvaro Graves
Inria Chile
Av. Apoquindo 2827, piso 12
Santiago, Chile
alvaro.graves@inria.cl

Javier Bustos-Jiménez
NIC Chile Research Labs
Blanco Encalada 1975
Santiago, Chile
jbustos@niclabs.cl

ABSTRACT

The rise of Open Data initiatives has led to the publication of many datasets from different organizations and governments. These datasets cover a wide range of knowledge domains, from budget to education to health care. However, not all datasets have the quality, granularity or type of information that is relevant to each user. Moreover, in many cases the description or metadata does not specify clearly the content of a dataset, difficulting the exploration of datasets by stakeholders. In this paper we propose the use of dashboards and visualizations as a way to preview the content of datasets for easier exploration. The use of visualizations can provide a rapid way to select or discard datasets based on their content, reducing the potential datasets that a user may need to look in order to get what she needs.

Categories and Subject Descriptors

D.2.2 [**Software Engineering**]: Design Tools and Techniques—*User interfaces*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Methodology*; I.7.4 [**Document and Text Processing**]: Electronic Publishing

General Terms

Documentation,Human Factors,Open Government Data

Keywords

Open Government Data, Open Data, Preview, Data Visualization

1. INTRODUCTION

Over one million datasets [16] are currently available in different portals across the globe. Although the data is publicly available, their organization and structure is not clear for all the stakeholders necessarily. For example, at the time of this writing the search for "child obesity" in *Data.gov* and *Data.gov.uk* (the two largest Open Government Data portals) gives different results, as can be see in Figure 1: In

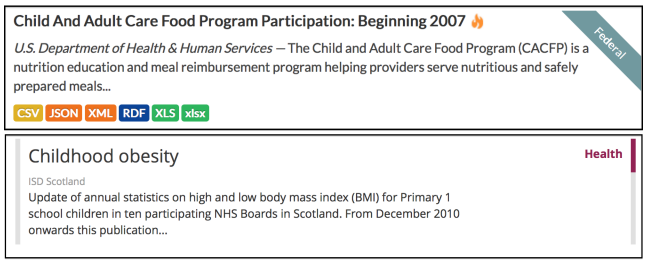


Figure 1: Restults provided by searching for `child obesity` in *Data.gov* (upper image) and *Data.gov.uk* (lower image)

Data.gov, only one dataset is available (in several formats). This dataset is described as "federal", however a closer look shows that the data is related to the state of New York only. In the case of *Data.gov.uk*, 16 results provide information related to child and obesity in PDF and Excel formats. Beyond these difference, it is not clear for a researcher or developer is these datasets are relevant to her needs; having a title and a description is useful, but does not clarify exactly what type of information, granularity and quality of the data is available.

For example, it is not clear what specific data is contained in a dataset, what structure is used or the scope of this dataset. As mentioned early,i n the case of the US dataset about child obesity, it is labeled as "federal", however the data describes only information about New York State; it is likely that other manually curated tags and descriptions may not be precise in terms of the content or scope of the datasets published. Thus, the question in this and many other cases is how can stakeholders know in advance what's in a dataset **before** downloading it? We propose the use of dashboards and visualizations to describe and preview the content of datasets; this visual representations will help stakeholder to decided whether a dataset is useful for them or not.

This paper is structured as follows: Section 2 describes related work found in the literature and state of the art technology. Section 3 discusses different pieces of information that can be used to create visual overviews from some of the more common file formats used to publish Open Government Data. In Section 4 we show a prototype developed as an example of what can be done to create visual overviews

of datasets using the information discussed previously. Section 5 presents the future challenges on our research and we discuss our conclusions in Section 6.

2. RELATED WORK

The problem of good data visualizations has been studied many years [24]. In terms of data exploration and visualization, Schneiderman [22] summarizes the Visual Information Seeking Mantra as *Overview first, zoom and filter, then details-on-demand*; humans need to get the "big picture" of a dataset first in order to decide where to explore next. Thus, a visual overview of a dataset can be be useful for researchers and journalists to know "what's in there" before taking further action.

