Strategies and Approaches for Exploiting the Value of Open Data

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

von
Judie Attard
aus
Victoria (Gozo),
Malta

Bonn, 17.10.2016

Dieser Forschungsbericht wurde als Dissertation von der Mathematisch-Naturwissenschaftlichen Fakultät der Universität Bonn angenommen und ist auf dem Hochschulschriftenserver der ULB Bonn http://hss.ulb.uni-bonn.de/diss_online elektronisch publiziert.
Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter:	Prof. Dr. Sören Auer
2. Gutachter:	Prof. Dr. Marijn Janssen
Tag der Promotion:	21.03.2017
Erscheinungsjahr:	2017

Abstract

Data is increasingly permeating into all dimensions of our society and has become an indispensable commodity that serves as a basis for many products and services. Traditional sectors, such as health, transport, retail, are all benefiting from digital developments. In recent years, governments have also started to participate in the open data venture, usually with the motivation of increasing transparency. In fact, governments are one of the largest producers and collectors of data in many different domains. As the increasing amount of open data and open government data initiatives show, it is becoming more and more vital to identify the means and methods how to exploit the value of this data that ultimately affects various dimensions.

In this thesis we therefore focus on researching how open data can be exploited to its highest value potential, and how we can enable stakeholders to create value upon data accordingly. Albeit the radical advances in technology enabling data and knowledge sharing, and the lowering of barriers to information access, raw data was given only recently the attention and relevance it merits. Moreover, even though the publishing of data is increasing at an enormously fast rate, there are many challenges that hinder its exploitation and consumption. Technical issues hinder the re-use of data, whilst policy, economic, organisational and cultural issues hinder entities from participating or collaborating in open data initiatives.

Our focus is thus to contribute to the topic by researching current approaches towards the use of open data. We explore methods for creating value upon open (government) data, and identify the strengths and weaknesses that subsequently influence the success of an open data initiative. This research then acts as a baseline for the value creation guidelines, methodologies, and approaches that we propose. Our contribution is based on the premise that if stakeholders are provided with adequate means and models to follow, then they will be encouraged to create value and exploit data products. Our subsequent contribution in this thesis therefore enables stakeholders to easily access and consume open data, as the first step towards creating value. Thereafter we proceed to identify and model the various value creation processes through the definition of a Data Value Network, and also provide a concrete implementation that allows stakeholders to create value. Ultimately, by creating value on data products, stakeholders participate in the global data economy and impact not only the economic dimension, but also other dimensions including technical, societal and political.

Acknowledgements

Persuing a PhD was an eye-opening and rewarding experience. Germany proved to be quite challenging to settle into, yet it turned out to be a beautiful country that has much more to offer than beer and currywurst! Living by the Rhine provided a beautiful environment for relaxing (or otherwise going for 50km bike rides).

I would first like to thank Prof. Dr. Sören Auer, without whom this experience would not have been possible. Working within the Enterprise Information Systems group was an unforgettable experience. I would like to extend my gratitude towards Dr. Fabrizio Orlandi, who was a great co-supervisor and provided me with invaluable discussions and feedback throughout my PhD.

A special thanks goes to my friends in Bonn, particularly Simon, Fabrizio, Steffen, and Nicole. We passed many enjoyable moments together, and you made my stay in Germany so much more fun!

I would like to thank my parents, Joseph and Frances, and my brother Michael-George. Even though they made it very clear they want me back home in Malta, they supported me throughout my studies abroad, and I am very sure that when the time comes to return to Malta, they will be waiting with open arms.

Lastly, I want to express my deepest gratitude to my fiancé Jeremy, who has been an amazing partner for these five years we spent together, and throughout our shared academic journey away from home. I cannot wait to embark on the next great adventure together.

Prost!

Contents

List of Figures

List of Tables

Part I

Prologue

CHAPTER 1

Introduction

In our information-centric society, data has become an indispensable commodity that serves as a basis for many products and services. Huge amounts of data are constantly being created, such as what consumers are buying, flight travel plans, financial transactions, energy consumption, health records, etc. This flow of data is therefore becoming a more crucial part of the global economy, and many traditional sectors, such as health, transport, or retail, are benefiting from new-found opportunities based on digital developments.

Complementing the vast increase in the production of data, the relatively new trend of open data is becoming more and more popular. The goals of this open data movement are to make data publicly available for re-use, and is usually motivated by societal goals, such as improving the transparency and accountability of institutions, reducing poverty, and increasing innovation. This movement has prompted the foundation of a number of open data initiatives, such as the Open Data Institute[1] and the Open Knowledge Foundation[2]. Such initiatives advocate and campaign for the release of data to the public, and many times result in starting a chain of changes which ultimately have a real impact on open data, such as the establishment of new policies and laws.

