Modelling the Cost of Open Data

Jolon Faichney, Bela Stantic
Yasaman Moaven, Sanjeev Hiremath
School of Information and Communication Technology
Griffith University
Gold Coast, Australia
email: {j.faichney, b.stantic}@griffith.edu.au
email: {yasaman.moaven, sanjeev.hiremath}@griffithuni.edu.au

John Galvin
Organisational Services
City of Gold Coast Council
Gold Coast, Australia
email: jgalvin@goldcoast.qld.gov.au

***Abstract*—The basic principle of *Open Data* is that data should be freely available to the public to use it without restrictions from copyright or other mechanisms of control. Open Data has benefits including improvements in transparency, productivity, integrity, and accountability. However, at what cost do these benefits come? Relatively little work has been done in quantifying the costs of Open Data in comparison to quantifying the benefits. In this paper we provide a case study on the Open Data initiatives within the City of Gold Coast council. We provide a detailed analysis and description of the processes and people involved in opening data sets and provide estimates for the time involved for each participant in the process. We also explore methods to reduce the time and costs involved through the use of automation. By providing cost models for the Open Data process, organisations will be better equipped to formulate and budget for Open Data strategies.**

Keywords-open data; case study; cost modelling.

I. INTRODUCTION

Open Data is a broad term that has been described by the Open Data Institute as "accessible at marginal cost and without discrimination, available in digital and machine-readable format, and provided free of restrictions on use or redistribution" [1]. Even though the term, Open Data, is used to describe all forms of Open Data, it is commonly associated with Government Open Data [2].

Open Data provides both economic and non-economic benefits. By making data openly available to the public, there is more transparency within the government providing the potential for reduced levels of corruption [3]. In 2007, $3.2 billion of misused funds were detected in Canada through the use of Open Data [4].

A report by McKinsey Global Institute [5] found that Open Data can unlock $3 trillion in economic value annually across seven sectors including: education, transportation, consumer products, electricity, oil and gas, health care, and consumer finance. In the United Kingdom, publishing data on cardiac arrests has estimated to have reduced mortality rates, which in turn has an economic value of £400 million per annum, an example of both economic and non-economic benefits [6].

Despite the many benefits of Open Data, the processes involved in making data openly available come at a cost. Given the recentness of Open Data, little work has been done in capturing the cost. However, it is important for organisations to understand the costs involved to make strategic decisions in their Open Data strategies in terms of what data will be made available, how it will be published, and how frequently it will be updated.

In this paper, based on an ongoing collaboration between Griffith University and the City of Gold Coast, we investigate the processes involved in opening data and provide a model for estimating the costs involved.

In Section II, we describe the requirements of Open Data in more detail providing an understanding of the deliverables of an Open Data process. We also explore existing attempts at quantifying the cost of Open Data. In Section III, we specifically focus on the City of Gold Coast's Open Data strategies which we have been working closely with since its inception. In Section IV, we consider the drivers generating demand for Open Data. In Section V, we describe the current process used by the City of Gold Coast to make its data open. In Section VI, we attempt to capture the costs involved in activities, actors, and time in the Open Data process. In Section VII, we look at ways to reduce the cost of the Open Data process through automation. In Section VIII, we discuss the results from our investigation into the cost of Open Data. In Section IX, we provide conclusions and directions for future work as a result of this study.

II. BACKGROUND

In this section, we describe state-of-the-art definitions and standards of Open Data. The requirements of Open Data have an impact on the processes involved in producing it, and hence the cost. The definition of Open Data first begins with the definition of 'Open'.

A. Open Definition

The Open Knowledge Foundation provides the Open Definition, now at version 2, as "*Knowledge is open if anyone is free to access, use, modify, and share it subject, at most, to measures that preserve provenance and openness*" [7]. The Open Definition does not describe how the data is to be made available, but focuses on the policies of the availability of the data. Existing organisations often have a culture where data is not open by default. Therefore, part of the Open Data process is to adopt new policies around openness and educating data custodians to adopt a new culture around Open Data.

B. Sunlight Foundation Open Data Principles

In 2010, the Sunlight Foundation defined 10 principles of Open Data (extending the previous 8 Sebastopol Principles): Completeness, Primacy, Timeliness, Ease of Physical and

 ISBN: 978-1-61208-457-2

Electronic Access, Machine readability, Non-discrimination, Use of commonly Owned Standards, Licensing, Permanence, and Usage costs [8].