One of the seminal works in dataset preview was made by Doan et al. in 1999. They studied the effects of visual previews of queries for NASA's EODIS datasets [19], concluding that the main advantages of these visual strategies were:

- "*eliminate zero-hit queries*,
- *reduces network activity and browsing effor by preventing the retrieval of undesired datasets*,
- *represents statistical information of database visually to aid comprehension and explorarion*,
- *support synamic queries, which aids users to discover dataset patterns and exceptions*, and
- (they are) *suitable to novice, intermittent, or expert users*".

A generalization of query previews is presented in the work of Tanin et al. [23], complementing the work of Doan et al. with barcharts in order to show data distribution.

In the beginning of this century, similar conclusions were reached by Green et al. in their study about how previews and overviews allow users to rapidly discriminate useful information from those not for interest [10], applying their findings in the interfaces provided by the Digital Library of Congress and concluding that "*previews should be available at a high level within a site so users get a taste of what is to come early in their visit*".

Nowadays, the principles behind above works seems to be suitable for open data publication, as it has been reported to be for web searching by the work of Dörk et al.[7], where they studied performance and benefits of a new approach called *visual exploration* for information seeking on the Web (Figure 2).

From the perspecting of the Open Government Data, visualizations are valuable and useful artifacts for users [12]; visualizations can provide feedback and help on the decision making process related to public policies. A survey [9] showed that many stakeholders found that users were interested in interacting with data via the use of visualizations. Hence, there is reasonable evidence to support our hypothesis that preview visualizations can be a useful tools for Open Government Data stakeholders.

Figure 2: Visual exploration interface proposed by Dörk et al.[7], which includes data collection choosers, visualization widgets, text query box and the current set of results.

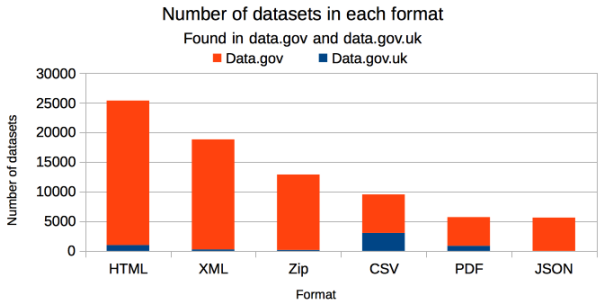


Figure 3: Number of datasets available in *data.gov* and *data.gov.uk* by format.

3. CONTENT FOR VISUAL OVERVIEWS

Different formats provide different support for data, metadata, annotations and other extra information that can be helpful for users to identify datasets that area valuable for them. In order to understand what format are more often used to publish Open Government Data, we looked at *Data.gov* and *Data.gov.uk*, two of the largest government data portals. We took the most popular formats reported by these portals and we found that most datasets are published in HTML, followed by XML, ZIP, CSV, PDF and JSON, as can be seen in Figure 3. It is important to note that in many cases a dataset is published in multiple formats, so these numbers are not related to the number of datasets available.

It is reasonable then to focus our efforts on the most common formats in order to cover an important number of datasets with our study. For this work, we do not considered ZIP files as part of the list of datasets to study, due to the fact that ZIP files are actually archives containing other files, such as CSV. Hence, for this study a ZIP file can be considered only as an "extra layer" of communication, and not a file format that we should study.

3.1 Data, metadata and annotations

We identify three different sources of information in a dataset that can be used to create visual overviews: data, metadata and annotations. We understand metadata different from annotations in that the former is aimed to provide machine-processable data about the dataset (e.g., creation date, author of the dataset), while the latter is more focused on explaining to a human reader certain aspects of the data (e.g., what does a field mean or information about how the data was collected). As mentioned before, different data formats provide different levels of support for data and metadata; thus, extracting data, metadata and annotations from different file formats present different challenges.