Although still in its early days, the open data movement has resulted in a large number of open datasets in a plethora of different domains. This data is used to create products and services that have a number of different impacts and benefits, such as government data portals[3], reviews, feedback, and product suggestion on e-commerce websites, weather emergencies forecast[4], patient monitoring[5], citizen participation and decision-making[6], etc. Open datasets can be created by different stakeholders, such as institutions, companies, and individuals, but governments or public entities are usually the largest producers of data. Yet, whether the data is geospatial, environmental, weather, transport and planning, statistical, budget, or otherwise, it has social and commercial value. In fact Carrara et al. [23] have estimated the (total) market size of open data in the European Union to be between 193 and 209 billion Euro for 2016. Manyika et al. [83] also estimate that open data can help unlock between 3 to 5 trillion U.S. Dollars in economic value, annually. The benefits of this economic value include increased efficiency, development of new products and services, cost savings, and better quality products. For example, Mastodon C (a big data company) used open data to identify unnecessary spending in prescription

[1] http://theodi.org/ (Accessed: 30 August 2016)
[2] https://okfn.org/ (Accessed: 30 August 2016)
[3] https://open-data.europa.eu/en/data/ (Date accessed: 2 August 2016)
[4] http://centrodeoperacoes.rio/ (Date accessed: 2 August 2016)
[5] http://www.immunizeindia.org/ (Date accessed: 2 August 2016)
[6] https://www.fixmystreet.com/ (Date accessed: 2 August 2016)

medicine[7]. This will result in potentially huge savings from the National Health Service in the UK. Although not yet quantified, open data also results in a number of social benefits, such as improvement of teaching approaches, more efficient public transport, increasing competitivity between businesses, better healthcare provision, increase in citizen social control, and hindering corruption.

In recent years, in order to reflect this dataification [29], the concept of *data value chains* was introduced, building upon the concept of traditional value chains for tangible products [111]. The rationale of a data value chain is to extract the highest possible value from data by modifying, processing and re-using it. Value creation is especially relevant since open data has no value within itself unless it is used [59]. Value can be added to the generated raw data to make it re-usable or more fit for the intended use. This results in the data being a product within itself. The exploitation of this data with added value has the potential to feed a chain of innovative information products and services, making the data value chain the centre of the knowledge economy.

1.1 Problem Definition and Motivation

Open data already provides advantages to entities who embrace its potential. Open data can improve service provision, such as patient monitoring[8], it increases competitiveness [59], it can be used to help preserve and showcase cultural identity, such as in the case of the German Digital library[9], and in a government setting open data also helps hinder corruption and increase citizen social control [78]. Yet, although the use of open data has seen a drastic increase in recent years, there are still some major challenges which hinder the full potential of open data. Various dimensions of these barriers are covered in existing research, however we can aggregate them in five dimensions as follows:

1. **Technical** - This dimension regards aspects concerning the nature of the data itself. For example, the use of PDF to publish data or data of low quality (e.g. incomplete or ambiguous) would act as a disincentive for re-use.

2. **Policy and Legal** - Existing laws or policies impact the resulting creation or use of open data. For example, some licenses restrict data use, and the incompatibility of licences between datasets further aggravates the issue.

3. **Economic and Financial** - Monetary issues mostly affect the creation and publishing of open data. Being a relatively new concept, there might not be enough budget allocated to such efforts.

4. **Organisational** - This dimension is especially relevant within existing institutions who want to start an open data endeavour. In this case, data can be created in various parts of the institution, hence the challenge here is to implement an appropriate strategy for its aggregation and management.

5. **Cultural** - Some preconceptions about open data still exist in the general public. For example, some stakeholders might not understand the potential of open data, so they are not motivated enough to publish it. Other concerns arise in the business sector, where the publishing of open data might be considered to be unfair competition from rival companies.

Whilst all of the above-mentioned challenges can have a big impact on the success of an open data initiatives [27, 159], in this thesis we focus on the technical dimension. Since the data itself forms the

[7] http://theodi.org/news/prescription-savings-worth-millions-identified-odi-incubated-company (Date accessed: 2 August 2016)

[8] http://www.immunizeindia.org/ (Date accessed: 2 August 2016)

[9] https://www.deutsche-digitale-bibliothek.de/?lang=en (Date accessed: 2 August 2016)

basis of any open data initiative, this dimension is vital as it not only affects the other challenges, but also affects the potential value that can ultimately be exploited from open data, and therefore the resulting benefits. Whilst open data can be taken to be any data that is publicly accessible, the process of opening data with the aim of enabling the exploitation of its value requires various non-trivial steps. These include the selection of the data to be published (including the removal of sensitive or private data), the curation of this data to make it more fit for the intended use, and the actual publishing of the data which makes it available for discovery and consumption by the public. Data can also be consumed in a myriad of ways, such as in decision-making, visualisations, and service creation, and each method provides the end user with different value, whilst also presenting the stakeholder with various challenges. This means that even though open data use is on the rise, we have no guarantee that the potential value behind its re-use is actually being fully exploited.

In this thesis our motivation is therefore to encourage and enable the use and exploitation of open data to its full potential. This can have substantial impacts (direct and indirect) on various dimensions, including economic, societal, and political. For example, good quality data would result in more re-use and consequently in more benefits. This will eventually act as a motivation for other stakeholders to participate in an open data initiative. We therefore assess the current situation of open data, including the actual processes for publishing and consuming data, challenges in exploiting open data, impacts of open data initiatives, guidelines on how to fully exploit open data, roles of participating stakeholders, and extracting value from data. While these various topics were previously tackled in existing literature, we comprehensively cover all these aspects in context of value creation on open data. The main aim of this research is hence to identify the key factors that ultimately influence the outcome of an open data initiative, that is, any effort towards opening data for public use.

