Challenges and Supports for Accessing Open Government Datasets

Data Guide for Better Open Data Access and Uses

Fanghui Xiao
University of Pittsburgh
Pittsburgh, PA
fax2@pitt.edu

Daqing He
University of Pittsburgh
Pittsburgh, PA
dah44@pitt.edu

Yu Chi
University of Pittsburgh
Pittsburgh, PA
yuc73@pitt.edu

Wei Jeng
National Taiwan University
Taipei, Taiwan
wjeng@ntu.edu.tw

Christinger Tomer
University of Pittsburgh
Pittsburgh, PA
ctomer@pitt.edu

ABSTRACT

The importance of open government data is often associated with increased public trust, civic engagement, and accountable administrations. While there is a myriad of benefits, the existing literature suggests that many open government datasets lack accessibility and usability for diverse users. This study seeks to explore what contextual information users require when they access these datasets. Using mixed methods, we aim to discover the challenges of accessing data, and the necessary contextual information needed by the users to overcome these challenges. As the outcome of this study, we propose a framework called "Data Guides", which is composed of the identified important contextual information. In future work, we will test the effectiveness of the Data Guide in aiding users' accessing and understanding open government data.

CCS CONCEPTS

• **Information systems** → *Users and interactive retrieval*;

KEYWORDS

Open government data (OGD), structured data, accessibility, understandability, contextual information, Data Guides

ACM Reference Format:
Fanghui Xiao, Daqing He, Yu Chi, Wei Jeng, and Christinger Tomer. 2019. Challenges and Supports for Accessing Open Government Datasets : Data Guide for Better Open Data Access and Uses. In *2019 Conference on Human Information Interaction and Retrieval (CHIIR '19), March 10–14, 2019, Glasgow, Scotland Uk,* Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter (Eds.). ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3295750.3298958

1 INTRODUCTION

While the Internet and communication technology facilitates data sharing, there are numbers of online open data for general public or parties with commercial interests to access or consume. Many of those datasets are shared and managed by governments at various levels. Open Government Data (OGD) is available to anyone to access, reuse, and distribute without copyright restrictions[6, 13]. Being recognized as treasured resources for enhancing policy transparency and government accountability, OGD has been widely accessed for many projects aiming for both economic and social development [2, 11, 12]

However, there are challenges in accessing OGD. In 2012, Zuiderwijk et al. [17] reviewed existing literatures and found that "difficult to access" is a major impediment to OGD. At that time, the results were mainly attributed to the limited openness of the data (e.g., access to the data is restricted to a particular group of users), emphasizing more on the publishers' side. With the open data related policies and rules being developed since then, which contributed an increasing number of datasets available online, the accessibility barriers become more about locating and making use of the appropriate datasets. Thus the issues are discussed more from users' perspectives. For example, through working on Globe Open Data Index (GODI) 2016/2017, LÃďmmerhirt et al. [9] identified three critical barriers, and the most important one is about findability : "data is hard to find." Some example difficulties are: citizens have to check many different websites to find all the data they need; users have to try many queries to access the needed dataset due to bad naming or website indexing. Koesten et al. [8] also pointed out that finding datasets in many practical situations are not always straightforward and may need the datasets from different sources.

Unlike online webpages, datasets contain various forms of data often without sufficient contextual information. Therefore, conventional content indexing methods used in web search engines do not always work for the datasets [8].

Furthermore, although many datasets could come with certain metadata schema, which provides limited capabilities of text description, these structured descriptions presented in a metadata schema often suffer problems such as lacking detail descriptions, containing irrelevant elements for access, or missing certain important elements[4, 17, 18]. This is because metadata is mainly created

CHIIR '19, March 10–14, 2019, Glasgow, Scotland Uk
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6025-8/19/03…$15.00
https://doi.org/10.1145/3295750.3298958

for collection management or long-term preservation, which often is not fully associated with supporting access for a particular user.