Many of the Sunlight Foundation principles are now covered in the Open Definition 2.0, specifically the last five principles listed above. The first five principles however introduce a burden on the data custodians to ensure that the data they provide is in formats that machines can understand. Providing data in raw, primal, machine-readable form may at first appear simple, however rarely do organisations simply export their data in raw format. For example, much data today is stored in relational tables and simply exporting it would introduce problems such as interpreting the internal schema and exposing private fields. In reality database views must be constructed to produce the Open Data. However, if the data is already made available publicly, for example in PDF form, it is possible that the database views used to generate the data in the PDF will already exist and can be used for the export.

C. 5-Star Linked Data

Based on our experience, raw, unprocessed data can make Open Data less accessible [9]. Tim Berners-Lee introduced the 5-star Linked Open Data framework with an emphasis on technical accessibility [10]. Each level makes the data more accessible to applications. The five levels of the Linked Open Data framework are shown below:

1) Make the data available on the web in any format with an open license.
2) Make it available as structured, computer-readable data (not in image or PDF formats).
3) Use non-proprietary formats such as CSV and XML.
4) Use URIs within data so that other websites can point to resources
5) Link data to other data to provide context.

Berners-Lee's focus on linked data is related to his work on the semantic web [11]. The requirement to provide URIs within data which point to other resources and provide context creates another burden for Open Data providers.

D. Open Data Accessibility Framework

Based on their work with the City of Gold Coast, Faichney and Stantic [9] proposed the Open Data Accessibility Framework (ODAF), which can be seen as an expansion of the third level of the 5-star Linked Data. In our experience it is more useful for Open Data consumers to improve the technical accessibility of Open Data than providing linked data. The ODAF is described using the following six criteria:

1) Resource Naming.
2) Data Coalescing.
3) Data Filtering.
4) Data Consistency.
5) Data Formats.
6) API Accessibility.

The above criteria improve usability of the Open Data for Open Data *consumers* but places an extra burden on the Open Data *providers*.

E. ODI Certificates

The Open Data Institute (ODI) has developed the Open Data Certificates [12] which combine the Sunlight Foundation Principles and 5-star Linked Data frameworks into four levels of Open Data access, which are:

Raw – A great start at the basics of publishing open data.
Pilot – Data users receive extra support from, and can provide feedback to the publisher.
Standard – Regularly published open data with robust support that people can rely on.
Expert – An exceptional example of information infrastructure.

The Expert level technical requirements can be summarised as follows:

- Provide database dumps at dated URLs,
- provide a list of the available database dumps in a machine readable feed,
- statistical data must be published in a statistical data format,
- geographical data must be published in a geographical data format,
- URLs as identifiers must be used within data,
- a machine-readable provenance trail must be provided that describes how the data was created and processed.

F. Quantifying the Cost of Open Data

As can be seen in the previous subsections a lot of work has been done in determining the requirements of Open Data, and providing mechanisms to evaluate and rate the quality of Open Data, primarily with the Open Data consumer in mind. However, how much will it cost the Open Data producers to fulfil the preceding requirements?

The Open Data Institute has identified that there are costs associated with technical work, administration and governance, and building skills capacity [13]. However, no attempts were made at quantifying the costs.

The Transit Co-operative Research Program (TCRP) conducted a survey of 60 respondents working with transit data and reported a broad range of hours required to work on Open Data [14]. The survey identified the following types of costs associated with Open Data:

- Staff time to update, fix, and maintain data as needed
- Internal staff time to convert data to an open format
- Staff time needed to validate and monitor the data for accuracy
- Staff time to liaise with data users/developers
- Web service for hosting data
- Publicity/marketing
- Consultant time to convert data to an open format

 ISBN: 978-1-61208-457-2

III. Case Study: City of Gold Coast

In this paper we investigate the costs of opening data through a case study with the City of Gold Coast Council, located within the state of Queensland, Australia. The City of Gold Coast is the second largest council in Australia. In this section we provide an overview of the City of Gold Coast's Open Data strategy.

In 2013, the City of Gold Coast appointed an Enterprise Architect with the purpose of implementing an Open Data strategy. Their commitment to Open Data was also demonstrated by sponsoring the GovHack Gold Coast competition in 2013. GovHack is a national hackathon organised by the federal government. The City of Gold Coast have since sponsored GovHack in 2014 and 2015.

In addition to implementing an Open Data strategy the City of Gold Coast supported and sponsored three apps developed by Griffith University which utilise Open Data: Access GC, GC Dog Parks, and GC Heritage. The three apps all utilise geospatial Open Data integrated with other data sets. Griffith University's work with City of Gold Coast Open Data led to the development of the ODAF presented in the previous section.