3.2 HTML

HTML is a markup language aimed to write "scientific documents, although its general design and adaptations over the years have enabled it to be used to describe a number of other types of documents" [25]. While not a data format *per se*, it has been widely used to publish data in a way that it is easy to consume by humans, via a web browser. There are multiple sources of data, metadata and annotations that we can use to represent visually.

- **Data:** Representing data in HTML can be done in multiple ways, from HTML tables to full web applications. In the most basic case, data can be presented as a list or a table, structured using the `<ul>`, `<ol>` or `<table>` elements. The process of extracting data from HTML documents is know as *Web Scraping* and there are many tools to do so [17][1]. This data can feed visual overviews to give insights about the actual content of the dataset.
- **Metadata:** HTML provides a mechanism to store metadata, by using the `<meta>` element. In the case of well-formed HTML tables, the header of these tables contain valuable metadata as well; the `<th>` element on a table will describe the name of each column, something that will tell a user if the dataset is useful for her purposes or not. These metadata elements can be extracted with web scraping techniques as well and used to give more insight about the structure of the data as well as more information about the provenance of it.
- **Annotations:** HTML supports comments in the code between <!– and –> strings sequences. These annotations can be used to extract information about the document and the data described in it as well. For example, it is possible to obtain the most relevant words in the comments and visualize them using a word cloud. It is important to note that in the case of HTML, many annotations might be related to the JavaScript code used in the document; a smart heuristic could discard potentially confusing annotations of this type.

3.3 XML

The *Extensible Markup Language* [4] is a language focused on structuring data for the Web, by providing a set of rules on how to encode such data. XML defines a tree-like structure where each node is a user-defined tag which may have content and attributes, as can be seen in Figure 4. There are several entities that can be extracted from a valid XML document to be used on a visual overview.

```
<persons>
 <person>
  <!-- this is a comment -->
  <name>John</name>
  <lastname>Doe</lastname>
  <language iso="EN">English</language>
 </person>
</persons>
```

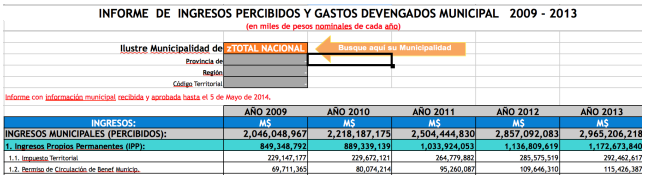
Figure 4: Example of a XML document.

- **Data:** It is possible to check for common words, numbers or phrases that occur in the content of XML tags. One way to do so is by using XPath [5], a query language aimed to extract data from XML documents. Similar to the case of HTML, the data can be used to inform the user about the actual content of the dataset.
- **Metadata:** There are at least two sources of information that can be used for a visual overview. First, the words used as tags and attributes are descriptive of the type of content that is about a dataset. For example in Figure 4, the words `person`, `name` and `lastname` give a good insight of what the data is about. Preprocessing the XML schema with Natural Language Processing techniques (e.g., Term frequency [20] or entity extraction [18]) can provide better insight on what type of information is contained in the dataset. Also, the structure how the data is organized is valuable in itself to understand the dataset; identifying the most common patterns in a XML structure and represent it visually, could give insight to users of what type of data is available, without the need to download the dataset.
- **Annotations:** XML allows comments in a similar way as in HTML (See Figure 4). XML Schema [14] also provides a series of non-mandatory mechanisms to annotate XML documents, by using the `xsd:annotation` tag. Applying NLP techniques as described above could help identify key entities related to this dataset (e.g., countries, contributors, organizations).

