Open Government Data and File Formats: Constraints on Collaboration

Anne L. Washington PhD
Schar School of Policy and Government
George Mason University
Arlington, VA
washingtona@acm.org

David Morar
Schar School of Policy and Government
George Mason University
Arlington, VA
dmorar@gmu.edu

ABSTRACT

This exploratory interpretive case study investigated the collaborative potential of open government data available through data.gov, the US federal open data catalog. Open data is a central aspect of open government collaboration because it fosters exchange and communication between governments and the public. Government organizations that release open data make choices about file formats that have a substantial impact on the potential for collaboration. A file format, such as a document or a spreadsheet, is a constraint on which programs can read the file and what actions a user can do with the file. Overall, we found data.gov formats with limited collaboration potential but files that could be accessed by people with a wide range of skills. The findings are incorporated into suggestions for future iterations of open data policy. The advantages and limitations of using file formats for open data research are considered. The exploratory findings raise questions about future user-centric open data evaluations.

CCS Concepts

• **Information systems~Data exchange** • *Information systems~Users and interactive retrieval*

Keywords

Open data; data.gov; United States; user-centric data; file formats.

ACM Reference format:

A L. Washington, D. Morar 2017. Open Government Data and File Formats: Constraints on Collaboration. In *Proceedings of dg.o '17, Staten Island, New York, June 2017 (dg.o 2017)*, 5 pages.
DOI: 10.1145/3085228.3085232

dg.o '17, June 07-09, 2017, Staten Island, NY, USA

ACM ISBN 978-1-4503-5317-5/17/06...$15.00
http://dx.doi.org/10.1145/3085228.3085232

1. INTRODUCTION

Collaboration is a critical aspect of open data. Funded research projects and government agencies now produce a stable supply of internal digital files as open data. Open data policy encourages data producers to cooperate with outside partners [35] Data files, unlike static documents, require active engagement to use. Open data has become an engine of innovation for scientists [22], policy advocates [6], entrepreneurs [19] and governments [9], [12]. Increasingly public access to information is an expected aspect of contemporary democracy. However, a 2015 Pew research survey [14] found that only 17% of Americans thought open data could improve government.

Open government data are digital files produced with public funds [16] and made available with few licensing restrictions [31]. The wide availability of open data is designed to promote public engagement yet there are few assessments about how digital material is used by the public. Initially the release of any digital government information was evaluated as an achievement [23]. The number of visitors [18] nor the number of files released [29] have been sufficient to understand open data's potential for collaboration.

An exploratory interpretive case study investigated the collaborative potential of open government data through an examination of file formats. Files provide some indication of possible public engagement. For instance, a table available in a spreadsheet format provides more interaction than a picture in an image format. Bertot [4] argued for digital government research that assesses the quality of community service and delivery mechanisms. File formats deliver government information as databases, spreadsheets, transaction logs, semi-structured texts, streaming geospatial values, or other digital files [8], [15]. Each file format has associated skills and tools needed to manipulate the file in appropriate software programs.

The United States open data catalog is an extreme exemplar case [36] that provides unique opportunities for case study research. Established in 2009, data.gov was the first US national repository of open government data. It quickly became the flagship for the data dot gov phenomenon and part of the worldwide open government movement [25]. This study evaluates the potential for open government collaboration by

examining the file formats released by US federal agencies on data.gov. This case study demonstrates how file formats can impact the reuse of open data.

Overall, we found file formats with limited collaboration potential but that can be accessed by people with a wide range of skills. There was a wide breadth of file formats across all agencies. Individual agencies tended to concentrate their offerings around a few formats. The advantages and limits of using file formats for empirical investigation are discussed. The findings are incorporated into suggestions for future iterations of open data policy.

2. BACKGROUND

The US federal government built a critical mass of open data through its information policy. On the first day in office, a new United States presidential administration outlined a federal information policy that emphasized open data [28]. It was the first of a series of memos, executive orders, guidance, and directives issued by the Executive Office of the President. Government agencies were required to release open data to the public. It was during the 2009 era of open government that the flagship data catalog, data.gov, became the central repository of open data in the United States [25].

Open data repositories rely heavily on compliance with information policies. Suppliers of open data are assessed on delivery of material [16], [33]. Maturity models provide benchmarks for the readiness of the open data supplier to create stable flows of data [31]. The Common Assessment Methods for Open Data (CAF) is one of the practical methodologies for analyzing open data based on timeliness, planning, and other quantitative metrics [2]. The OECD recently began to capture the OUR Index (Open, Useful, Reusable Government Data) that tracks availability and support for innovating through the reuse of data.