1.2 Research Questions

Whilst we tackle various topics in context of open data, in this thesis we have a single motivation, as discussed in Section 1.1. Hence, the core research question we answer in this thesis can be defined as follows:

> *What strategies, methods and technologies can be used to maximise the exploitation of open data?*

This question guides our research towards exploiting open data to its highest value potential, where the value potential refers to the possible outcomes and impacts of using open data. This means that while certain data can provide more benefits or impacts than other data, our aim is to enable the maximal exploitation of the data in question. Considering the somewhat generic nature of this research question, we further define more specific research questions in order to better direct our research. Each of these sub-questions is then reflected in the contributions in the rest of this thesis.

Our first aim within this thesis is to obtain a clear picture of the current situation of open data initiatives. We here focus on the tools and approaches for publishing and consuming open data since these processes are vital in any open data initiative: without the existence or creation of open data, there is no open data initiative. Our first research sub-question is hence the following:

Research Question 1:

What are existing approaches and techniques that enable the publishing and consumption of open data?

The aim of this research question is to identify the various factors that make up an open data initiative. Since the scope of open data is vast, for this research question we decided to narrow this exploratory search to open government data. As a subset of open data, open government data is still representative of its broader counterpart. Being a popular and common use case of open data, open government initiatives can provide us with crucial information on the open data life cycle. This information will hence provide the building blocks to define an open government data life cycle, and specify guidelines and recommendations on what are the best methods to publish and consume open (government) data, as well as how to get the most out of it.

The main challenge in exploiting open data and releasing value is that open data has no value in itself, yet it becomes valuable when it is used [59]. Although there is ample research, guidelines, and tools on the publication of open data, the research on the consumption of open data is quite lacking, especially in the case of non-experts. For this reason in the next research question we target the consumption of open data:

Research Question 2:

How can we enhance the consumption process of a data product in order to enable further value creation?

We here consider a *data product* to be any data that through its use will facilitate the end goal. For example, public transport timetable data can be used to create a journey planning mobile application that has the purpose of enabling a person to arrive to the desired destination at the desired time. Using the above question allowed us to investigate current methods for open data consumption, with the aim of identifying any strengths and weaknesses. We also explore the role of Linked Data technologies in the consumption process, based on the hypothesis that such technologies can improve the value creation potential of the data in question. This research question hence guides us towards enabling and encouraging stakeholders in exploiting open data by creating value.

Once stakeholders are able to consume open data, endless possibilities are available. Our next research question therefore has the aim of identifying and defining processes that are used by different stakeholders to create value upon a data product. In context of more traditional value creation on tangible products, the processes by which an entity adds value to a product, including its production, marketing, distribution, and after-sales service, form what is called a value chain. In the context of data, we here focus on defining a data value chain that is capable of representing existing sequences of value creation on data products:

Research Question 3:

What aspects and processes play a role in value creation on a data product?

To answer this research question we require to explore existing value chains, with the aim of identifying the best value chain specifically suitable for a data product. Such a value chain is particularly effective in enabling the full exploitation of open data since it provides insight on the specific processes that make data more useful. The delineation of these processes and the related stakeholder roles within the value chain will then act as a guide for participating stakeholders, and they can align their contribution within the value chain accordingly. After identifying the various processes for value creation, we require to concretely determine how value is created upon a data product. Again, due to the broad nature of open data, for this research question we take open government data as a use case. The delineation of the resulting impacts of the value creation process will then provide us with a better perspective on why

the value creation process is vital in our data-based economy. Finally, we aim to explore methods that allow us to measure the value potential of a data product, hence enabling entities to exploit data with the highest potential.

Research Question	Part	Contribution
1	II	A systematic analysis of existing open government data initiatives
2	III	A data consumption and re-use framework based on Linked Open Data
3	IV	The identification and specification of a Data Value Network as a methodology to create value on a data product
3	IV	A Demand and Supply Service that enables stakeholders to participate in the Data Value Network
3	IV	A Value Creation Assessment Framework to analyse the value creating potential of open data initiatives

Table 1.1: Overview of the contributions and research questions we tackle in the different parts of the thesis.

