Open for whom?
An overview of data.gov file formats

Anne L. Washington, PhD
washingtona@acm.org

School of Policy, Government, and International Affairs
George Mason University, Arlington VA

David Morar
dmorar@gmu.edu

School of Policy, Government, and International Affairs
George Mason University, Arlington VA

April 2016

Open for whom?

Public sector organizations who release open government data to a broad audience face a difficult task. Unlike scientists releasing open data about research, public sector organizations can make few assumptions about the skills or background of the recipients of their data. The user of open government data could be someone new to computers or it could be a sophisticated lobbyist. Public sector organizations have made considerable investments to create data catalogs and repositories yet we have few tools to evaluate these activities. If the release of data is intended as a form of communication, one way to evaluate these investments is to understand those who receive the information. An audience perspective considers how public sector organizations anticipate secondary use of their data sets. This research asks: who is the intended audience for open government data?

The flagship for the open government data movement is data.gov (http://data.gov) which has been hosted by the United States federal government since 2009. It has inspired open data sites with similar Internet addresses in public sector and civil society (ie. data.gov.uk, data.ny.gov, or data.un.org). Open data movements are built on the premise that knowledge should be shared. Open data is sometimes considered an engine of innovation when scientists, governments, or entrepreneurs share their data. Open data can also be evaluated not only by whether it is released but what is released. The file formats that are released provide some indication of the possible reuse. Table 1 is a ranked list of the most frequently appearing file formats on data.gov.

- - - - - - - - - - - INSERT TABLE 1 : Ranked List of Formats - - - - - - - - - - - -

The objective of this study is to investigate the intended audience for open government data through an analysis of available file formats. This unique approach to evaluating open government data will make it possible to do comparative analysis of strategies between specific organizations. Measuring progress in open government data through file formats reveals the organization's strategy to reach an intended audience. This research builds on past work about open data in scientific research (Borgman, 2015; Leonelli, 2013), government information (Clarke & Margetts, 2014; Zuiderwijk & Janssen, 2014), and

transparency (Kitchin, 2014; Lord, 2006).

A file format anticipates the moment of interaction. The selection of a file format may simply be based on the most convenient format for the data publisher. However, that choice also represents the software availability, user skills, and context for reuse. While counting the increasing number of available files is one evaluative approach, we argue that the file formats may best anticipate future activity. This study suggests that file formats serve as a proxy for an anticipated audience and provide some insight into how government organizations may have imagined their data audience.

- - - - - - - - - - - INSERT TABLE 2 Dimensions Defined - - - - - - - - - - - - -

The most frequently appearing file formats on data.gov were aggregated into groups based on two dimensions: access and structure. The first dimension, access, indicates whether there were any limitations for beginning to use the file. A limited file format requires additional knowledge or proprietary software. An unlimited or open, format can be used by multiple software applications. For instance, a TXT text file can be read in many word processing programs or statistical software. The second dimension, structure, indicates whether the document contains elements for machine-readable semantic interpretation. For instance, an unstructured file is an image-only PDF file while a structured file is a spreadsheet. Table 2 defines each dimension. Table 3 shows the relationships between dimensions.

- - - - - - - - - - - - INSERT TABLE 3 Dimension Relationships - - - - - - - - - - - - -

We counted the appearances of specific file formats on data.gov and aggregated those counts into four groups. These groups are loosely based on the 5-star open data deployment advocated by Tim Berners-Lee (2006) who is credited with creating the technology for the world wide web. Berners-Lee currently advocates for a new innovation in the web technology: linked open data. As hyperlinks created networked environments for documents, linked data would create dynamic connections between smaller collections of Internet information.

The five star system used in this study builds on Berners-Lee (2006) condensing it along the dimensions of access and structure. This framework provides a means for evaluating which file formats are suitable for specific activities. Given that few governments have attained the ideal of linked open data (Berners-Lee, 2006), we instead added a zero star element for file formats that were obscure. A zero star is given for records that contain an unknown file format or do not contain sufficient information about formats. The first star is for unstructured formats. We include HTML in the first star although HTML is a structured mark-up language. We spot checked many HTML items in the catalog and found that they connect to a web page that describes but does not contain data. The second star is for proprietary formats. The third star is for structured unlimited data. The fourth star is for advanced file formats that are open and structured but require additional skills to use. Table 4 shows how the file formats were divided into stars.