In 2015 a new Enterprise Architect for Open Data was appointed initiating increased collaboration with external organisations. For example they are active participants of the ODI Queensland branch and hold regular Open Data Working Groups for the Gold Coast region. The City of Gold Coast's philosophy is *Open by Default*, a concept promoted by ODI. The work in Open Data is broadening to now include Smart Cities, recently signing a Letter of Intent with the Open and Agile Smart Cities initiative.

The City of Gold Coast Open Data is published on the data.gov.au national data portal hosted by the federal government. The City of Gold Coast has published 61 data sets and is ranked 5th in Australia according to the Open Data Census [15].

In the following sections we detail the processes involved to make a data set open and then identify the costs associated with the process.

IV. Sources of Demand

The concept of Demand-Driven Open Data (DDOD) has recently been promoted by the US Department of Health and Human Services (HHS) as the main driver for opening data [16]. The purpose of DDOD is to create value for the 'customer'. The process is managed with *use cases*, which define a clear and concise definition of a desired outcome.

In the City of Gold Coast, three sources of demand for Open Data have been identified, as shown in Figure 1:

1) External Entities
2) Business Users
3) Open Data Team

As in DDOD, external entities may make requests for Open Data. However, so far this has represented only a small portion of requests for Open Data. The majority of requests have come from the Open Data Team themselves. The Open Data Team conducted a survey where participants indicated their interest in data sets listed in the information register of publicly available data sets and ranked the data sets by interest. The information register of publicly available assets existed before the Open Data initiative within the council, however it is worth noting that even though the data was 'publicly' available, it was not necessarily 'open data' in terms of being available electronically and in a machine readable formats. The Open Data Team has been progressively releasing data based on the demand indicated from the survey.

Finally the Business Users, i.e., people within a Business Unit within the organisation, may make a request for Open Data themselves. This may be motivated by a reduction in costs associated with the existing process of other entities requesting data. By making the data open, the costs of managing that process will reduce.

V. High-Level Open Data Process

The process for opening data is outlined in Figure 2. The Open Data process begins with the Business User making a request for a data set to be opened. Note that the Business User's request may have been initiated by one of the three sources of demand in the previous section. It is also important to note that the Business User is the custodian of the data.

A. Request Data Publication

The Business User begins by making a request for a data set to be opened. This may either occur electronically via email to the Open Data Team or may involve interaction with a Business Relationship Officer. It is important to capture the possible interaction of the Business Relationship Officer in the request process in terms of determining the cost of Open Data.

The Open Data Team receives a request for Open Data which includes details on the types and forms of data required. After reviewing the data set the Open Data Team may elect to agree to publish the data.

The remaining flow is determined by how the data is stored:

1) Relational Database
2) Geospatial Information System
3) Spreadsheet or other file format

B. Opening Relational Data

Data in relational databases is relatively easy to publish once a database view has been created. Creating the database view however may be challenging as it can involve complex SQL queries that may cross multiple tables and databases. In the City of Gold Coast a large portion of the business data is stored within SAP databases.

C. Opening GIS Data

In the City of Gold Coast there currently isn't a mechanism to create a 'view' of GIS data in the same way as relational databases. As a result the data must be exported from the GIS system as a file in a common GIS format such as KML, SHP, or GeoJSON.

D. Opening File Data

Other data may be stored in individual files, such as spreadsheets. These files can be published as is. However, if they need to be modified this will be a resource intensive phase, generally more so than the publication of database or GIS data.

