IJDC | *General Article*

Emerging Roles for Optimising Re-Use of Open Government Data

Fanghui Xiao
School of Computing and Information,
University of Pittsburgh

Liz Lyon
School of Computing and Information,
University of Pittsburgh

Ning Zou
School of Computing and Information,
University of Pittsburgh

Robert M. Gradeck
Center for Social and Urban Research,
University of Pittsburgh

Abstract

This paper describes a small-scale study to investigate the missions, services and operational tasks provided by four open government data centers: NYC OpenData (New York Open Data Center), DataSF (open data portal of San Francisco), WPRDC (Western Pennsylvania Regional Data Center) and the London Datastore (Greater London open data portal). The findings are used to propose three emerging specialist data roles for open government data (OGD) centers. The methodology used was an analysis of the textual content of the data center websites to identify the common elements of the mission and services. A common mission across all four open government data centers was 'to improve the use of data'. The range of data center services and tasks identified and extracted from the websites could be classified into five common categories: Availability, Understandability, Technical Help, Social Engagement, and Improve User Data Literacy. Three new specialist open government data roles were proposed, which were framed to facilitate the delivery of the services identified in this study: Data Interpreter, Data Consultant and Data Visual Assistant. In parallel with existing research data policies and guidelines, these three specialist OGD roles could be extended and applied across other open data portals and domain-based data centers, including research data repositories, to optimise the delivery of open data, to facilitate greater value from data sharing, to maximize the understanding of complex data and to minimize the subsequent misuse of data.

Received 22 January 2018 ~ *Accepted* 22 January 2018

Correspondence should be addressed to Liz Lyon, University of Pittsburgh, School of Computing and Information, 135 N Bellefield Ave, Pittsburgh PA 15260. Email: elyon@pitt.edu

An earlier version of this paper was presented at the 13th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v13i1.609
DOI: 10.2218/ijdc.v13i1.609

Introduction

Acknowledging the value of Open Government Data (OGD), open data centers have been rapidly proliferating in the United States, Europe, and Asia. These centers publish increasing volumes of datasets which have been collected or used by governments, such as transportation and environmental data. With the growth of OGD, the functions of the associated infrastructure platforms are not limited to simply supporting data accessibility. The broader use of these data has become the main goal of the centers; indeed the full value of the centers cannot be realized until these datasets are widely used. Manyika et al. (2013) suggest that by applying advanced data analytics, citizen use of open data could produce $3.2 to $5.4 trillion in economic value per year across several domains. Therefore, in order to empower the use of open government data, these data should not only be available in consistent and easily usable formats, but also be understandable. Given this challenging goal of many OGD projects and the current types of open data (mainly complex quantitative data), open data centers have created several positions so that data portals function efficiently. These roles may be divided into general roles and specialist roles. General roles include data center manager, programmer, data analyst, and training specialist. However, new specialist job types are also appearing, which are the primary focus of this study.

In this paper, we have explored four local level OGD centers: NYC OpenData[1] (New York Open Data Center), DataSF[2] (open data portal of San Francisco), WPRDC[3] (Western Pennsylvania Regional Data Center) and the London Datastore[4] (Greater London open data portal). Local level portals were selected for this study because these platforms are likely to be more connected with civic organizations, neighborhoods, and communities. The field of information and data science has a growing interest in this domain, since the diverse challenges of curating and managing these data, facilitating access and reuse of data through dedicated user tools and services, plus the need to train people to improve their information or data literacy skills, are critical current themes for iSchool research and education programs. Three research questions are addressed here:

1. What are the common missions of open government data centers?
2. What user services and supportive tasks are provided by open government data centers?
3. Which specialist job types are needed to deliver these OGD services?

We begin this paper with a Literature Review followed by sections describing the Methodology, Results and Data Analysis, Discussion, Conclusions and Next Steps.