- - - - - - - INSERT TABLE 4 -- FILE FORMATS for each Star - - - - - - -

Methods

The observations in this study are the records stored in data.gov which is the United States federal data catalog. The data.gov website is a catalog and also a repository. As a catalog, it describes data just as a library card catalog describes books in a library. As a repository, it points to specific file locations just as a table of contents directs the reader to specific pages. Government organizations complete the catalog record with flexibility for input which means that the same file formats are represented in multiple ways.

The data.gov website hosts records for files for other governments. There are state, city, and local data sets hosted on data.gov. For this study, we focused on the data sets representing the United States federal government. In addition, the data is divided into non-geospatial and geospatial types of data. Geospatial data can be used in mapping software to understand relationships between locations. We considered that geospatial data used specialist software that required additional skills. For this study, we chose to focus on federal non-geospatial data sets.

Figure 1 : Sample data.gov Record

The data.gov catalog contains a field for file format to indicate that an item is directly available. A single data catalog record can contain multiple files and file formats. For instance, the same information may be available as HTML and CSV as in the image in Figure 1. In this white paper, we report on the files in the data catalog, not the records. The total number of catalog records is generally smaller than the number of files attached to those records. We found that some data catalog records do not contain any file format. The Federal Student Loan Program Data (https://catalog.data.gov/dataset/federal-student-loan-program-data), which had several hundred views at the time of our study, did not contain any file formats.

We built URLs based on the query syntax available at catalog.data.gov for each file format (i.e. http://catalog.data.gov/dataset?res_format=PDF). The data.gov website makes it possible to do direct queries for specific file formats, however, we were unable to systematically find items without any file format using our methods. We built the same queries for records tagged as federal organizations and non-geospatial data types. Each query was run and the counts were pulled from each table in one sitting on April 21, 2016. At first we did not count the unusual file formats but realized they were a significant part of our findings. The final results for each format were aggregated into groups. The number of file formats available were compared using descriptive statistics.

Our unit of analysis was the file format not the data record. This brings additional limitations to the study. Because a single record contains multiple file formats, we may over represent file formats. Also without downloading the entire data catalog, it is possible that records may have changed during data capture.

We intentionally chose to not track all of the widely diverging descriptions of the same file formats. For instance the Microsoft spreadsheet program was listed as Excel, xls, xlsx, application/xls or zipped xls. We accepted this limitation, because our intention was to use the site as any user would without digging deep for obscure file format descriptions. We did not track all iterations of each file format. Although not thorough, it approximates what a reasonable user with limited knowledge might do. The file format approach is an approximate for user audience. See Table 5 for the complete list of the most frequently appearing file formats aggregated into stars.

- - - - - -- INSERT TABLE 5 – Star Aggregation of Frequent Formats - - - - - - -

Analysis

The US federal data catalog primarily published file formats that follow open standards. Proprietary files are rare. The most frequently occurring formats were PDF and HTML in our study however, they were followed by XML and structured data formats.

The data.gov catalog had 244,689 files listed on April 21, 2016. The total number of data catalog records was 199,666. The file format which appeared the most was HTML with 77,217 files, followed by XML with 42,846 files. The Star 1 category contained the most files, 148,537, which means information was primarily available in unstructured formats. Star 3 contained the second most files with 55,229.

The data.gov catalog had 59,312 non-geospatial files listed from federal organizations. The total number of federal non-geospatial records was 53,537. The file format which appeared the most was PDF with 20,428 files, followed by Application/Octet-Streaming with 18,179 files and XML with 8,422. The Star 1 category contained the most files, 27,990, which means information was primarily available in unstructured formats.