3.4 CSV

Comma-separated values is a loosely used term to define plain text files structured as tables, using separators (usually a comma, but semicolon and the *tab* character are not uncommon). CSV files are popular due to its simplicity in terms the readability and processing of the data, done both by humans and computers. In many cases, CSV are the result of exporting Spreadsheet files (such as Microsoft Excel) into text. In many cases it is possible to observe headers that defined the columns of a CSV file.

- **Data:** Since a CSV file is basically a table, it is possible to extract the most common terms found in the cells and display them as a bar chart or a word cloud



INFORME DE INGRESOS PERCIBIDOS Y GASTOS DEVENGADOS MUNICIPAL 2009 - 2013

(en miles de pesos nominales de cada año)

Ilustre Municipalidad de ¿TOTAL NACIONAL | Busque aquí su Municipalidad

Provincia de

Región

Código Territorial

Informe con información municipal recibida y aprobada hasta el 5 de Mayo de 2014.

INGRESOS:	AÑO 2009 M$	AÑO 2010 M$	AÑO 2011 M$	AÑO 2012 M$	AÑO 2013 M$
INGRESOS MUNICIPALES (PERCIBIDOS):	2,046,048,967	2,218,187,175	2,504,444,830	2,857,092,083	2,965,206,218
1. Ingresos Propios Permanentes (IPP):	849,348,792	889,339,139	1,033,924,053	1,136,809,619	1,172,673,840
1.1. Impuesto Territorial	229,147,177	229,672,121	264,779,882	285,575,519	292,462,617
1.2. Permiso de Circulación de Benef Municip.	69,711,365	80,074,214	95,260,087	109,646,310	115,426,387

Figure 5: Example of a spreadsheet version of a dataset. The CSV version does not respect the table structure, due to the titles and headers that are exported along with the rest of the data.

or other way to present it as a visual overview of the dataset. There are tools and libraries for virtually any programming language to read and extract data from CSV files.

- **Metadata:** Due to its simplicity, little metadata can be found in a CSV file. However, as mentioned before, in many cases CSV files contain headers that can be used to identify the topics described in the dataset.
- **Annotations:** CSV does not support annotations, however in many cases, the direct translation from a spreadsheet, such as Microsoft Excel, carries the title and other comments available on it (see Figure 5 as an example). These annotations break the table structure of the CSV file and makes it difficult to read it by programs. Still, these annotations can provide useful information about the content of the file. An heuristic to obtain such annotations could be the following: Read each line of a CSV file and consider it as an annotation, until the header is found.

3.5 PDF

The use of PDF files to publish data is a common practice among practically all governments and organizations, although it is widely discouraged and criticized [15][8]. One of the main reasons is that PDF is a *document* format, not a data format. In this sense, PDF does not comply with the Open Government Data principle [11] that states that data should be in a machine-processable format. Still, many efforts like Tabula [2] have been developed to extract data from PDF files.

- **Data:** As mentioned before, in the best of cases PDF files contain data tables that can be extracted semi-automatically to generate visualizations, similar to the case of CSV files.
- **Metadata:** Although PDF supports metadata and embeddable raw data [13], common tools for creating PDFs do no include metadata but some basic authorship information. It is not clear what type of metadata may be available in the general case to use for an visual overview.
- **Annotations:** Similar to the case of XML and HTML documents, annotations in PDF can be used to identify relevant terms that can be later used create a visual overview.

```
{
  persons: [
       {
         name: "John",
         lastname: "Doe",
         language: {
                     value: "English",
                     iso: "EN"
                   }
       }
  ]
}
```

Figure 6: A possible JSON representation of the data shown in Figure 4 as XML.

3.6 JSON

The JavaScript Object Notation JSON, is an open standard format that has gained popularity, especially in the Web development community, due to the simplicity for consumption by humans and machines alike. JSON provides a mechanism to transmit objects that can be use to communicate different types of variables. Many see JSON as a simpler, easier-to-use alternative to XML [6]. An example of a JSON document can be seen in Figure 6.