[1] NYC OpenData: https://opendata.cityofnewyork.us/overview/
[2] DataSF: https://data.sfgov.org/about
[3] Western Pennsylvania Regional Data Center: http://www.wprdc.org/about
[4] London Datastore: https://data.london.gov.uk/about/

Literature Review

There has been much prior discussion of the requirements to develop workforce capacity and capability for data science and data stewardship (Lazer et al., 2009; Pryor and Donnelly, 2009; Bakhshi, Mateos-Garcia and Whitby, 2014; National Research Council, 2015). This literature has also explored the nature and functions of a range of supporting roles and positions, using a varied taxonomy to categorize the different job types. Six broad data scientist roles were described by Lyon and Brenner (2015) – data analyst, data archivist, data engineer, data journalist, data librarian, data steward/curator – and their likely organizational locations plus a brief summary of their key tasks were proposed. These data science roles were explored in more depth in two further studies (Lyon, Mattern, Acker and Langmead, 2015; Lyon and Mattern, 2016). These two reports describe an analysis of the real-world requirements for a range of positions across different job sectors and highlight the specific qualifications, knowledge, experience, skills and competencies for each role.

Despite the substantive research on broader data science roles, there appears to be a lack of research on these roles within the open government data context. Since the development of Open Government Data initiatives, and in particular the development of OGD portals, which have proliferated since the mid-2000s both at federal and local government levels, governments are actively seeking ways to make their data more easily accessible, usable and re-usable by all (Ubaldi, 2013). One complex challenge of open data is understandability; sometimes, data users find that it is difficult to interpret the data. The data in open data platforms are most often available in raw data formats (Weerakkody et al., 2017); also the users are unfamiliar with definitions or categories that are adopted to present the data. Another challenge is that users are required to have a certain level of skills to use the data (Kapoor, Weerakkody and Sivarajah, 2015). In general, user studies have found that the potential open data users lack the professional knowledge or skills to interpret or use the data (Martin, 2014; Janssen et al., 2012).

This current study seeks to begin to remedy the lack of research around OGD roles and to contribute to the field by providing a small-scale analysis of selected OGD portals, their associated user services and requirements for supporting data roles.

Methodology

This study focuses on four local-level open data portals: NYC OpenData (New York Open Data Center), DataSF (open data portal of San Francisco), WPRDC (Western Pennsylvania Regional Data Center) and London Datastore (Greater London open data portal). These four particular open data platforms were chosen by considering the following perspectives. First, city size, scale and geographical distribution: New York City (NYC), the City of San Francisco (SF), Pittsburgh and London are substantial metropolitan urban areas. NYC and SF are located in the east and west of the United States respectively, Pittsburgh is located in a more central US location and London is an international city in the United Kingdom. Taken together, these four cities represent a broad geographical spectrum, whilst all being cities of significant size with substantive local citizen populations. As a result, the OGD centers within these cities are able to collect and provide access to large amounts of data through their infrastructure platforms and services. A second perspective is the maturity of these OGD platforms. For example, DataSF was launched in 2009, the original London Datastore was launched in 2010 (Arthur, 2010) and NYC OpenData was set up in 2012. As a result, these centers have been exploring and developing the methods and services which that can facilitate data reuse for many years. The Pittsburgh-based WPRDC was established in 2015, and whilst it is the most recently-established OGD center, it references many efficient operational methods, standards and data practices from those relatively mature data platforms. Therefore, WPRDC is a well-formed OGD center. Furthermore, from the perspective of familiarity, the first author worked for WPRDC for a year as a graduate student researcher, and the fourth author is the current director of WPRDC. As a result, we have an excellent understanding of the work of the WPRDC and of open government data centers in general.

The methodology utilized a content analysis of the four selected open government data center Web sites. The content analysis collected, examined and analyzed three key classes of information, including the mission statements, the range of user services, and the associated supporting tasks provided by the four portals. In the first step, to collect data for the in-depth analysis, one coder manually examined and extracted the three classes of information from the four platforms' official websites on January 10th, 2018. For each website, the coder first located and identified the relevant information, and then classified this information into different categories according to the thematic similarity. The coding results were then verified by the second coder, to ensure that both coders were working consistently. All the collected data was stored in a MS Excel spreadsheet and then manually analyzed. The four websites examined are listed below in Table 1.