1.3 Research Map

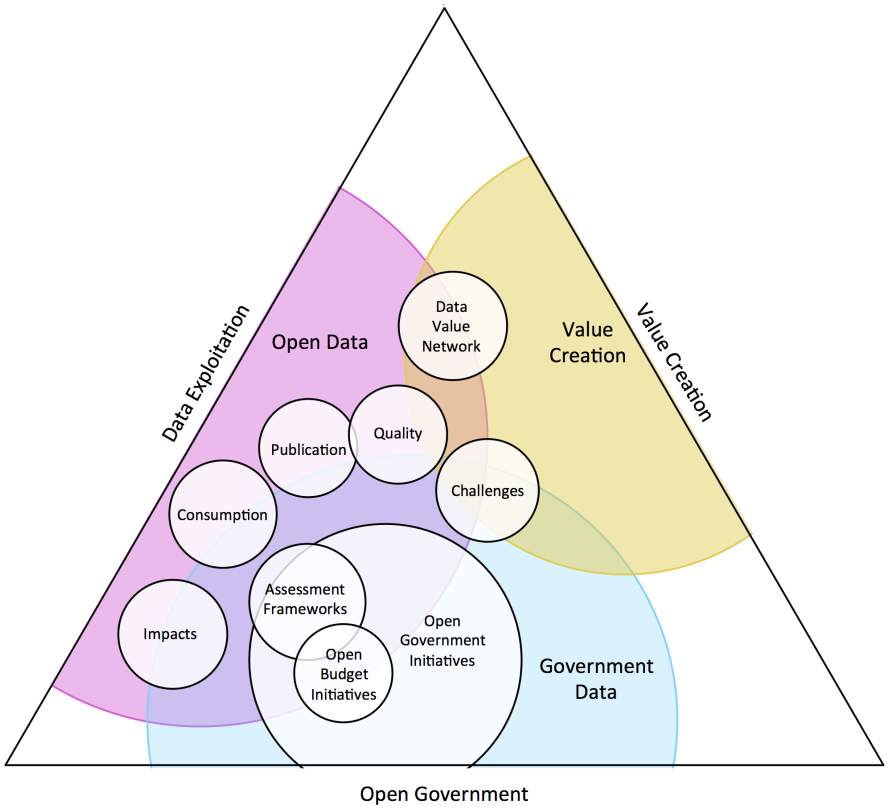

Figure 1.1: Research areas (three sides of the triangle), topics (circles in colour), and aspects (circles in white) indicating the various directions we focus on within this thesis.

In this section we provide a description of the main contributions provided in this thesis, as the results of the research questions defined in Section 1.2. Figure 1.1 provides an overview of the research areas we explore, namely Data Exploitation, Value Creation, and Open Government, and how the pertaining topics within these areas; Open Data, Government Data, and Value Creation, intersect. We also portray the different aspects of the topics we focus on. Whilst not accurately representing the degree of intersection

between the research areas, topics, and aspects we explore in this thesis, this diagram gives a good overview of our research focus.

Contributions:

1. *A systematic analysis of existing open government data initiatives*
   This contribution was determinative towards creating a baseline for the rest of this thesis. Following the direction provided by Research Question 1, we systematically analyse a number of existing open government data initiatives, whilst also identifying any flaws or strengths within the implemented approaches and techniques. We consequently formulate and define an Open Government Data Life Cycle. Focusing on the consumption and publication of open data, the two most essential processes within the life cycle, we identify challenges and issues which hinder the success of an open government data initiative. Apart from providing solutions or ways to mitigate these challenges or issues, we also identify the various impacts (direct and indirect) that open data can have on its stakeholders. This contribution hence acts as the first step towards motivating and encouraging the use of open data. This contribution is published in the following publications: [9, 143].

2. *A data consumption and re-use framework based on Linked Open Data*
   This framework was developed in order to fill in the niche in existing tools that aid and enable stakeholders to consume open data, particularly non-experts who are not familiar with the SPARQL querying language, RDF, or the underlying schema of the open dataset. The aim of this framework is therefore to enhance the consumption process of a data product, as defined in Research Question 2. The framework basically allows users to query datasets through a user-friendly SPARQL endpoint, and download the results in a number of formats. We define an ontology which is used to persist the information generated in this process, in order to maintain provenance information. This allows users to re-use and edit existing queries. This contribution is a concrete step towards enabling stakeholders to more easily consume open data. This contribution is published in the following publications: [5, 10, 97].

3. *The identification and specification of a Data Value Network as a methodology to create value on a data product*
   In order to characterise the various value creation processes upon a data product, we define a Data Value Network. In response to Research Question 3, we here highlight the processes that improve upon a data product with the aim of making it more useful, and we also identify how these processes concretely achieve this improvement. We distinguish the various roles though which stakeholders can participate, as well as the different entities that usually partake in open data initiatives. We also point out the dimensions that affect an open data initiative, as well as the resulting impacts. This contribution is published in the following publications: [6–8].

4. *A Value Creation Assessment Framework to analyse the value creating potential of open data initiatives*
   This contribution focuses on the impact of value creation, where we provide a number of aspects of open (government) data initiatives that we recommend should be assessed in order to determine the potential of value creation. This assessment framework is intended to act as a baseline for identifying or establishing initiatives with the highest probability of being successful, as well as resulting in the highest impact through the use of open data. This contribution is published in the following publications: [6, 8].

5. *A Demand and Supply Service that enables stakeholders to participate in the Data Value Network* As a concrete implementation of the Demand and Supply Model that we define based on the Data Value Network, we developed a Demand and Supply as a Service. This service acts as a broker between data publishers and consumers, allowing the former to 'advertise' their data products, and the latter to more easily discover the data they require. We define and implement an ontology to persist information about existing datasets and their use cases, as well as requested datasets. Thus this service provides an entry point to participate in value creation on a data product. This contribution is published in the following publications: [7].

1.4 Publications

The work described in this thesis was partially covered by or stemmed from the following publications:

1. **Judie Attard**, Fabrizio Orlandi, Simon Scerri, Sören Auer. *A systematic review of open government data initiatives*. In Proceedings of the Government Information Quarterly Journal, 2015.