Similar to XML, JSON provides a tree-like structure, but supports different data types, arrays and other objects as well. Thus, it is possible to extract similar information as in the case of XML to later be visualized.

- **Data:** The values in a JSON document can be used to obtain the most significant words or phrases that can be used later to create a visualization.
- **Metadata:** Collecting the words used as keys can give insights on what type of data is presented in the document. Also, the tree structure could be used to identify how the data is modeled.
- **Annotations:** JSON does not provide a way to annotate or comment documents.

4. PROTOTYPE

As a way to test our ideas, we developed a demo tool that creates a visual overview of a dataset. This visual overview consist on a sample of the data and a dashboard based on the information extracted from a dataset. Due to simplicity, our prototype only works with CSV files, but the principles shown are the same for the other file formats described in Section 3. We implemented this demo using JavaScript and the D3.js library [3]. The prototype is available at `https://github.com/niclabs/visual-overview` as open source software.

4.1 Rationale

The prototype presents three different levels of detail of the data contained in a dataset. First, we considered useful to give the user a sample of the data, so she can get an idea of what it looks like as a table. To do so, we included the first three rows of the dataset.

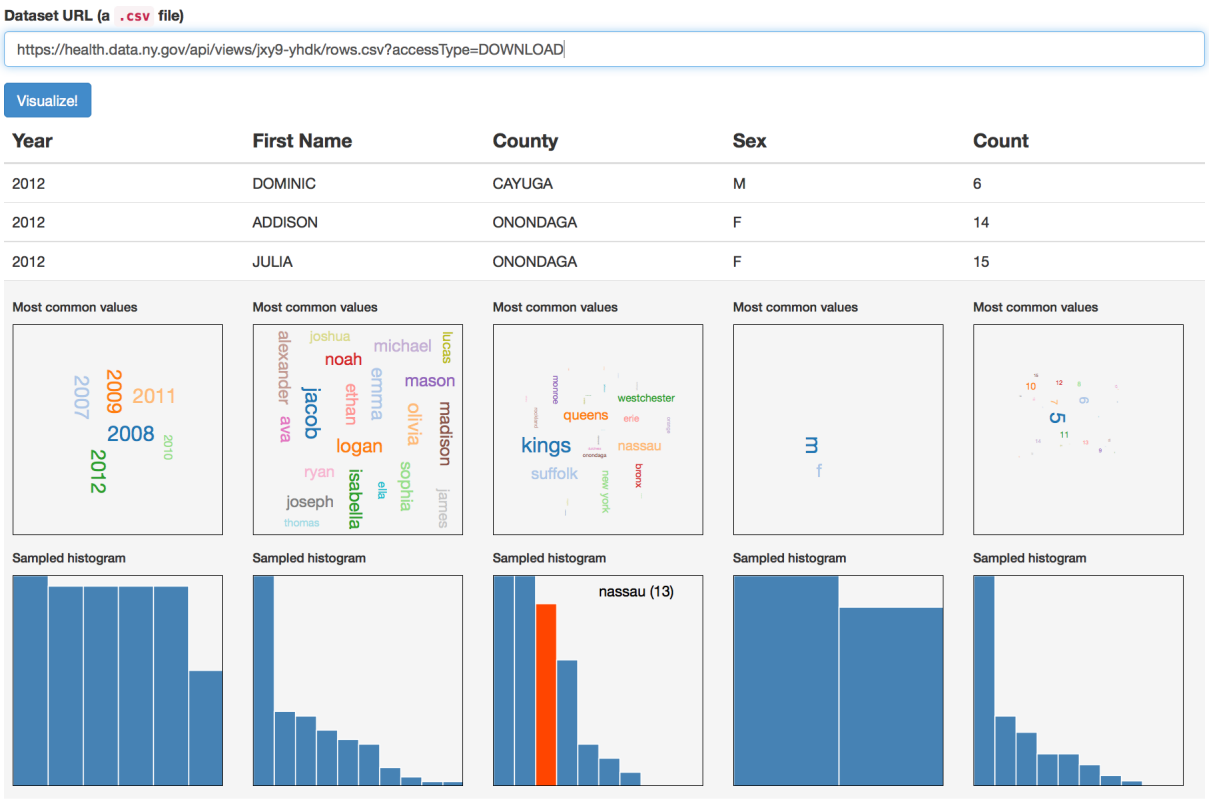


Figure 7: Screenshot of our prototype. A user can indicate a CSV file available and the system will render several statistics related to the values present in the data, as well as the headers available.