While supply-side assessments can determine the amount of transparency, they do not adequately address the needs of those who use open data. People who actively integrate open data need to know more than the number of files available. The demand for open data involves assessing each data set's potential for reuse. Releasing data that has the potential for secondary use requires collaboration between the supply and demand for open data.

3. LITERATURE

Collaboration, along with participation and transparency, defined open government in the United States in 2009 [23], [28], [34]. Transparency is the outward action of the government towards the public while participation is the action of the public reaching in towards the government [34]. Although inherently slow [26], participative decision-making can increase trust in government [24] and improve decision-making.

Wirtz and Birkmeyer [34] suggest that collaboration succeeds when data are integrated in partnerships between public sector organizations, citizen, social groups, or corporations. A transparent government makes its information accessible for public consumption. Collaboration signifies cooperation between transparent governments and a participating public.

Open data initiatives fit into a larger scope of information policy. Information policy is a set of guidelines for creating and analyzing objects that contain recorded knowledge. Braman [5] analyzes the information policy implications of making and reproducing "facts" along four dimensions. The first dimension deals with the entity that is doing the knowing, the who, or perceptual identity. This is exemplified in the move towards an information society. The second dimension focuses on the out-in direction of knowledge and how it restricts or liberates the experience of the real. The third-dimension articulates the in-out direction or the projection of our knowledge unto the world. Technological change and policy issues are inherent actors in this dimension. Finally, the fourth dimension deals with the process of establishing a common platform for knowledge. Technology simply realigns and recreates the platform for discussing knowledge. Open data in this way serves several roles within information policy.

Collaboration emphasized the dynamic aspects of governing that were reflective in an interactive relationship, especially new interactions available through technology. Technological innovation powers and enables open government [10] by providing technical infrastructure [21]. Importantly, technology uses innovative tools, methods and systems [24] to foster collaboration.

Most open data benchmarks focus on the originating organization [1], [7], [30], [32]. These benchmarks serve as organizational maturity models that assist information system managers in managing their assets. Computer-mediated open government [26] is often measured by benchmarks that emphasize supply not demand of open data. User-centered evaluations of open government initiatives are now receiving scholarly attention [17]. Still, little is known about the user perspective of open data despite notable exceptions that investigate stakeholders [11] and the usability of data platform [18].

Tim Berners-Lee [3] who is credited with creating the infrastructure to dynamically link documents through hyperlinks began to advocate for infrastructure to link data in 2006. With linked open data as the ultimate goal, Berners-Lee [3] suggests that governments release the same information in increasingly collaborative file formats which he presented as a "5-star framework". Governments can begin by releasing any digital file with little attention to reuse potential. Gradually governments would move from unstructured formats, like a text file, to structured formats like a spreadsheet. The 5-star framework [3] also encourages open standard formats instead of proprietary formats, which require commercial software. Eventually, data would be available in advanced machine-readable files and finally ready for semantic interpretation as linked open data. The 5-star framework [3] establishes the importance of file formats as the basis for collaborations between governments and the public.

We adopt the Berners-Lee model by combining the upper two tiers of the 5-star framework [3] to create four categories. First, simple unstructured formats are human readable only. Second, proprietary formats require the purchase of special software or hardware. Three, structured formats have some internal organizing logic that can be consistent processed by algorithms. Fourth, advanced machine-readable formats can be processed using computational tools such as those in Berners-Lee's ideal networked environment.

4. METHODS

This paper reports the findings of an exploratory interpretive case study on the collaborative potential of open government data. Case studies provide a comprehensive analysis of a complex research environment. Yin [36] characterized case studies as one of three types: exploratory, descriptive or explanatory. An exploratory case study gathers evidence to define later research directions.

The United States open data catalog was selected as the focus of this exploratory case study. The data.gov website is a data catalog containing a record for each dataset. As one of the earliest national data catalogs, it represents a unique opportunity for research. This extreme exemplar case [36] provides insight into a mature ecosystem [12] for open data.

The research design reflects a user-centered perspective by emphasizing file formats. Case studies in information systems research frequently focus on technology from the user's view [20], [27]. File formats appear in clusters with related but not identical file formats. For instance, the Microsoft spreadsheet program was listed as Excel, xls, xlsx, application/xls or zipped xls. Because our intention was to use the site as any user might, we intentionally did not track all iterations of each file format. We grouped file formats based on the three-four letter extensions that follow the file name. For instance .xls files were grouped but application/xls files were not.