Besides the above general challenges to access datasets, accessing OGD also has some unique difficulties. Users who want to access OGD can be an individual citizen who just wants to know more about the public work in her neighborhood, or a commercial startup company aims to build real-time traffic alert Apps on mobile devices. The former user would mostly be a user with little technology capabilities or data literacy[4], whereas the latter user could be an expert on data manipulation or analysis. Both groups of users should be supported in their access to OGD, which presents interesting, important and open challenges to the designers and managers of OGD repositories.

In addition to the challenge of accessibility, understandability is also one of the major difficulties for OGD users. For example, Zuiderwijk et al. [17] and Janssen et al. [4] pointed out that the main reasons that cause the difficulties of understanding OGD include lacking of information to interpret data, lacking of explanation for the meaning of data, and lacking of knowledge to make sense of data. If users cannot make sense of data, they would not be able to evaluate the data; let alone using the data they found.

Existing literatures suggest that guidance and help with understanding the content within OGD (e.g., descriptions of categories, what data is present, and so on) are important supports for improving the accessibility and understandability [13, 15]. For example, within the context of land use, Verburg et al. [14] stated that it is essential to have clear and extended documentation to help users understand the data they are engaging with. Janssen et al. [4] pointed out "an infrastructure is necessary which helps users to make sense of data." Koesten et al. [8] found out more detailed information on *how the original data was collected* can aid users in making decisions of trusting in the data. These studies emphasized the importance of providing information and context surrounding OGD. Therefore, we believe that a more detailed document beyond metadata should be utilized to help users in their access to and make use of the OGD.

However, what exactly the information should be included in the documentation is unrevealed. There has no systematic investigation on the specific elements in the documentation. In this paper, we call this type of documentation *Data Guide*, and we aim to examine the necessity of the Data Guide, and to explore the important content components for the Data Guide. Specifically, the research questions are: RQ1: what are the specific challenges when users accessing OGD? RQ2: what information should be included in the Data Guide to enable users better access and use OGD?

2 RESEARCH METHODS

An iterative exploratory-sequential mixed-method design is adopted in this research project. A semi-structured interview was conducted first to obtain deeper insights into the users' needs. Then, the results of the interview, the sense-making theories [1, 3, 10, 16] as well as the framework of human structured-data[8] were adopted to guide the development of the questionnaire. Finally, the instrument was disseminated to collect the perspectives of the OGD users.

There are two reasons that motivated us to start our study with an interview. Firstly, the qualitative approach can improve our understandings of the specific challenges that the users might encounter and the supports they expect for overcoming the challenges when accessing OGD. Secondly, as there is little research on systematically investigating contextual information for helping the accessibility and understandability of open data, we lacked sufficient guidance to frame a larger-scale quantitative survey. We therefore combining the results of semi-structured interviews with sense-making theories and a framework of human structured-data interaction to construct a theoretical basis for a web-based questionnaire survey.

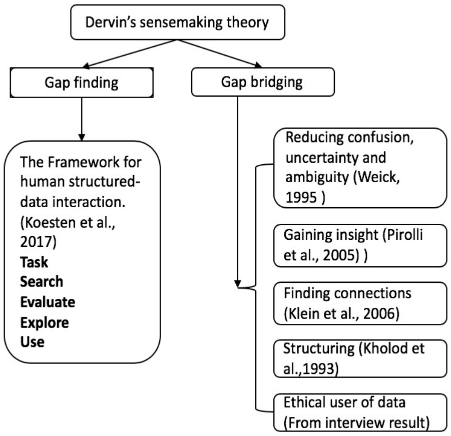
Dervin's sensemaking theory
Gap finding
Gap bridging
The Framework for human structured-data interaction. (Koesten et al., 2017)
Task
Search
Evaluate
Explore
Use
Reducing confusion, uncertainty and ambiguity (Weick, 1995)
Gaining insight (Pirolli et al., 2005))
Finding connections (Klein et al., 2006)
Structuring (Kholod et al.,1993)
Ethical user of data (From interview result)

Figure 1: The main structure of the questionnaire

Overall, the theories and framework guide us to build the structure of questions, and the interview results provide the options of each question. The structure of Gap finding and Gap bridging in the questionnaire was designed, as shown in Figure 1, and we will illustrate the design process in the following sections.