2. **Judie Attard**, Fabrizio Orlandi, Sören Auer. *ExConQuer Framework - Softening RDF Data to Enhance Linked Data Reuse*. In Proceedings of the ISWC 2015 Posters & Demonstrations Track co-located with the 14th International Semantic Web Conference (ISWC-2015), Bethlehem, Pennsylvania, USA, October 11, 2015.

3. **Judie Attard**, Fabrizio Orlandi, Sören Auer. *Value Creation on Open Government Data*. In Proceedings of the 49th Hawaii International Conference on System Sciences, HICSS 2016, Koloa, Hawaii, USA, January 5-8, 2016.

4. Alan Freihof Tygel, **Judie Attard**, Fabrizio Orlandi, Maria Luiza Machado Campos, Sören Auer. *"How Much?" is not Enough: an Analysis of Open Budget Initiatives*. In Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance, ICEGOV 2016, Montevideo, Uruguay, March 1-3, 2016.

5. **Judie Attard**, Fabrizio Orlandi, Sören Auer. *Data Driven Governments: Creating Value through Open Government Data*. In Proceedings of the Transactions on Large-Scale Data- and Knowledge-Centered Systems Journal, 2016.

6. Spiros Mouzakitis, Dimitris Papaspyros, Michael Petychakis, Sotiris Koussouris, Anastasios Zafeiropoulos, Eleni Fotopoulou, Lena Farid, Fabrizio Orlandi, **Judie Attard**, John Psarras. *Challenges and Opportunities in renovating Public Sector Information by enabling Linked Data and Analytics*. In Proceedings of the Information Systems Frontiers Journal, 2016.

7. **Judie Attard**, Fabrizio Orlandi, Sören Auer. *Data Value Networks: Enabling a New Data Ecosystem*. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Omaha, Nebraska, USA, October 13-16, 2016.

8. **Judie Attard**, Fabrizio Orlandi, Sören Auer. *ExConQuer: Lowering barriers to RDF and Linked Data re-use*. To appear in Proceedings of the Semantic Web Journal, accepted on 12 October 2016.

9. **Judie Attard**, Fabrizio Orlandi, Sören Auer. *Exploiting the Value of Data through Data Value Networks*. In Proceedings of the 10th International Conference on Theory and Practice of Electronic Governance, ICEGOV, 2017.

1.5 Document Structure

This thesis is divided in five parts. After providing the context and motivation of this thesis in Part I, we proceed to discuss open data initiatives whilst taking open governments as a context. In Part II we hence lead out a systematic survey on literature covering open government initiatives with the aim of identifying key aspects that determine their success or otherwise. We specifically focus on the publishing and consumption of data, two vital processes in such initiatives. Directing our efforts towards the consumption of open data by stakeholders, in Part III we explore existing approaches that aid and enable entities in consuming open data. As a crucial process in value creation, in this part we strive to enable stakeholders, particularly non-experts, to easily consume open data. In the next part, Part IV, we focus on the various value creating processes that stakeholders can participate in to enhance data products. This value creation helps stakeholders to exploit data to its fullest potential. Finally, in Part V, we provide the concluding discussion for the research in this thesis. We provide an overview of the relevant chapters at the beginning of each part.

Part II

Open Data in the Government Domain

In this part of the thesis we lead out a systematic survey with the aim of creating a baseline for the rest of the thesis. In the following chapters we provide an insight into open government data initiatives, with the aim of identifying key aspects that determine their success. We specifically focus on the government domain, rather than open data in general, since it is a relevant subset of open data and is a popular and common use case of open data. In Chapter 2 we provide the relevant context, the implemented research method, and the definition of the open government data life cycle. In Chapter 3 we systematically cover existing literature on open government initiatives, whilst also identifying existing challenges and impacts. In Chapter 4 we focus specifically on the publishing and consuming aspects of the open government data life cycle. We finally provide a use case and an assessment model of open budget initiatives in Chapter 5.

The chapters in Part II are based on the following publications:

- **Judie Attard**, Fabrizio Orlandi, Simon Scerri, Sören Auer. *A systematic review of open government data initiatives*. In Proceedings of the Government Information Quarterly Journal, 2015.
- Alan Freihof Tygel, **Judie Attard**, Fabrizio Orlandi, Maria Luiza Machado Campos, Sören Auer. *"How Much?" is not Enough: an Analysis of Open Budget Initiatives*. In Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance, ICEGOV 2016, Montevideo, Uruguay, March 1-3, 2016.

CHAPTER 2

Context of Systematic Survey

In recent years, a number of open data movements sprung up around the world, with transparency and data re-use as two of the major aims. To mention a few, there is the Public Sector Information (PSI) Directive[1] in 2003 in Europe, U.S. President's Obama open data initiative in 2009[2], the Open Government Partnership[3] in 2011, and the G8 Open Data Charter[4] in 2013. Open government data portals resulting from such movements, such as in the United Kingdom[5], the United States of America[6], and Singapore[7], provide means for citizens and stakeholders to obtain government data concerning the locality or country in question.