The unit of analysis was the file format not the data record. The total number of catalog records is generally smaller than the number of files attached to those records. Since the data.gov catalog hosts data from a large number of different organizations, not all records were selected for this study.

In order to evaluate national open data policy, this study focused on the data sets released from 2009-2016 US federal cabinet-level agencies. The choice of other federal government organizations was based on whose performance data was evaluated by the OMB, Office of Management & Budget, under the Chief Financial Officers CFO Act of 1990. Twenty-five agencies were included in this study.

We first built URLs for each file format based on the query syntax described in the documentation on the website catalog.data.gov (i.e. http://catalog.data.gov/dataset?res_format=PDF). We then ran each query within a single twenty-four-hour period. There is a limitation to this method. Without downloading the entire data catalog at once, it is possible that records may have changed during data capture.

A modified 5-star framework [3] provided a means for categorizing file formats that were suitable for specific activities. Formats were categorized in four groups that move from simple to complex: 1) simple unstructured, 2) proprietary, 3) structured, or 4) advanced machine-readable. We also noted files that were inadequately labeled or had ambiguous metadata.

5. FINDINGS

The data.gov catalog had 206,799 files available on April 21, 2016 for twenty-five federal government organizations. The total number of files in the entire data catalog was 244,689. The number of files posted by each organization ranged from 110,620 files to 21 files. The average number of files for each organization was 8,272.

The majority of files, 54.4%, were simple unstructured file formats such as text or PDF Portable Document Format. Structured file formats, such as CSV common separated values, accounted for 20.2% files. Advanced machine-readable formats, such as XML (eXtensible Markup Language), accounted for only 6.7%. Unfortunately, 17.1% files had insufficient metadata to determine the file format. See Table 1 for a complete list.

The file formats of each federal organization were analyzed and grouped. With the exception of one agency that released 84% in proprietary files, most agencies released few files that required the purchase of commercial software. The fewest number of files 1.5%, were in proprietary formats.

Table 1. Data.gov File Formats

1. Unstructured	2. Proprietary	3. Structured	4. Advanced
128,372	1,576	28,627	1,435
54.4%	1.5%	20.2%	6.7%
HTML,TXT, PDF	XLS, DOC	CSV, XML	RSS, LOD
Human Readable	Special Software	Internal Logic	Machine Readable

Organizations tended to concentrate their offerings around one type of file format. In other words, files were not evenly distributed across all possible file formats. It is important to note that HTML, Hypertext Markup Language, is listed in the unstructured category although the file format holds the possibility of containing its own internal logic. In the data catalog, HTML files mostly pointed to a web page, with no formatted content, back on the agency's website.

More than half of the agencies, 14, released 50% or more of their files were in simple unstructured formats. In fact, three agencies released over 97% of their files in simple unstructured formats or without sufficient metadata to determine a format. Only two agencies released fewer than 10% as simple unstructured formats. The Department of Commerce and NASA National Aeronautics and Space Administration were the only two agencies that released more than 50% of their files in structured or advanced machine-readable formats. Not one

agency in 25 released a majority of their files solely in advanced machine-readable formats as envisioned in the 5-star Framework [3]. See Table 2 for the government agencies that released the most files in each file format category.

Table 2. Agencies By Highest Percentage in Format Category

	Agency	Percentage of total	
1. Unstructured	DHS Homeland Security	99.06%	213 files
2. Proprietary	NRC Nuclear Regulatory	83.87%	31 files
3. Structured	Labor Department	69.01%	284 files
4. Advanced	HHS Health & Human Services	41.86%	2,876 files

Overall, a majority of agencies released information as simple unstructured files which are digital files ideal for reading. Two of the 25 released a majority of files in file formats ideal for interaction. Agencies mostly avoided releasing file formats that were proprietary. The US federal open data catalog primarily contains human-readable material not machine-readable data.

6. DISCUSSION

The findings of this exploratory case study lead to several questions about open data collaboration.

How can collaboration potential be assessed? In this study we considered file formats as a measure of open data's potential for sparking collaboration between governments and the public. Machine-readable files are released primarily by only a few agencies, limiting cooperation and partnerships to those topic areas. Innovators and people who actively reuse digital material prefer machine-readable structured file formats [19]. Data literate users would be frustrated by the constraints of PDF files, which represent the majority of files that are released by US federal agencies. Additional research might identify the needs of both government organizations and partner data users in successful collaborations.