Second, in our experience most CSV files describe data properties in terms of columns (in contrast to rows); a CSV column usually contains values related to a specific dimension (e.g., age, latitude, name). Thus, one reasonable approach is to create visualizations for each column. As a way to provide a visual representation of the values on each column, we used *word clouds* [21]; in this way, we present the most common values in each column to the user in a way that is easy to consume without any technical background.

Finally, in many cases it is important to provide more information about the distribution of values to answer questions, such as *Is the data normally distributed? Does it follow a long tail? Are all the values equally likely?*. Although the word cloud provides some insights on this respect, we think a clearer representation was needed. Thus, a histogram of the values in each column is provided. This histogram facilitates the understanding of how the data is distributed and what are the most/least common values.

It is important to note that as a prototype, there are many issues with this software. For example, a more sophisticated approach would consider the type of data (i.e., generic numbers, strings, geographical coordinates, time and dates) and use different visual strategies that are more suitable for each case. The variety of the available values may also affect what visual strategy could be used; for example, for the columns `sex` and *count* the use of word clouds is not necessarily the best strategy.

4.2 Use of the prototype

After a user has entered the URL of a dataset, the prototype will analyze the data in order to extract the more common terms. Although our prototype processes the data live, it is possible to imagine more sophisticated mechanisms that deal with larger datasets, such as offline or batch processing. As mentioned early, our prototype provides several visualizations for each column of the CSV file, including data sample, a wordcloud and a histogram for each column. A screenshot of our prototype can be seen in Figure 7.

5. FUTURE WORK

Our hypothesis is that these visualizations can facilitate the process of deciding if a dataset is useful for a person or not. Thus, we propose to perform a user study to evaluate how easy or hard is for a user to find valuable information in the presence/absence of visual overviews. Also, the effectiveness of visual overviews may also depend on the type of visualizations that are displayed in different scenarios. Further research is necessary in this regard.

From this prototype, we can also take several paths. We plan to include support for other data formats, as described in Section 3. Having a web-based service available to preview and give insights about a dataset can be a valuable tool for journalists, activists and Open Government Data

researchers in general. Another option is to promote the use of tools similar to our prototype to be part of government data portals by default. Most of government organizations already provide a series of tags to help people identify and understand what each dataset is about. Adding an visual overview will help them on that effort. Finally, a smarter set of heuristics could be included in our prototype to provide more suitable visual representations, based on the type of data available in each dataset. Also, the use of annotations in datasets could be used to highlight certain visualizations over others.

6. CONCLUSIONS

In this paper we have proposed the use of visualizations to preview and give insights about datasets that can be useful and valuable to many stakeholders. We showed that for most of the more common file formats used to publish Open Government Data, it is possible to extract valuable information that can be later used to create visual overviews. We also showed how these visual overviews can be created using a prototype developed by the authors that present a dashboard of visualizations based on the information obtained from a dataset. Finally, we discussed the different paths this work can take in the future.

7. REFERENCES

[1] (2011) ScraperWiki. ScraperWiki.2011.http://scraperwiki.com/.

[2] "Tabula," http://tabula.nerdpower.org/, 2013.

[3] M. Bostock, V. Ogievetsky, and J. Heer, "D^3 Data-Driven Documents," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 12, pp. 2301–2309, Dec. 2011.