 ISBN: 978-1-61208-457-2

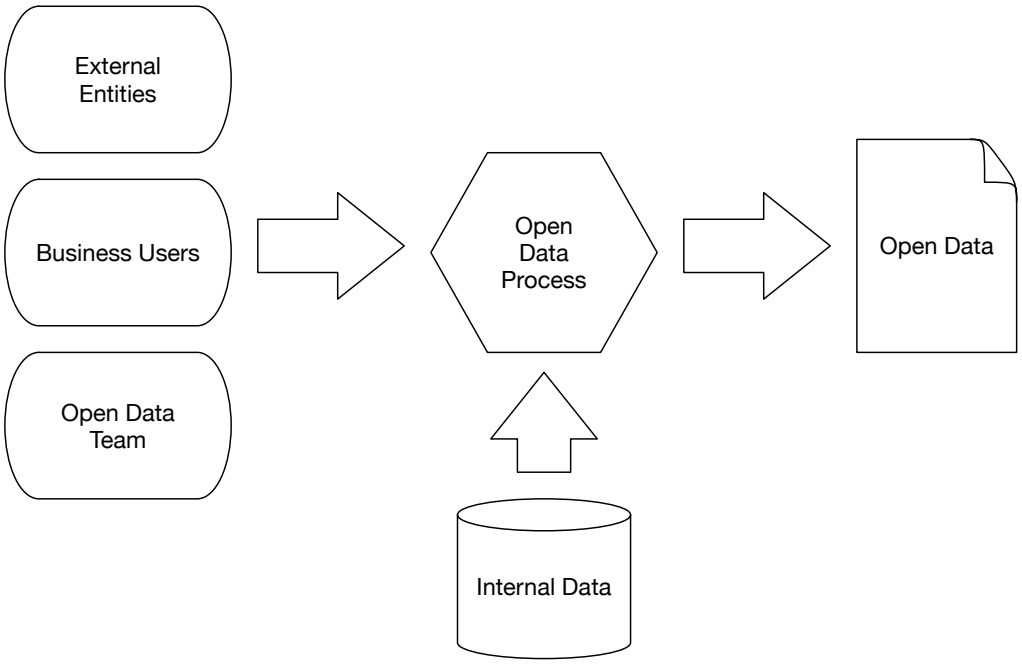

Figure 1. Sources of demand in the Open Data process.

E. Privacy Concerns

Data may need to be de-identified to ensure privacy policies are not breeched. This may involve the removal columns or tags from data sets, or only releasing aggregate data views.

F. Test Data Review

Before publication the Business User is sent a sample extract of data to be published. The Business User confirms whether the extracted data is correct.

G. Automation

If data is to be released periodically, an automation process can be established. Currently in the City of Gold Coast data publication is only automated if it is updated more frequently than yearly. Database views are relatively simple to automate. GIS and file-based data currently still involves human intervention. Automation is discussed in more detail in Section VII.

H. Approve for Publishing

Once the automation process is implemented, the Business User may approve the data for publishing. Data is currently published to the national Open Data portal data.gov.au. It is made available on data.gov.au initially with private access to ensure the processes are working correctly. Once the Business User approves the publication of data, the data set is made public by the Open Data Team.

VI. Cost of Open Data

In this section we aim to model the cost of the Open Data process. The main cost in the Open Data process is staff time. Since the cost of staff time varies between organisations, cities, and countries, we will model our costs as a proportion of a staff member's time. Table I shows our estimate for the number of full-time equivalent (FTE) days spent for a single data set. Note that the total of 6.5-16 days is not the wall clock time required to release a data set as some work can be performed in parallel and multiple data sets can be released simultaneously if multiple staff members exist on the team, likewise the wall clock time may be longer if there are delays in the process such as organising meetings at a future date.

The City of Gold Coast has one member in the Open Data Team being the Enterprise Architect for Open Data. The Enterprise Architect has other responsibilities, such as developing strategies and policies for Open Data and engaging with the community. This limits the amount of time dedicated to releasing new data sets. Approximately 40% of the Enterprise Architect's time is dedicated to releasing data sets. The above table indicates that 1-3 days are required to release a data set, which approximately correlates with the current output of around 40 data sets a year by the City of Gold Coast.

The ranges provided indicate variations in complexity. Data sets which involve more files will take longer to publish. Each data set published to the data portal requires the data set to be registered and a unique key provided which is used for subsequent updates to the file. The complexity will also depend on the data. Database views are generally the most complex to formulate. GIS data exports are often simpler as the required

TABLE I. Days required to release a single data set.

Actor	FTE Days/Data Set
Open Data Team	1-3 days
Business Users	2-5 days
Business Relationship Officer	0.5-1 day
Analyst	2-5 days
Integration Analyst	1-2 days
Total	**6.5-16 days**