Table 1. The OGD official websites.

OGD Portal	Website
NYC OpenData	https://opendata.cityofnewyork.us
DataSF	https://datasf.org
WPRDC	http://www.wprdc.org
London Datastore	https://data.london.gov.uk

To further illustrate the text extraction and analysis process, the coder extracted raw data about the missions of the selected OGD centers, the services they provided, and the tasks they perform to support those services. In this first step, the coder simply collected relevant data into the three classes, Table 2 shows an example of part of this data collection.

Table 2. OGD website data collection exemplar.

ODG Portal	Mission	User Service	Support Task
DataSF	Our mission is to empower use of data. At DataSF, we seek to transform the way the City works through the use of data.	Fundamental: search for data; browse by Category, Publisher, View Types, and Tags; guideline of API documentations; training for enhance skills in data use; data management and process improvement; Metadata: data dictionary; ask questions; leave comments and request datasets Technologies: APIs, and code; show case.	1. Developing a catalog of datasets that can be made public 2. Monitoring and responding to comments on public datasets 3. Monitoring and tracking the rollout of public datasets on the open data portal 4. Ensuring that processes are followed to exclude private, confidential or proprietary data.
…	…	…	…

In a second step, based on the nature of the data center services and tasks, the coder classified this information into one of five categories: Availability, Understandability, Technical Help, Social Engagement (Interactive) and Improving User Data Literacy. Table 3 explains the five categories.

Table 3. The five categories used to classify OGD website content.

Category	Rationale
Availability	Availability is viewed as the first step in releasing data. To ensure the data are open, the government data should be available to anyone who wants to access and use them i.e. easily discoverable and downloadable.
Understandability	Understandability refers to providing more information (such as contextual information) or tools, for assisting users to understand the data and use them. London Datastore claims that “[r]aw data often doesn’t tell you anything until it has been presented in a meaningful way and most people don’t have the tools to do this.”
Technical Help	Technical Help focuses on creating APIs or tools to better serve various types of data users to use and apply data, such as software engineers and interested citizens.
Social Engagement (Interactivity)	Social Engagement and interaction is seen as one of the significant purposes for publishing government data. User inquiries or comments can also help local government to improve service provision.
Improve User Data Literacy	Enhancing user data skills is a necessary requirement for empowering data use. This work could consist of training and answering questions.

Results and Data Analysis

The missions extracted from the official websites of the chosen OGD centers are shown in Table 4.

Table 4. Missions of selected open data platforms.

ODG Portal	Mission	Source
NYC OpenData	Open Data is an opportunity to engage New Yorkers in the information that is produced and used by City government.	https://opendata.cityofnewyork.us/overview/
DataSF	Our mission is to empower use of data. At DataSF, we seek to transform the way the City works through the use of data.	https://datasf.org/about/
WPRDC	Our goal at the Western Pennsylvania Regional Data Center is to make community information easier to find and use.	http://www.wprdc.org/performance- management/
London Datastore	We want everyone to be able access the data that the GLA and other public sector organizations hold, and to use that data however they see fit – for free	https://data.london.gov.uk/about/

Based on the missions, the four platforms provide corresponding common services for data users. The extracted services were selected based on at least three platforms offering similar functions. These functions were then mapped onto one of the five service categories as shown in Table 5.

Table 5. Services provided by selected open data platforms.