2.1 Questionnaire Design

The questionnaire includes four sections: the challenges of accessing OGD (Gap finding), the expected contextual information (Gap bridging), participants' data literacy skill, and their demographic information. The motivations for this design are several folds. Firstly, we adopted Dervin's sense-making theory as the fundamental theory, in which she claimed that one needs to identify the gap (i.e., gap finding) and make use of the corresponding bridge to achieve the outcome of moving to the other end of the gap via the bridge (i.e., gap bridging) [3]. We believe that accessing a relevant dataset to satisfy one's particular need follows a similar process too. Besides, we collected the participants' data literacy skill and the demographic information to investigate whether these factors would affect users' access to data. In this preliminary paper, the analysis focused on the questions of gap finding and gap bridging, while the data of the other two sections will be analyzed in the future.

After completing the interviews, we performed an initial comparative analysis to examine transcripts for the interviews. First, we identified the challenges and useful elements that brought up by interviewees, and then the information was coded into different categories (see Table 1 & Table 2, 2nd Column).

2.1.1 Questions regarding Gap Finding. To bridge the users' gaps in accessing data, we investigated what their gaps are first. Koesten et al. [8] conducted a study to see the information seeking behaviors of people searching for the sources of structured data online, including the behaviors of data search and the identification of relevant datasets. According to the findings, they, ultimately, proposed a framework for human structured-data interaction that concludes the whole process of users interacting with OGD. Therefore, we adopted this framework to design questions to learn the difficulties during the entire processes of accessing data. Based on the framework for human structured-data interaction and the interview results, we identified 9 challenges that were presented in the questionnaire. We also have open-ended questions to ask if they have other ideas, but no other challenges were brought up, so we assume the options we provided are sufficient. (see Table 1):

Table 1: Gap finding questions and corresponding options

Perspectives	Options
Search	A. Forming queries for searching
Evaluate	B. Finding the content-relevant datasets C. Finding the usable datasets
Explore	D. Knowing what data is about in the datasets that I found E. Inferring deeper information according to the dataset that I found F. Interpreting the visualizations that websites provide G. Trusting the credibility
Use	H. Using the tools or functions within the portals that are provided for users using data, such as analyzing data or visualizing data I. Using the data to my work

2.1.2 Questions regarding Gap Bridging. Next, we began to find out the useful information to bridge the gaps. Alsufiani et al. [1] examined the different sensemaking theories that proposed by different scholars, including reducing confusion, uncertainty and ambiguity [16], gaining insight [10], finding connections [7], structuring [1], and gap-finding & gap-bridging [3]. They proved that sensemaking involves all these processes. Therefore, we adopted these theories to design the questions regarding gap bridging. In addition, 12 elements were collected from the interview results and were proposed in the questionnaire.

For both Gap Finding and Gap Bridging questions, we employed typical five-level Likert scores, from *Not at all challenge/important* (score 1) to *Extremely challenge/important* (score 5).

2.2 Data Collection

The participants of our study mainly came from users population of a regional data center in the northeast region of the US. The data center serves mainly to local communities, and currently hosts hundreds of datasets (such as transportation and environment data) from the local county, other public-sector agencies, academic institutions, and nonprofit organizations.

We adopted the snowball sample method for recruiting participants with the experience of using open data. We made initial contacts with users who are qualified and willing to participant the interviews and the questionnaire. Then, these participants recommended their colleagues or other suitable people to be the new subjects. In total, we interviewed 14 users (9 females and 5 males, 3 researchers, 3 data scientists, 6 librarians, 1 educator, and 1 business people, age from 18 to 45 years old), and collected 14 valid responses to our questionnaire (3 females and 11 males, 6 students, 3 researchers, 2 developers, 1 librarian, and 1 data analyst, age from 18 to 45 years old) by the end of September 21, 2018.