[4] T. Bray, J. Paoli, C. M. Sperberg-McQueen, E. Maler, and F. Yergeau, "Extensible markup language (xml)," *World Wide Web Consortium Recommendation REC-xml-19980210. http://www. w3. org/TR/1998/REC-xml-19980210*, 1998.

[5] J. Clark, S. DeRose *et al.*, "Xml path language (xpath) version 1.0," 1999.

[6] D. Crockford, "Json: The fat-free alternative to xml," in *Proc. of XML*, vol. 2006, 2006.

[7] M. Dörk, C. Williamson, and S. Carpendale, "Navigating tomorrow's web: From searching and browsing to visual exploration," *ACM Transactions on the Web (TWEB)*, vol. 6, no. 3, p. 13, 2012.

[8] M. Fioretti, "Open data: Emerging trends, issues and best practices," *Laboratory of*, 2011.

[9] A. Graves and J. Hendler, "Visualization tools for open government data," in *Proceedings of the 14th Annual International Conference on Digital Government Research*. ACM, 2013, pp. 136–145.

[10] S. Greene, G. Marchionini, C. Plaisant, and B. Shneiderman, "Previews and overviews in digital libraries: Designing surrogates to support visual information seeking," *Journal of the American Society for Information Science*, vol. 51, no. 4, pp. 380–393, 2000.

[11] O. G. W. Group *et al.*, "Principles of open government data," in *Workshop held in Sebastopol, CA, USA. http://www. opengovdata. org/home/8principles*, 8.

[12] J. Hoxha and A. Brahaj, "Open government data on the web: A semantic approach," in *Emerging Intelligent Data and Web Technologies (EIDWT), 2011 International Conference on*. IEEE, 2011, pp. 107–113.

[13] King, James C., "Role of PDF and Open Data," in *Open Data on the Web, Campus London, Shoreditch, 2013*, 2013.

[14] A. Malhotra and P. Biron, "XML schema part 2: Datatypes," *World Wide Web Consortium Recommendation REC-xmlschema-2-20041028*, 2004.

[15] Manning, Nathaniel. (2013) Bad metrics and PDF graveyards: why development needs open data. http://www.theguardian.com/global-development-professionals-network/2013/oct/21/development-open-data-action.

[16] C. Peng. (2012, Aug.). int. open government data search data analytics. *linking open government data*. [Online]. Available: http://logd.tw.rpi.edu/iogds_data_analytics[RetrievedNov.24,2013]

[17] R. B. Penman, T. Baldwin, and D. Martinez, "Web scraping made simple with sitescraper," 2009.

[18] M. Pennacchiotti and P. Pantel, "Entity extraction via ensemble semantics," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, 2009, pp. 238–247.

[19] C. Plaisant, B. Shneiderman, K. Doan, and T. Bruns, "Interface and data architecture for query preview in networked information systems," *ACM Transactions on Information Systems (TOIS)*, vol. 17, no. 3, pp. 320–341, 1999.

[20] G. Salton and M. J. McGill, "Introduction to modern information retrieval," 1983.

[21] C. Seifert, B. Kump, W. Kienreich, G. Granitzer, and M. Granitzer, "On the beauty and usability of tag clouds," in *Information Visualisation, 2008. IV'08. 12th International Conference*. IEEE, 2008, pp. 17–25.

[22] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Visual Languages, 1996. Proceedings., IEEE Symposium on*. IEEE, 1996, pp. 336–343.

[23] E. Tanin, C. Plaisant, and B. Shneiderman, "Broadening access to large online databases by generalizing query previews," *The craft of information visualization: readings and reflections*, p. 31, 2003.

[24] E. R. Tufte and P. Graves-Morris, *The visual display of quantitative information*. Graphics press Cheshire, CT, 1983, vol. 2.

[25] W3C, "HTML5, A vocabulary and associated APIs for HTML and XHTML," http://www.w3.org/TR/html5/introduction.html, 2014.