Could file formats measure the potential for demand? File formats give insight into how knowledge is constructed [5] within the government as a whole or by individual agencies. File formats also have the advantage of being readily available for observation. Counting the number and types of formats is a way to create cross-site comparisons. Heald [13] argues that policy should be designed with an understanding of multiple types of transparency, each which can be quantified or measured differently. File formats, which constrain activity, also provide a means to track different types of transparency. Real-time transparency [34] might be better served by measuring the number of streaming file formats while retrospective transparency [13] might be sufficiently served by static PDF files. The large variety of file formats available could support many demands.

What types of file formats best represent the goals of open government? Ten years of open government did not produce the network of linked data envisioned by early pioneers [3]. US federal agencies in 2016 chose to release unstructured human-readable file formats over the structured machine-readable file formats. While open government policy emphasized collaboration and participation, human-readable digital formats do not reflect those goals. On the other hand, open data policy was successful in setting an expectation of transparency through the release of non-proprietary file formats. The files that are available represent open standards and will reach the maximum number of people and machines. In this way, open data catalogs efficiently deliver digital documents to the average citizen. A file format distills a wide range of open data policy goals into a single object. To meet differing goals, open data policy might consider encouraging the release of a range of file formats.

How do we evaluate open data investments? Funding agencies, civil society, and governments have poured considerable effort into updating information policy and establishing open data infrastructure. These investments have significantly increased the amount of digital material available to the public in a central location. The findings in this study suggest that federal open data investments support transparent government by providing files that require few special skills or software. Future investments might consider how to balance transparency and collaboration through additional funding incentives.

Open data policy could improve the range of file formats available in data catalogs. A simple mandatory quota for file formats might not be practical. An agency that releases geospatial data might be unnecessarily compelled to release spreadsheets instead. This might cause additional burdens for a mandate that is already unfunded. An alternative would be open data policy that encourages agencies to identify relevant file formats for their constituents. Another approach would require agencies to be transparent about the predominant file formats they release. File format transparency sets expectations and also focuses on certain types of collaboration. Some public sector agencies are suited to partnerships with specialized data innovators while others are suited to cooperating with the wider public.

