Integrating User Feedback with Open Data Quality Models

Paper and visual presentations: Leave author and other identifying information blank for double-blind peer review

First Author Name
Affiliation
Address
e-mail address

Second Author Name
Affiliation
Address
e-mail address

ABSTRACT

User feedback is critical to improving the quality of open data. However, most open data publishers gather only anecdotal evidence about user experiences. This unstructured and informal commentary is, consequently, difficult to translate into actionable steps towards improving data quality. Drawing on user comments collected from Data.Gov - an open data portal providing access to thousands of datasets published by city, state, and federal government agencies in the USA – we inductively develop a classification of reported data quality issues. This poster presents preliminary findings from applying this coding scheme to all issues that users filed on Data.gov in 2015 and 2016. We suggest that our classification scheme can help open data publishers collect structured, actionable information to improve data quality.

Keywords

Open data, data quality, civic technology.

INTRODUCTION

Over the last decade open data initiatives have demonstrably increased access to public sector information (Davies, 2010); are shown to correlate strongly with more efficient and impactful basic science research (Piwowar and Vision, 2013); and, can act as a driver of private sector innovation in critical areas of our economy such as healthcare and renewable energy (Chan, 2013). Yet, many previous studies have shown that open data remains difficult for lay users to discover (Martin et al., 2013), access (Janssen et al., 2012), and use (Braunschweig et al, 2012). Some barriers to the meaningful use of open data include poor metadata quality (Neumaier et al. 2016), insufficient provenance information (Umbirch et al. 2015), and poorly organized data portals that are supposed to match data producers with data consumers (Thorsby et al., 2017). Collectively, these barriers have been described as a function of the overall quality of open data (Saez-Martin, Rosario, & Perez, 2016).

The goal of our on-going research is to better understand how to communicate a relative notion like 'quality' to potential open data users. We believe that by avoiding rigid definitions of quality, and instead focusing on descriptive features of a dataset, we can enable users to make better judgements about a dataset's fitness for use. The preliminary work described in this poster is focused on creating an empirically informed model of user feedback, focused specifically on users of data.gov. Formally stated our research questions ask:

> What data quality 'issues' do users report to data.gov?
>
> How can this unstructured user feedback be generically classified in order make data quality actionable for data publishers?

RESEARCH DESIGN

Users of data.gov can provide feedback about their use of a dataset in a number of ways - through email, social media (e.g. Twitter), or directly on the data.gov website. The latter option includes a button prominently placed at the top of each dataset that reads "Report Data Issue". Users selecting this option are given the ability to create a free text description of the issue, and publish the issue to data.gov for public viewing.

Data

Using the RVest package (Wickham, 2015) for R we scraped all data issues that were published to data.gov from January 01, 2015 - December 31, 2016 (n = 956). For each issue, we collected the date the issue was reported, the status of the issue (open or closed), and a user's free-text response to the "issue" prompt.

Inductive Coding

Our research team met and discussed themes that were reported in the free text of 20 randomly selected user issues. We then developed an initial coding scheme to classify each issue. Next, we selected a subset of issues from the scraped data (n = 50), and separately coded each issue with this initial coding scheme. We compared the applied codes for each issue, and revised our coding scheme to account for differences in our application of one

ASIST 2017, Washington, DC | Oct. 27-Nov 1, 2017

particular code over another. A new subset was selected (n = 50) and again coded separately. For the second round of coding we calculated a Kappa score (Carletta, 1996) to measure inter-coder reliability, achieving an agreement score of 0.93. We again simplified our coding scheme to account for differences in the application of codes, selected a third subset (n=100), and separately coded each issue. For the third round we achieved a kappa score of 0.94 - giving us confidence in the final coding scheme (presented in Table 1). Finally, we divided and separately coded the remaining issues (n= 856). The results presented below include only those issues coded during our third and final round of coding.

RESULTS

Users of data.gov reported issues in three broad categories – Data availability, Data Quality, and Documentation. Data availability accounted for 81% (n=610) of all issues reported, followed by unclear (as in the report was not specific enough to code) (n=136), data quality (n=73), and documentation (n=26) issues. As shown in figure 1, of the 610 reported data availability issues, over half were a result of a broken link / 404 error (n=327).

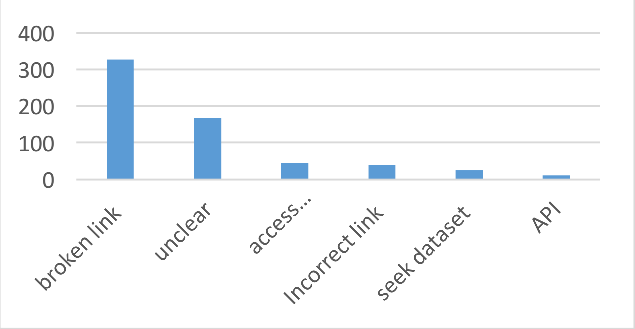

Figure 1. Distribution of reported data availability issues.

Data quality and documentation issues were reported less frequently. As shown in Figure 2, data quality issues were distributed somewhat evenly over around incorrect (n=25), obsolescent (n=13) and formatting problems (n=11).

DISCUSSION AND FUTURE WORK

The answer to our first research question is evident in the classification scheme that we inductively developed through coding user reported data issues; Users report problems with the availability, quality, and documentation of open data hosted by data.gov. Results from applying these codes show that a majority of the issues reported by users of data.gov are concerned with data availability, and in particular broken links for data that should be accessible via the data.gov portal.