Availability	Understandability	Technical Help	Social Engagement (Interactivity)	Improve User Data Skills
Platform for publishing data	Guidelines for using the platform for new users	APIs		Training to enhance skills in data use
Search for data	Guidelines for API documentation for data veterans		Ask questions, leave comments and request datasets	Providing technological training-related information
Browse by category, publisher, new datasets, popular datasets	Metadata and data dictionary	Data visualization tools		
Download data	Showcases (data analysis report)			Help desk

The data for supportive tasks was primarily extracted and analyzed from WPRDC and DataSF websites, because only these two data centers provided detailed information about their staff and their work on the official websites. The extracted tasks were classified into the same categories as the services, since the tasks that the OGD centers have performed are to directly support the services. The classification of the specific tasks is illustrated in Figure 1 below.

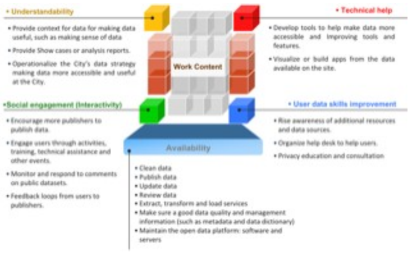
• Understandability
• Provide context for data for making data useful, such as making sense of data
• Provide Show cases or analysis reports.
• Operationalize the City's data strategy making data more accessible and useful at the City.
• Technical help
• Develop tools to help make data more accessible and improving tools and features.
• Visualize or build apps from the data available on the site.
Work Content
•Social engagement (Interactivity)
• Encourage more publishers to publish data.
• Engage users through activities, training, technical assistance and other events.
• Monitor and respond to comments on public datasets.
• Feedback loops from users to publishers.
• User data skills improvement
• Rise awareness of additional resources and data sources.
• Organize help desk to help users.
• Privacy education and consultation
Availability
• Clean data
• Publish data
• Update data
• Review data
• Extract, transform and load services
• Make sure a good data quality and management information (such as metadata and data dictionary)
• Maintain the open data platform: software and servers

Figure 1. Supporting tasks performed by the open data platforms.

Discussion

Returning to the first of our research questions, in Question 1 we asked: What are the common missions of open government data centers? From the extracted content describing the missions, we can see that the four OGD platforms have a common mission that is '*to improve the use of data*'. Whilst this common mission is expressed and articulated using subtly different semantic language, the ultimate goal is the same for each open government data center (Figure 2).

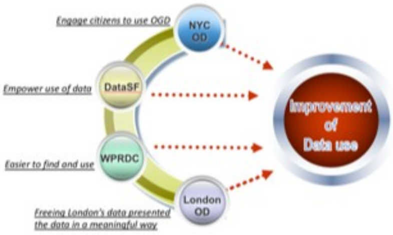

Figure 2. The common mission of the four open data platforms.

Our second research question asked: What user services and supportive tasks are provided by open government data centers? In order to achieve their common mission, the OGD platforms have set out to provide a range of services which are not limited to simply publishing data. In addition to the fundamental work of ensuring access to data (i.e. availability), the platforms carry out many micro-practices for increasing the use of data. For example, to help to make the data understandable, platforms have begun to provide showcases and data analysis reports to help users learn about the data. Some data centers have produced user guides in addition to providing metadata about their datasets and a data dictionary. The WPRDC offers Data Guides that contain contextual information about datasets. The Data Guides are primarily created for assisting users in making sense of the open data, and in particular about the complex quantitative datasets. Additionally, the four centers support users to visualize the datasets available through their platforms, by using a range of online tools. Although this function is still under development, it represents a trend which OGD platform developers are following. For improving user data literacy, some data centers provide Help Desk support, answer user questions and deal with many technical issues. All of these services contribute to improving the use of the data, as stated in the common mission.

The third and final research question asked: Which specialist job types are needed to deliver these OGD services? Our content analysis of the four OGD websites and the identification of the specific services provided and operational tasks, have led us to propose three specialist open government data roles or positions, which are described here in more detail.

Data Interpreter: The goal of this role is '*to make sense of data for users'*. A data interpreter's specific activities consist of working with data providers to create data guides, collecting and creating data-related blogs and data stories, working independently or with programmers to make maps or other visualizations, and informing data-related policies.