 ISBN: 978-1-61208-457-2

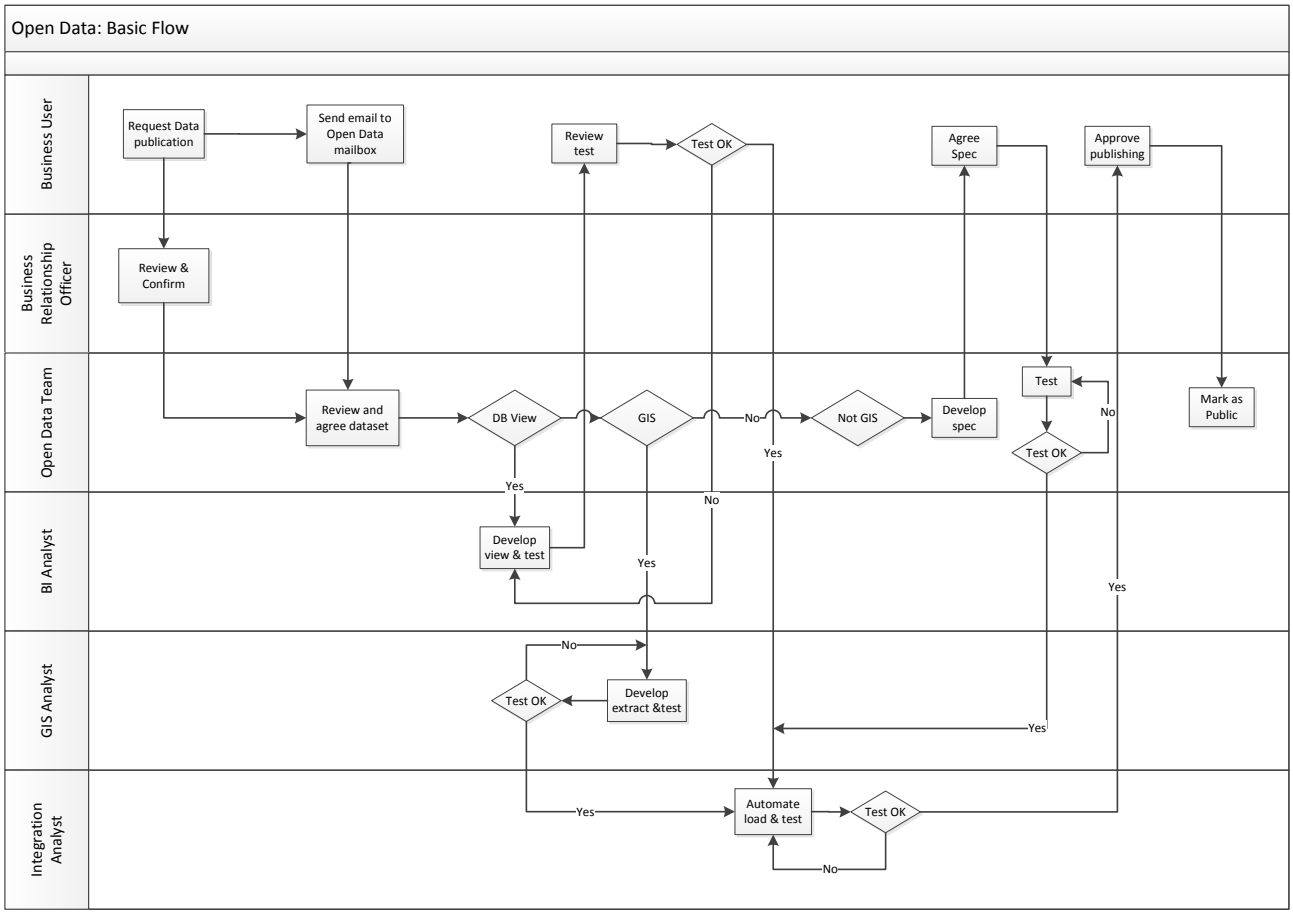

Figure 2. Process used by the City of Gold Coast to open existing data sets.

data exists in a layer and can be exported in its entirety.

VII. AUTOMATION

Automation can be utilised to reduce the cost of releasing periodic data and can also reduce human errors that can be introduced when repeatedly releasing data. The City of Gold Coast has two primary mechanisms for automating the release of data:

1) Database Views
2) Automatic File Upload

Both approaches upload data to the data portal at regular intervals. The data portal utilises the CKAN content management system. The first approach is completely automated. Database views are generated and the results uploaded to the data portal via a script. GIS and other file data is currently produced manually resulting in a file to be uploaded to the data portal. To simplify this process, a network directory is monitored, when a file is copied to the network directory it is automatically uploaded to the respective section of the data portal. This reduces the time required by the staff that administer the data portal.

Automation is able to reduce the cost of releasing data on an ongoing basis. However it will require further time upfront to establish the automation process. This additional time requires the Integration Analyst to implement and test the automation procedure and the Business User to confirm that it is working.