While not being the only motivation, initially corruption was one of the main issues that prompted the founding of open government data initiatives such as the above. Corruption is a global issue that seriously harms the economy and society as a whole, affecting people's lives and often infringing fundamental human rights. The democracy of many countries around the world is undermined by deep-rooted corruption, which also affects the economic development. While the total economic costs of corruption cannot be easily calculated, the 2014 European Commission Anti-Corruption Report[8] states that corruption can be estimated to cost the European Union economy 120 billion Euros per year. In places where there is widespread belief that corruption prevails, the people end up losing faith and trust in those entrusted with power. As the Global Corruption Barometer 2013[9] shows, corruption can be identified running through the democratic and legal process in many countries. This results in people losing trust in key institutions such as political parties, the judiciary and the police. While transparency cannot be regarded as an end [165], it can be regarded as a means to act as a disincentive to corruption.

Collectively, there are three main reasons for opening government data[10]:

[1] http://ec.europa.eu/digital-agenda/en/european-legislation-reuse-public-sector-information (Date accessed: 2 August 2016)

[2] http://www.whitehouse.gov/open/documents/open-government-directive (Date accessed: 2 August 2016)

[3] http://www.opengovpartnership.org/ (Date accessed: 2 August 2016)

[4] https://www.gov.uk/government/publications/open-data-charter (Date accessed: 2 August 2016)

[5] http://www.data.gov.uk (Date accessed: 14 August 2016)

[6] http://data.gov (Date accessed: 14 August 2016)

[7] http://data.gov.sg (Date accessed: 14 August 2016)

[8] http://ec.europa.eu/dgs/home-affairs/what-we-do/policies/organized-crime-and-human-trafficking/corruption/anti-corruption-report/index_en.htm (Date accessed: 2 August 2016)

[9] http://www.transparency.org/gcb2013 (Date accessed: 2 August 2016)

[10] https://okfn.org/opendata/ (Date accessed: 2 August 2016)

1. **Transparency** - In order to have a well-functioning, democratic society, citizens and other stakeholders need to be able to monitor government initiatives and their legitimacy. Transparency also means that stakeholders should not only be able to access the data, but they should also be enabled to use, re-use and distribute it. The success to achieve transparency results in a considerable increase in *citizen social control*;

2. **Releasing social and commercial value** - Governments are one of the largest producers and collectors of data in many different domains [2, 58]. All data, whether addresses of schools, geospatial data, environmental data, transport and planning data, or budget data, has social and commercial value, and can be used for a number of different purposes which can be different than the ones originally envisaged. By publishing such data the government encourages stakeholders to innovate upon it, and create new services;

3. **Participatory Governance** - Through the publishing of government data citizens are given the opportunity to actively participate in governance processes, such as decision-taking and policy-making, rather than sporadically voting in an election every number of years. Through open government data initiatives such as portals, stakeholders can also be more informed and be able to make better decisions [118].

The above motivations, while not being the sole ones, are the foundations for most open government data initiatives. Such initiatives act as a preventive policy and give stakeholders the opportunity to scrutinise and re-use the available information in a number of ways, including identifying patterns in the data and creating new services. This results in an increased accountability that in turn hinders corruption. Besides, through the creation of new services based on open government data, users add value to the data itself, which can also be commercialised. The participation of citizens in decision-making processes is also a very important aspect of opening governmental data, as it empowers citizens and thus enables governments to be more citizen-centred. However, citizen participation is not only limited to the decision-making process. Open government initiatives may also allow stakeholders to provide feedback on government actions or collaborate in policy-making.

Although the number of public entities seeking to publicly disclose their data has seen a drastic increase, it is still a major challenge to achieve the full potential of open government data and support all interested parties with the publication and consumption of this data. A number of barriers, including technical, policy and legal, economic and financial, organisational, and cultural barriers, also contribute to this challenge [27, 159]. Yet, a major stumbling block for the full exploitation of open government initiatives remains the *heterogeneous nature* of data formats used by public administrations, which include anything from images, PDF documents, CSV files and Excel sheets, to more structured XML files and database records. This heterogeneity is a technical barrier to both data providers and data consumers, and hinders society from realising government data transparency. Open government data portals also suffer from the *large number of diverse data structures* that make the comparison and aggregate analysis of government data practically impossible. The *diversity of tools* to present, search, download and visualise this government data is also nearly as diverse as the number of existing portals. Past efforts have sought to overcome this situation by creating comprehensive and connected European transparency portals such as PublicData.eu[11]. However, the diversity of transparency standards across Europe, which proved to be a bottleneck, highlighted the need that platforms beyond the state-of-the-art also need to be more than just direct entry points to government data analysis. They also need to provide a platform for advocacy towards common transparency standards at the highest level across several jurisdictions.

[11] http://publicdata.eu (Date accessed: 14 August 2016)

Government data portals also experience a number of cultural obstacles which hinder them from reaching their full potential. For example, public entities might be *unwilling to publish their data*. This may be so for a number of reasons, including the perception that it requires a lot of resources and effort, or that the release of government data might backfire. This disposition is, however, slowly changing world-wide, mostly due to advocacy of civil society initiatives.