The role starts with open government data. A data interpreter is responsible for interpreting data in various ways, such as providing contextual information. The data will then be more explainable and understandable. In addition, the interpreted information not only helps users understand data, but also lowers the concern of data providers regarding misinterpreted data.

Data Consultant: The goal of this role is '*to directly assist users to understand the data and teach them technical skills to accurately use the data'*. This job requires that the consultant will hold help-desk hours each week to help users who have difficulties with the data, especially from the perspective of technology; organizes meetings to collect information from various groups of people, including the data category needs or the required tools, and then finds the technical solutions to meet their needs.

This role starts with OGD users. One of the goals of OGD platforms is to reach more citizens and thus to increase the use of open data. Most of the OGD is raw data, and using the raw data requires a certain level of data processing skills. However, the data literacy levels of OGD staff are often very different from those of the public. Data literacy levels may also vary between different members of the public. Hence, data consultants can directly help users to understand and use data based on the specific user questions, and then ultimately improve public data literacy.

Data Visual Assistant: The goal of this role is '*to assist users to visualize or manipulate data by developing tools that can be directly and easily used by users'*. This role's focus is to develop software tools and apps that can (easily) create data visualizations, such as line charts, bar charts, maps and other infographics. Users can then use the graphical tools and apps to create the specific visualizations they want by simply selecting the parameters and applying them to the whole dataset or data sub-set.

This role starts with tools. A data visual assistant makes the open government data more visible, more discoverable, more understandable and helps civic users to get credit for their 'mash-ups'.

Furthermore, these three OGD roles do not operate in isolation; rather they work together as an effective OGD team; the data interpreter and the data visual assistant both help users to gain new insights through expert exemplar interpretation and the provision and application of customized user tools. The data consultant offers public users professional help, to assist them in acquiring an in-depth understanding of open government data and its value, plus the opportunity to enhance and build their own individual data skills. Figure 3 summarizes the connections between the OGD mission, user services, supporting tasks and the new specialist OGD roles which, it is hoped, will greatly contribute to solving the OGD challenges previously identified in the literature.

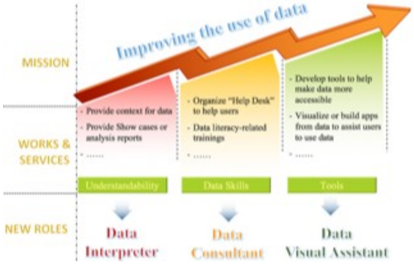
Improving the use of data
MISSION
WORKS & SERVICES
NEW ROLES
- Provide context for data
- Provide Show cases or analysis reports
- ......
- Organize "Help Desk" to help users
- Data literacy-related trainings
- ......
- Develop tools to help make data more accessible
- Visualize or build apps from data to assist users to use data
- ......
Understandability
Data Skills
Tools
Data Interpreter
Data Consultant
Data Visual Assistant

Figure 3. Critical inter-relationships for optimizing the re-use of open government data.

Looking beyond open government data, although scientists agree with the potential benefit of data sharing/reuse for scientific progress, the majority are reserved when it comes to practical implementation. Researchers who are reluctant to share data with others, reported major concerns with legal issues, misuse of data, and incompatible data types (Tenopir et al., 2011). In spite of many research data centers and publishing platforms (Scientific Data, F1000Research, DataOne, etc.) offering data policies and guidelines in support of data sharing/reuse, there are still a large number of researchers responding that there is a risk that data may be misinterpreted due to the complexity of data (Tenopir et al., 2015). This risk will be reduced when data interpreters work with data providers to create contextual information for a particular data set. The OGD specialist data roles identified in this study, could also be applied to research data centers and repositories. In collaboration with existing research data policies and guidelines, data interpreter/data consultant roles in each domain could help to maximize the understanding of complex data and minimize the subsequent misuse of data.