Some of the automation procedures can be reduced across data sets. The City of Gold Coast spent 20 days building their current automation system.

VIII. DISCUSSION

As can be seen in the previous sections, Open Data has a cost. Do the benefits of releasing Open Data outweigh the costs? The literature so far indicate that the benefits of Open Data outweigh the costs, this conclusion has been determined by estimating the overwhelming benefits without providing finer grained analysis of the processes and costs involved in releasing Open Data. In this paper, we have looked at staff time which correlates with a financial cost and can be evaluated against a financial benefit. However there are other non-financial benefits to releasing Open Data such as transparency and social benefit. Can we evaluate the non-financial benefits against the financial costs? We don't think it is necessary to draw a connection between the financial cost of Open Data and the non-financial benefits. Modelling the cost of Open Data is

 ISBN: 978-1-61208-457-2

sufficiently important for organisations to help plan their Open Data strategies.

IX. Conclusions and Future Work

In this paper we have reported our collaboration with the City of Gold Coast in capturing the costs involved with releasing Open Data. Some studies have reported the macro-economic benefits of Open Data, but relatively little has been done in capturing the cost. In this paper we report the processes currently used by the City of Gold Coast and estimate the roles involved in opening data and the FTE time required by staff. This will help organisations budget and plan for Open Data rollouts and transitions.

Further work will investigate in greater detail the causes of variations in complexity in releasing Open Data, such as the type of data (database, GIS, file), the number of files within the data set, additional data processing, and the time required to deal with cultural resistance to releasing data openly.

Additional work will also investigate how various automation techniques can be used to further reduce the cost of Open Data, particularly in the cases of GIS and spreadsheet-based data.

Acknowledgment

The authors would like to thank the Open Data Working Group of the City of Gold Coast Council for their collaboration and support during this project.

References

[1] M. Heimstädt, F. Saunderson, and T. Heath, "From toddler to teen: Growth of an open data ecosystem." eJournal of eDemocracy & Open Government, vol. 6, no. 2, 2014, pp. 123–135.

[2] J. Kloiber, "Open government data - between political transparency and economic development," Master's thesis, Utrecht University, 2012.

[3] N. Rajshree and B. Srivastava, "Open government data for tackling corruption-a perspective," in Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012, pp. 21–24.

[4] D. Eaves, "Case study: How open data saved Canada $3.2 billion," 2012, retrieved: January, 2016. [Online]. Available: http://eaves.ca/2010/04/14/case-study-open-data-and-the-public-purse/

[5] J. Manyika et al., "Open data: Unlocking innovation and performance with liquid information," McKinsey Global Institute, Tech. Rep., 2013.

[6] Deloitte, "Market assessment of public sector information," UK Department for Business Innovation and Skills, Tech. Rep., 2013.

[7] Open Knowledge Foundation, "Open definition 2.0," 2014, retrieved: January, 2016. [Online]. Available: http://opendefinition.org/od

[8] Sunlight Foundation, "Ten principles for opening up government," 2010, retrieved: January, 2016. [Online]. Available: http://sunlightfoundation.com/policy/documents/ten-open-data-principles/

[9] J. Faichney and B. Stantic, "A novel framework to describe technical accessibility of open data," in The First International Conference on Big Data, Small Data, Linked Data and Open Data (ALLDATA), 2015, pp. 52–57.

[10] T. Berners-Lee, "Is your linked open data 5 star?" 2010, retrieved January, 2016. [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html

[11] T. H. C. Bizer and T. Berners-Lee, "Linked data – the story so far," International Journal on Semantic Web and Information Systems, 2009, pp. 1–22.

[12] The Open Data Institute, "Open data certificates," 2013, retrieved: January, 2016. [Online]. Available: https://certificates.theodi.org

[13] ——, "Estimating the cost of a government open data initiative," 2014, retrieved: January, 2016. [Online]. Available: https://theodi.org/blog/estimating-the-cost-of-a-government-open-data-initiative

[14] C. L. Schweiger, "Open data: Challenges and opportunities for transit agencies, a synthesis of transit practice," Tech. Rep., 2015.

[15] O. K. Foundation, "Local open data census for Australia," 2015, retrieved: January, 2016. [Online]. Available: http://au-city.census.okfn.org

[16] D. Portnoy, "Identifying and harnessing demand to drive open data," 2015, retrieved: January, 2016. [Online]. Available: http://www.hhs.gov/idealab/2015/03/16/identifying-harnessing-demand-drive-open-data/

 ISBN: 978-1-61208-457-2