Table 2: Gap Bridging questions and corresponding options.

Reducing confusion, uncertainty and ambiguity. Weick [16]	A. Data collection and creation process B. More detailed description of the dataset C. Information about the publisher D. Additional information, such as tips, tricks, and cautionary notes
Gaining insight. Pirolli and Card [10]	A. The publisher's purposes for collecting and publishing the datasets B.Examples of how the data has been used C. Potential uses of a dataset
Finding connections. Klein et al. [7]	A.The relationships (scenario-based connections) with other datasets within the same data center in the datasets that I found B. Cross-linking to other sites' data that related to the data that you are looking at
Structuring. Alsufiani et al. [1]	A. Database schema with a dataset or a database B. The governance history of data formats
Ethical use of data (From interview results)	How to cite the dataset that you want to use

3 RESULT ANALYSIS

In total, we distributed 36 invitations of the questionnaire, and received 14 valid responses, with a response ratio of 38.9%. Because we focus on a special group of people, who are the users of the local OGD site, the initial contact size was limited, so the final sample size is small despite we actually had a pretty high response rate. Reviewing the current positions of the participants, we found that four of them are data scientists who may search or use data frequently, but the rest may not access data very often in their work. Therefore, it is reasonable to think that our participants are similar to general dataset search users, who have the information needs for accessing or using datasets, but do not have much experience with dataset search

3.1 Top challenges of accessing OGD

Table 3 presents the five most challenging processes. The most difficult one is *inferring deeper information.* Also, our results, such as *Finding the usable/content-relevant datasets*, and *Using the data to my work* are consistent with Koesten et al. study in 2017[8]. They learned that people have difficulties in finding datasets, and lack of the information for evaluating the fitness of use, or the information that helps interpret data out of context. This finding would help data portals learn more about their users' difficulties, and then develop training or tools to help users overcome these challenges.

3.2 The important components of a Data Guide

In the questionnaire, we asked the question *What are the following statements describing the next action(s) you would probably take when you cannot understand a dataset? (Select all that apply).* All 14 participants chose answer *Look for related documentations within data portals.* This result further confirms the need that mentioned

Table 3: The top five challenging process

Processes	Mean	Mode	SD
Inferring deeper information according the dataset that I found	3.7	4	0.9
Using the data to my work	2.9	3	1.0
Finding the usable datasets	2.9	2	1.1
Finding the content-relevant datasets	2.8	2	1.1
Knowing what data is about in the datasets that I found	2.7	3	1.0

in the previous literatures that a clear and extended documentation to help users understand the data they are engaging with.

Table 4 presents the important elements that participants selected for conquering their difficulties. *More detailed description of the data* has been recognized as the most essential element (Mean=4.1, Mode=5, SD=0.9). In fact, some elements are almost equally important, such as *Examples, Potential uses, and Data collection and creation process.* Additionally, we built the connections between the challenges and the supportive elements (see Figure2).

Table 4: The important contextual information

Elements	Mean	Mode	SD
More detailed description of the dataset, such as data metrics	4.1	5	0.9
Additional information, such as tips, tricks, and cautionary notes	3.6	3	0.9
Database schema	3.5	3	1.2
Examples of how the data has been used	3.4	4	1.2
Potential uses of a dataset	3.4	4	1.3
Data collection and creation processes	3.4	4	1.1
How to cite the dataset that you want to use	3.4	3	1.2
Cross-linking to other sites' data that related to the data that you are looking at	3.4	3	1.2
The relationships (scenario-based connections) with other datasets within the same data portal	3.1	2	1.3
The publisher's purposes for collecting and publishing the datasets	2.9	2	1.3
Information about publishers	2.4	2	1.0
The governance history of data formats	2.1	2	0.9