Conclusions and Next Steps

This exploratory research has proposed three new and specialist OGD roles, and builds on prior work which has described generic data science roles. We acknowledge that this has been a small-scale study examining the websites of just four open government data centers, however the methodology used in this study could be extended and applied across other open data portals, to provide a more substantive baseline reference. Furthermore, the findings of this study may provide valuable indicators for open government data portal managers in developing strategy, planning operational services and in allocating resources for new positions to deliver on such plans. The high-level descriptions for the three new specialist roles, together with description of the micro-tasks which they may deliver, provide a good foundation for putative job descriptions for open government data centers to use in the future.

We plan to carry out a further study to investigate the concrete skills, competencies and knowledge that are required for the three specialist OGD roles proposed in this paper. We believe that the role requirements would not only contribute to OGD centers to help these organizations to effectively find suitable candidates and to develop their work plans, but will also benefit data science educators e.g. faculty within iSchools, in re-designing the curriculum for well-established graduate programs, such as the Masters in Library & Information Science (MLIS).

References

Arthur, C. (2010). Boris Johnson to launch London 'Data store' with hundreds of sets of data. Guardian Newspaper, 6 January 2010. Retrieved from https://www.theguardian.com/technology/2010/jan/06/london-datastore-launch-johnson-mashups

Bakhshi, H., Mateos-Garcia, J., & Whitby, A. (2014). Model workers: How leading companies are recruiting and managing their data talent. *Nesta*, (July).

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. doi:10.1080/10580530.2012.716740

Kapoor, K., Weerakkody, V., & Sivarajah, U. (2015). Open data platforms and their usability: Proposing a framework for evaluating citizen intentions. *Lecture Notes in Computer Science,* 9373, 261-271. doi:10.1007/978-3-319-25013-7_21

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.L., Brewer, D., ... Van Alstyne, M. (2009). Social science: Computational social science. *Science*, *323*(5915), 721–723. doi:10.1126/science.1167742

Lyon, L. & Brenner, A. (2015). Bridging the data talent gap: Positioning the iSchool as an agent for change. *International Journal of Digital Curation, 10*(1), 111-122. Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/10.1.111/384

Lyon, L., Mattern, E., Acker, A. & Langmead, A. (2015). Applying translational principles to data science curriculum development. iPres Conference Proceedings, Chapel Hill, November 2015.

Lyon, L. & Mattern, E. (2016). Education for real-world data science roles (Part 2): A translational approach to curriculum development. *International Journal of Digital Curation, 11*(2), 13-26. Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/11.2.13

Martin, C. (2014). Barriers to the open government data agenda: Taking a multi-level perspective: Barriers to the open government data agenda. *Policy & Internet, 6*(3), 217-240. doi:10.1002/1944-2866.POI367

Manyika, J., Chui, M., Farrell, D., Kuiken, V.S., Groves, P., & Doshi, E. A. (2013). Open data: Unlocking innovation and preformation with liquid information. Mc Kinsey Report (2013). Retrieved from http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information

National Research Council. (2015). Preparing the workforce for digital curation. Washington, DC: The National Academies Press. doi:10.17226/18590

Pryor, G. & Donnelly, M. (2009). Skilling up to do data: Whose role, whose responsibility, whose career? *International Journal of Digital Curation*, *4*(2), 158–170. doi:10.2218/ijdc.v4i2.105

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., . . . Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PloS One, 6*(6), e21101. doi:10.1371/journal.pone.0021101

Tenopir, C., Dalton, E.D., Allard, S., Frame, M., Pjesivac, I., Birch, B., et al. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS One* 10(8): e0134826. doi:10.1371/journal.pone.0134826

Ubaldi, B. (2013). Open government data: Towards empirical analysis of open government data initiatives. OECD. doi:10.1787/5k46bj4f03s7-en

Weerakkody, V., Irani, Z., Kapoor, K., Sivarajah, U., & Dwivedi, Y. K. (2017). Open data and its usability: An empirical view from the citizen's perspective. *Information Systems Frontiers, 19*(2), 285. doi:10.1007/s10796-016-9679-1