2.1 Research Method

In Part II we use a systematic methodology to gather the relevant literature. By following this formal method with explicit inclusion and exclusion criteria, we intend to provide a replicable research review with minimal bias arising from the review process itself. Our approach for this systematic survey is based on the guidelines proposed in [39] and [69]. The procedure we undertake to find relevant literature is as follows:

1. Define search terms;
2. Select sources (digital libraries) on which to perform search;
3. Application of search terms on sources; and
4. Selection of primary studies by application of inclusion and exclusion criteria on search results.

Identifying the research questions is essentially what distinguishes a systematic review from a traditional review. Asking predefined questions is not only required for determining the content and structure of the review, but it also aids in guiding the review process. This includes the techniques used for identifying studies, the critical reviewing of studies, and the ensuing analysis of the results.

As part of the overall aims of this thesis, the goal of this survey is to analyse existing open government data initiatives, tools, and approaches, for publishing and consuming open government data. We can then use this information to define an open government data life cycle, with the further aim of specifying guidelines and recommendations on what are the best methods to publish and consume open (government) data, as well as how to get the most out of it. We therefore use the following research question to guide our research in the right direction, as defined in Section 1.2:

Research Question 1:

What are existing approaches and techniques that enable the publishing and consumption of open data?

This generic question can be further divided into more specific sub-questions which will be tackled in the next chapters in Part II.

1. What are the characteristics of existing implementations of open government initiatives?
2. What are the supported technical aspects, features and functions in existing approaches?
3. Are there any defined guidelines for the publishing or consumption of open government data?
4. What are existing challenges within open government initiatives?
5. What are possible impacts of open government initiatives on the relevant stakeholders?

2.1.1 Search Strategy for Systematic Survey

In order for our systematic survey to yield the largest spectrum of relevant publications possible, we identified and used the most extensively used electronic libraries, namely:

- ACM Digital Library
- IEEE Xplore Digital Library
- Science Direct
- Springer Link
- ISI Web of Knowledge

Although we considered Google Scholar for this systematic review, we decided against including it since its content is indirectly obtained through the listed electronic libraries, thus making the use of Google Scholar redundant.

Based on the research questions, we led out some pilot searches and consulted with experts in the field in order to obtain a list of pilot studies. The latter were then used as a basis for the systematic review in order to find the search terms which would best answer our research questions. The following are the search terms used in this systematic review:

1. “government data portal”;
2. “government public portal”;
3. “government open data”;
4. “government open data portal”;
5. “government open data publishing”;
6. “government data publishing”;
7. “public government data”;
8. “consuming open government data”;
9. “consuming open data”;
10. “public open data”;
11. “open data consumption”;
12. “open data publication”;
13. “open data portal”; and
14. “consuming public data”.

To construct the search string, all the search terms were combined by using the "OR" Boolean operator. The reason this conjoining method was implemented for the query construction was to keep the query as simple as possible, with as few Boolean operators as possible. This made the query more flexible to use in different electronic library search tools.

The next step in defining the search strategy was to find suitable metadata fields on which to apply the search string on. Searching in the publication title field alone does not always provide the relevant publications, mostly due to low precision rates. While the search on the title retrieves a potentially larger number of results, the results might not all be relevant. Thus by adding the search on the abstract, irrelevant results would be reduced, while other relevant publications which do not have the search terms in the title are also retrieved. We therefore decided to lead the search on both the title and abstract fields of publications.

2.1.2 Study Selection

Some of the results obtained using the above method might still be irrelevant for our main research question and the extracted sub-questions (Section 2.1), even if the search terms appear in either the title, abstract, or both. Therefore, a manual study selection has to be performed, retaining only those results which are relevant to the research questions. We hence defined inclusion and exclusion criteria as follows:

Publications that satisfy any of the inclusion criteria are selected as primary studies:

- I1. A study that focuses on open government portals, open government data, or its publishing or consumption;
- I2. A study that describes open government data initiatives.

Publications that meet any of the following criteria are excluded from the review:

- E1. A study that only mentions some of the search terms, but does not focus on government data or its publishing or consumption;
- E2. A study that focuses on open data in general (not limited to government data);
- E3. A study that describes portals that exploit only non-governmental data.

The procedure for selecting the primary studies for this review was conducted in October 2014. Consequently, this review only includes studies that were either published or indexed before that date. We also limited our search to publications written in the English language that were published after 2002. This year was selected as a delimiter since the preliminary search indicated that there were no relevant results before that date. As shown in Figure 2.1, we started by applying the search string in each data source separately. Since the results included a couple of proceedings, we resolved them by including all publications within the proceedings, resulting in 368 publications. Subsequently, the results were merged, and duplicate studies were removed. This left us with 338 publications. We then proceeded to manually go through the titles of the remaining studies, removing those entries whose title indicated that they were not relevant to our review. This reduced the amount of potential primary studies to 159. The following step was to manually scan the abstracts. Yet again, the number of studies was reduced to 103. Finally we went through the full-text of the studies, whilst applying the Inclusion and Exclusion criteria defined above. This resulted in 75 studies, which represented our final set of primary studies.