Given that few governments have attained the 5-star ideal of the Berners Lee (2006) linked open

data vision, we instead added a zero star element for file formats that were obscure. These obscure formats contained file formats listed as Unknown, Application/Unknown, Application/Octet-Streaming, Originator Data Format. For the catalog overall, there were 46,838 files in our 0-star category for obscure file formats or 60.06% of the records. For the non-geospatial federal records, there were 18,329 files in our 0-star category or 44.72% of the records. See Table 6 for the complete ranking of star categories.

- - - - - - - - - - - INSERT TABLE 6 -- STAR RANK - - - - - - - - - - -

Discussion

The results confirm that public sector organizations have the technology infrastructure in place to regularly output information and contribute to a data catalog. In summary, a unit of analysis based on the files merely counts the number of times a file format appears in the data catalog. The files represent non-proprietary formats that may be visually structured more than machine-readable. This preliminary analysis raises questions about the imagined audience for open government data.

If the goal of open government data is to release as much as possible into a central location, they have met that goal. The data catalog has 244,689 files. Over eight years, approximately 28,000 files were released each year or over 70 files a day. Given the state of data distribution before data.gov, there is no doubt that a central repository of this size is a significant institutional accomplishment.

If the goal of open government data is to reach as many people as possible, they have succeeded. The majority of the files are open source formats, such as HTML or XML, that have been negotiated through open standards processes. Using unrestrictive file formats automatically invites a wide range of users. It also prevents any monopoly control over public information by a single software vendor who provides proprietary software. The release of non-proprietary file formats that conform to open standards is an important aspect of access. The publication of PDF means that information is easily human-readable by the average citizen with basic software applications. PDF is a file format that privileges visual arrangement of information which may be sufficient for those not manipulating data.

If the goal of open government data is machine readable structured files, there may be a legitimate concern about the large number of PDF and HTML files. The innovators and the data entrepreneurs expect structured machine-readable data. They are data literate people who work in industry, journalism, academics, commerce and oversight. XML was one of the top file formats, however proportionately rather small. XML represented 18 % of all files or 14 % of federal non-geospatial files. XML provides the flexibility for the organization to produce one file but publish multiple formats. XML can be converted into a print-like PDF publication, a CSV for a spreadsheet, or a simple TXT file with content only. XML allows for machine readable options and is also human readable.

If the goal of a data catalog is to provide clarity about what is available, data.gov may be a concern. The number of catalog records that contained sets with obscure or unknown format information, star zero, was alarmingly high, at least 44%. Highly motivated users who are familiar with the topic may be more inclined to download and investigate unlabeled data. Without consistency in cataloging, however, a casual user may overlook important data sets. Instead of making file format a reported field, it might be automatically categorized when the data set is harvested.

File formats are an important aspect of sharing information. They reflect the data publisher's intentions for reuse yet may serve as a barrier of entry for data users. These results are the first stage in an investigation into open data audiences.

Conclusion

Public sector organizations must make decisions about their audience. Will the next generation of open government data appeal to the average person or the technologist? As open government data continues to grow, additional research is needed to refine our understandings about what is available to whom and why. Who is the audience for open government data? The file formats available in the federal data catalog inform the English literate public more than the data literate who want machine-readable information.

Governments attempt to satisfy both the average user, with simple accessible formats, and the sophisticated data consumer, with structured machine-readable formats. Open government data has

established an important pattern of considering both the least and the most sophisticated users. This study suggests that we need a broader conversation about who the data audience will be in the context of open government.

References

Berners-Lee, T. (2006, July 27). Linked Data - Design Issues [blog] http://www.w3.org/DesignIssues/LinkedData.html

Berners-Lee, T. (2006) Linked data 5-star Open Data scheme. http://5stardata.info/en/

Borgman, C. L. (2015). Big data, little data, no data: scholarship in the networked world Cambridge, MA: MIT Press.

Clarke, A., & Margetts, H. (2014). Governments and Citizens Getting to Know Each Other? Open, Closed, and Big Data in Public Management Reform. Policy & Internet, 6(4), 393–417.