4 DISCUSSION

Components of a "Data Guide". Based on the survey and interview results, we propose the best practices for composing a Data Guide for OGD platforms with twelve preferred components: 1) detailed information of data collection and creation processes, 2) more detailed description of the dataset, 3) information about publishers, 4) additional information such as tips and cautionary notes, 5) data publishers' purposes for collecting and publishing the datasets, 6) examples of how the data has been used, 7) potential uses of a dataset, 8) the relationships (scenario-based connections) with other datasets within the same data portal, 9) cross-linking to other sites' data that related to the data that a user is looking at, 10) database schema, 11) the way to cite the dataset, and 12) the governance history of data formats. It is worth noting that these guidelines were confirmed with the real-world OGD users' perception.

Managerial Implications. Not surprisingly, participants from different perspectives reported what they value more differently. For instances, we found that researchers indicated that 1) detailed information of data collection and creation processes are critical since it can help evaluate the trustworthiness of a certain dataset. These findings are aligned with the prior work based on research data [5]. In contrast, the participants who reported as developers in the interview paid more attention to 12) the governance history of data formats and 10) database schema. We thus encourage data publishers and data portals to consider distributing more contextual information by following the Data Guide along with their data publishing process. However, to avoid the OGD users being overwhelmed by too much information appending to a Data Guide itself, we suggest the components in the Data Guide to be presented optionally and can be manipulated by individual users, e.g., they could collapse and expand some certain components.

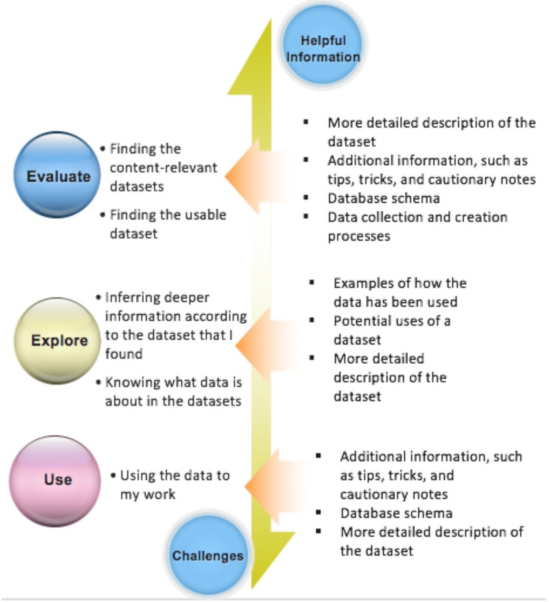
Helpful Information
Evaluate
• Finding the content-relevant datasets
• Finding the usable dataset
▪ More detailed description of the dataset
▪ Additional information, such as tips, tricks, and cautionary notes
▪ Database schema
▪ Data collection and creation processes
Explore
• Inferring deeper information according to the dataset that I found
• Knowing what data is about in the datasets
▪ Examples of how the data has been used
▪ Potential uses of a dataset
▪ More detailed description of the dataset
Use
• Using the data to my work
▪ Additional information, such as tips, tricks, and cautionary notes
▪ Database schema
▪ More detailed description of the dataset
Challenges

Figure 2: The connections between the challenges and the contextual information

Increasing the opportunities of accessing OGD. For the online dataset search, especially numerical data, the related descriptive information is usually insufficient. From this perspective, this Data Guide will enhance the search engines' performance in datasets search by providing the related contextual information.

5 CONCLUSION & FUTURE WORK

In this study, we examined the challenges for users to access open government data (OGD), and we also collected the important contextual information that could facilitate users in their accessing to OGD. The results of the study help us to design a basic framework, which we called Data Guides, to help users.

For future work, we have a short-term and a long-term plan. Our short-term plan includes increasing our sample size and conducting further analysis on whether there are associations between users' perceived challenges and their perceived data literacy skills, as well as between users' occupations and their needs for ODG. For the longer-term goal, we aim to validate and improve the Data Guide by working with professionals affiliated with OGD centers.