This finding is unsurprising given that data.gov is designed to act, in part, as a content aggregator that harvests metadata from other open data portals, and provides a unified metadata standard and search mechanism. However, a significant number (n=168) of data availability reports were simply too vague or general for us to classify. For example, on June 11, 2015 a user's free text reports "Nothing happens when I click the download link" – it's unclear from a report like this whether a resource has moved, a link has been deleted, or the user simply has lost internet connection and their browser is not properly resolving the link.

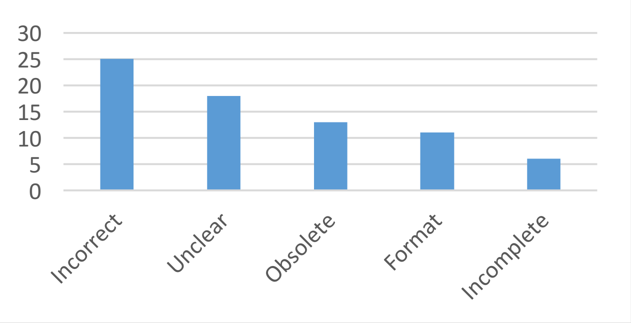

Figure 2. Distribution of reported Data quality issues

In total, almost half of all reported issues (n=323) were too vague or incomplete for us to classify. The answer to our second research question thus indicates a need to give users better direction in reporting data issues. In particular, a simple menu of choices – such as the classification scheme provided here - would likely help users more accurately report the problems that they face in using open data.

Contrary to our assumption, most users are not filing issues about the quality of the data they consume, but instead problems they encounter while trying to access data. This may be a result of the ease in filing an issue at the point of access (e.g. being denied access, one simply needs to click the "Report Data Issue" button). It may be too much to expect that a data consumer downloads a dataset, explores in some way, discovers an error, and then returns to the data.gov site to report the issue.

We therefore believe data quality issues may be reported in other locations – through email, social media, or on public forums, such as StackExchange, which has recently launched a dedicated question and answer site for developers and researchers interested in open data (https://opendata.stackexchange.com/). In future work we plan to also harvest user issues reported on these platforms, and apply the coding scheme developed in this preliminary work.

REFERENCES

Braunschweig, K., Eberius, J., Thiele, M., & Lehner, W. (2012). The state of open data. WWW2012, Lyon, France: ACM.

Code	Refinement	Definition
Data Availability		Issue reports a problem with the accessibility of data.
	broken link	404 or broken link is reported
	access forbidden	403 or lack of credentials is reported
	Incorrect link	The link provided resolves to a dataset or resource that is different from the one described
	API	A user reports an error with the data.gov API, or the failure of a wget method
	Seeking dataset	The user is requesting a dataset be made available
	Unclear	User has reported an issue with accessing a dataset, but it is unclear if the problem is on behalf of data.gov or a user error.
Data Quality		Issues report a problem with using available data
	obsolescence	The dataset is outdated, or has not been updated
	incomplete	The dataset is missing values
	incorrect	The dataset has errors, or the file is corrupt in some way
	format	The user reports an issue with the file format that the data are served
Documentation		Issues report a problem with the existing documentation for a dataset
	metadata	The user reports missing, incorrect, or incomplete metadata for a dataset
	contact information	The user reports missing, incorrect, or incomplete contact information for a dataset
Others or Unclear		Issue reports a problem that is out of scope for data.gov (e.g. irrelevant complaint about government secrecy)

Table 1: Coding Scheme used to classify data issues.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. Computational linguistics, 22(2), 249-254.

Chan, T.-C., Teng, Y.-C., Kuo, C., Yeh, Y.-H., & Lin, B.-C. (2017). Leveraging the Niche of Open Data for Disease Surveillance and Health Education. Online Journal of Public Health Informatics, 9(1).

Chen, Y., Wen, C.-Y., Chen, H.-P., Lin, Y.-H., & Sum, H.-C. (2011). Metrics for Metadata Quality Assurance and Their Implications for Digital Libraries. In ICADL (Vol. 7008, pp. 138–147). Springer.

Davies, T. (2010). Open data, democracy and public sector reform. A Look at Open Government Data Use from Data.Gov.Uk. (Dissertation)

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. Information systems management, 29(4), 258-268.

Martin, S., Foulonneau, M., Turki, S., & Ihadjadene, M. (2013, June). Open data: Barriers, risks and opportunities. In Proceedings of the 13th European Conference on eGovernment: ECEG (pp. 301-309).

Sáez Martín, A., Rosario, A. H. D., & Pérez, M. D. C. C. (2016). An international analysis of the quality of open government data portals. *Social Science Computer Review*, *34*(3), 298-311.

Neumaier, S., Umbrich, J., & Polleres, A. (2016). Automated quality assessment of metadata across open data portals. Journal of Data and Information Quality (JDIQ), 8(1), 2.

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. PeerJ, 1, e175.

Thorsby, J., Stowers, G. N., Wolslegel, K., & Tumbuan, E. (2016). Understanding the content and features of open data portals in American cities. Government Information Quarterly.

Wickham, H. (2015). rvest: Easily Harvest (Scrape) Web Pages. R package version 0.3. 1.

Umbrich, J., Neumaier, S., & Polleres, A. (2015, March). Towards assessing the quality evolution of open data portals. In Proceedings of ODQ2015: Open Data Quality: from Theory to Practice Workshop, Munich, Germany