Kitchin, R. (2014). Data revolution: big data, open data, data infrastructures and their consequences. [S.l.]: Sage Publications.

Leonelli, S. (2013). Why the Current Insistence on Open Access to Scientific Data? Big Data, Knowledge Production, and the Political Economy of Contemporary Biology. Bulletin of Science, Technology & Society, 33(1-2), 6–11.

Lord, K. M. (2006). The perils and promise of global transparency : why the information revolution may not lead to security, democracy, or peace. Albany: State University of New York Press.

Zuiderwijk, A., & Janssen, M. (2014). Open data policies, their implementation and impact: A framework for comparison. Government Information Quarterly, 31(1), 17–29.

NOTES

HTML – Hypertext Markup Language

PDF – Portable Document Format

TXT – Plain text file

XML – eXtensible Markup Language

RDF – Resource Description Framework

CSV – Comma Separated Values

Tables

Table 1. Frequent Formats on data.gov

Ranked List of File Formats
1. HTML
2. XML
3. PDF
4. originator data format
5. application/octet-stream
6. ZIP
7. TIFF
8. MrSID
9. CSV
10. WMS
11. JSON
12. XYZ
13. RDF
14. WCS
15. application/jpg
16. Esri REST
17. TXT
18. NetCDF
19. iwxmm-us
20. JPEG

The most frequent file formats on data.gov ranked by number of files.

Table 2. Dimensions Definitions

Dimension	Definition
ACCESS	**Barriers to use**
Limited Access	Proprietary format or requires specialty software
Unlimited Access	Open standards-based format and readable in multiple software applications.
STRUCTURE	**Semantic context**
Structured	Contains semantic information, mark-up or machine readable structure
Unstructured	Any arrangement. May contain visual formatting or human-readable patterns

Table 3. Relationship between Access and Structure Dimensions

	ACCESS	
	Limited	**Unlimited**
Structured	Contains semantic information in proprietary formats.	Contains semantic information in open formats.
Unstructured	Contains visual information in proprietary formats.	Contains semantic information in open formats.

Table 4. File Formats in each Star Category.

Stars	Property	Formats
0 Stars	Unknown	Unknown application/unknown App/Octet-Stream ODF
1 Stars	Unstructured	ascii, txt, html, pdf
2 Stars	Proprietary	doc, docx, xls, xlsx
3 Stars	Structured Unlimited	xml, rss, json,
4 Stars	Linked Open Data	rdf, lod

The file formats are loosely arranged into 5 stars from zero to four based on Berners-Lee (2006) 5-star deployment of open data.

Table 5. Frequent File Formats Aggregated into Stars

Star	File Format	Whole data.gov	Federal Non-Geospatial data.gov
1*	HTML	77,217	6,475
	PDF	34,831	20,428
	ZIP	17,269	707
	TXT	4,827	374
	ascii	737	0
	TIFF	13,656	6
2*	XLS	757	741
	xlsx	289	288
	DOC	71	55
	docx	15	15
	Excel	3,067	279
3*	CSV	12,383	1,684
	RTF	0	0
	XML	42,846	8,422
	XBRL	0	0
4*	JSON	10,855	840
	RDF	7,482	659
	rss	10	10
0*	App/Unknown	133	133
	Unknown	17	17
	App/Octet-Str	18227	18179
	ODF	28461	0
	TOTAL	**244,689**	**59,312**
	TOTAL on page	199,666	53,537

The file formats were aggregated into stars for the whole data catalog and for records that represented only Federal Non-Geospatial data.

Table 6. Star Rankings

Whole data.gov		**Federal Non-Geospatial data.gov**	
Star1	**148,537 files**	*Star1*	**27,990 files**
Star3	**55,229 files**	*Star0*	**18,329 files**
Star0	**46,838 files**	*Star3*	**10,106 files**
Star4	**18,347 files**	*Star4*	**1,509 files**
Star2	**4,199 files**	*Star2*	**1,378 files**

The file formats were aggregated into stars and then the stars were ranked for the whole data catalog and for records that represented only Federal Non-Geospatial data.