ACKNOWLEDGEMENT

This work was partially supported from the Young Scholar Fellowship Program by Ministry of Science and Technology (MOST) in Taiwan, under Grant MOST107-2636-H-002 -002.

REFERENCES

[1] Kholod Alsufiani, Simon Attfield, and Leishi Zhang. 2017. Towards an instrument for measuring sensemaking and an assessment of its theoretical features. In *Proceedings of the 31st British Computer Society Human Computer Interaction Conference*. BCS Learning & Development Ltd., 86.

[2] Ahmad Assaf, Raphaël Troncy, and Aline Senart. 2015. HDL-Towards a Harmonized Dataset Model for Open Data Portals.. In *USEWOD-PROFILES@ ESWC*. 62–74.

[3] Brenda Dervin. 1999. Chaos, order and sense-making: A proposed theory for information design. *Information design* (1999), 35–57.

[4] Marijn Janssen, Yannis Charalabidis, and Anneke Zuiderwijk. 2012. Benefits, adoption barriers and myths of open data and open government. *Information systems management* 29, 4 (2012), 258–268.

[5] Wei Jeng, Eleanor Mattern, Daqing He, and Liz Lyon. 2016. Unpacking the ”Black Box”: A Preliminary Study of Visualizing Humanists and Social Science Scholars' Data and Research Processes. *Proceedings of iConference* (2016).

[6] Maxat Kassen. 2013. A promising phenomenon of open data: A case study of the Chicago open data project. *Government Information Quarterly* 30, 4 (2013), 508–513.

[7] Gary Klein, Brian Moon, and Robert R Hoffman. 2006. Making sense of sensemaking 1: Alternative perspectives. *IEEE intelligent systems* 4 (2006), 70–73.

[8] Laura M Koesten, Emilia Kacprzak, Jenifer FA Tennison, and Elena Simperl. 2017. The Trials and Tribulations of Working with Structured Data:-a Study on Information Seeking Behaviour. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 1277–1289.

[9] Danny LÃďmmerhirt, Oscar Montiel, and Mor Rubinstein. 2017. The State of Open Government Data in 2017. (2017).

[10] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, Vol. 5. McLean, VA, USA, 2–4.

[11] Jeffrey Thorsby, Genie NL Stowers, Kristen Wolslegel, and Ellie Tumbuan. 2017. Understanding the content and features of open data portals in American cities. *Government Information Quarterly* 34, 1 (2017), 53–61.

[12] Barbara Ubaldi. 2013. Open government data: Towards empirical analysis of open government data initiatives. *OECD Working Papers on Public Governance* 22 (2013), 0–1.

[13] Nataša Veljković, Sanja Bogdanović-Dinić, and Leonid Stoimenov. 2014. Benchmarking open government: An open data perspective. *Government Information Quarterly* 31, 2 (2014), 278–290.

[14] Peter H Verburg, Kathleen Neumann, and Linda Nol. 2011. Challenges in using land use and land cover data for global change studies. *Global Change Biology* 17, 2 (2011), 974–989.

[15] Pieter Verdegem and Gino Verleye. 2009. User-centered E-Government in practice: A comprehensive model for measuring user satisfaction. *Government information quarterly* 26, 3 (2009), 487–497.

[16] Karl E Weick. 1995. *Sensemaking in organizations*. Vol. 3. Sage.

[17] Anneke Zuiderwijk, Marijn Janssen, Sunil Choenni, Ronald Meijer, and Roexsana Sheikh Alibaks. 2012. Socio-technical Impediments of Open Data. *Electronic Journal of e-Government* 10, 2 (2012).

[18] Anneke Zuiderwijk, Keith Jeffery, and Marijn Janssen. 2012. The potential of metadata for linked open data and its value for users and publishers. *JeDEM-eJournal of eDemocracy and Open Government* 4, 2 (2012), 222–244.