7. CONCLUSION

This case study explored how governments collaborate with the public through open data. It contributed to digital government research by providing an analysis of the constraints and potential of open data file formats. The exploratory findings raised questions about user-centric open data evaluations. This project is part of a body of research that gives scholarly attention to the user perspective of open government. More research is needed to understand collaborations between governments and the people who reuse open data.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the Doctoral Seminar on Open Government Data at George Mason University's School of Policy, Government, and International Affairs. This project is based, in part, on work supported by the National Science Foundation Grant No. # 1635449. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Attard, J. et al. 2015. A systematic review of open government data initiatives. Government Information Quarterly. 32, 4 (Oct. 2015), 399–418.
[2] Atz, U. et al. 2015. Benchmarking open data automatically. Technical Report #ODI-TR-2015-0000 UK20150105. Open Data Institute.
[3] Berners-Lee, T. 2006. Linked Data - Design Issues.
[4] Bertot, J.C. 2006. User-centered e-government: Challenges and benefits for government Web sites (an editorial). Government Information Quarterly. 23, 2 (2006), 163–8.
[5] Braman, S. 2011. Defining Information Policy. Journal of Information Policy. 1, 0 (Feb. 2011).
[6] Clarke, A. and Margetts, H. 2014. Governments and Citizens Getting to Know Each Other? Open, Closed, and Big Data in Public Management Reform. Policy & Internet. 6, 4 (Dec. 2014), 393–417.
[7] Cruz, N.F. da et al. 2016. Measuring Local Government Transparency. Public Management Review. 18, 6 (Jul. 2016), 866–893.
[8] Dawes, S.S. 2010. Stewardship and usefulness: Policy principles for information-based transparency. Government Information Quarterly. 27, 4 (Oct. 2010), 377–383.
[9] Dawes, S.S. et al. 2016. Planning and designing open government data programs: An ecosystem approach. Government Information Quarterly. 33, 1 (Jan. 2016), 15–27.
[10] Evans, A.M. and Campos, A. 2013. Open Government Initiatives: Challenges of Citizen Participation. Journal of Policy Analysis and Management. 32, 1 (Winter 2013), 172–185.
[11] Gonzalez-Zapata, F. and Heeks, R. 2015. The multiple meanings of open government data: Understanding different stakeholders and their perspectives. Government Information Quarterly. 32, 4 (Oct. 2015), 441–452.
[12] Harrison, T.M. et al. 2012. Creating Open Government Ecosystems: A Research and Development Agenda. Future Internet. 4, 4 (Oct. 2012), 900–928.
[13] Heald, D. 2006. Varieties of Transparency. Transparency: the key to better governance?. C. Hood and D. Heald, eds. Oxford University Press. 3.
[14] Horrigan, J.B. and Rainie, L. 2015. Users' Views on Potential Impacts of Open Data and Open Government. Pew Research Center's Internet & American Life Project.
[15] Jaeger, P.T. 2007. Information policy, information access, and democratic participation: The national and international implications of the Bush administration's information politics. Government Information Quarterly. 24, 4 (2007), 840–859.
[16] Janssen, M. et al. 2012. Benefits, Adoption Barriers and Myths of Open Data and Open Government. Information Systems Management. 29, 4 (Fall 2012), 258–268.
[17] Janssen, M. and Zuiderwijk, A. 2014. Infomediary Business Models for Connecting Open Data Providers and Users. Social Science Computer Review. 32, 5 (Oct. 2014), 694–711.
[18] Kapoor, K. et al. 2015. Open Data Platforms and Their Usability: Proposing a Framework for Evaluating Citizen Intentions. Open and Big Data Management and Innovation. M. Janssen et al., eds. Springer International Publishing. 261–271.
[19] Lakhani, K.R. et al. 2009. Data.gov. Technical Report #HBS No. 9-610-075. Harvard Business School Publishing.
[20] Lee, A.S. 1989. A Scientific Methodology for MIS Case Studies. MIS Quarterly. 13, 1 (1989), 32.
[21] Lee, G. and Kwak, Y.H. 2012. An Open Government Maturity Model for social media-based public engagement. Government Information Quarterly. 29, 4 (Oct. 2012), 492–503.
[22] 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 (Feb. 2013), 6–11.
[23] Linders, D. and Wilson, S.C. 2011. What is Open Government?: One Year After the Directive. Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times (New York, NY, USA, 2011), 262–271.
[24] Luna-Reyes, L.F. et al. 2014. Open Government, Open Data and Digital Government. Government Information Quarterly. 31, 1 (Jan. 2014), 4–5.
[25] McDermott, P. 2010. Building open government. Government Information Quarterly. 27, 4 (2010), 401–413.
[26] Meijer, A.J. and Thaens, M. 2009. Public information strategies: Making government information available to citizens. Information Polity: The International Journal of Government & Democracy in the Information Age. 14, 1–2 (2009), 31–45.
[27] Myers, M.D. 2009. Qualitative research in business and management. Sage.
[28] Obama, B. 2009. Transparency and Open Government.
[29] Peled, A. 2011. When transparency and collaboration collide: The USA Open Data program. Journal of the American Society for Information Science and Technology. 62, 11 (Nov. 2011), 2085–2094.
[30] Susha, I. et al. 2015. Benchmarks for Evaluating the Progress of Open Data Adoption Usage, Limitations, and Lessons Learned. Social Science Computer Review. 33, 5 (Oct. 2015), 613–630.
[31] Ubaldi, B. 2013. Open Government Data - Towards Empirical Analysis of Open Government Data Initiatives. Technical Report #22. Organisation for Economic Co-operation and Development.
[32] Veljković, N. et al. 2014. Benchmarking open government: An open data perspective. Government Information Quarterly. 31, 2 (Apr. 2014), 278–290.
[33] Viscusi, G. et al. 2014. Compliance with Open Government Data Policies: An Empirical Assessment of Italian Local Public Administrations. Information Polity: The International Journal of Government & Democracy in the Information Age. 19, 3/4 (2014), 263–275.
[34] Wirtz, B.W. and Birkmeyer, S. 2015. Open Government: Origin, Development, and Conceptual Perspectives. International Journal of Public Administration. 38, 5 (Apr. 2015), 381–396.
[35] Yang, T.-M. et al. 2014. How is information shared across the boundaries of government agencies? An e-Government case study. Government Information Quarterly. 31, 4 (Oct. 2014), 637–652.
[36] Yin, R.K. 2009. Case study research: design and methods. Sage Publications.