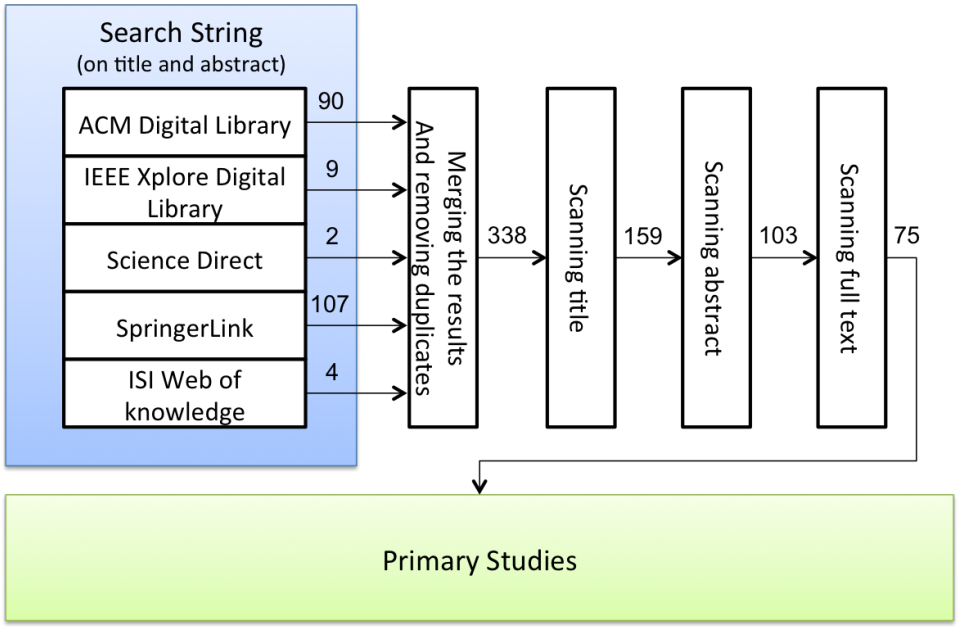

Figure 2.1: Procedure for identifying primary studies.

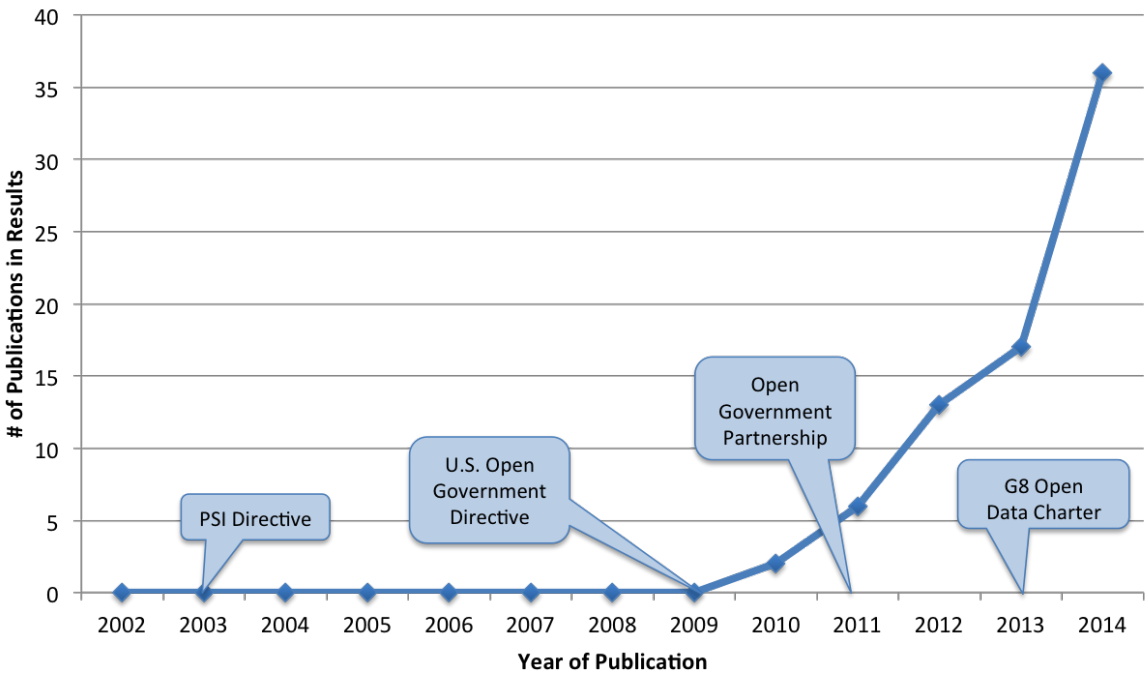

Figure 2.2: Resulting number of primary studies shown by year published.

2.1.3 Overview of Included Studies

The goal here is to execute a systematic analysis of existing literature within the field of open government data. We therefore discuss some statistics of the *relevant* literature resulting from the conducted systematic analysis. As shown in Figure 2.2, the period between 2002 and 2009 did not yield any relevant literature, however, the results increase significantly in the subsequent years. Even though a number of major open data initiatives were already established, such as the ones indicated in the figure, the surge in open government data literature may potentially be linked to U.S. President Obama's Open Government Directive at the end of 2009. As shown in the image, the year 2014 resulted in the highest number of related literature (as per the time the study was conducted), indicating that the awareness of open government initiatives is increasing at a fast pace.

2.2 Terminology

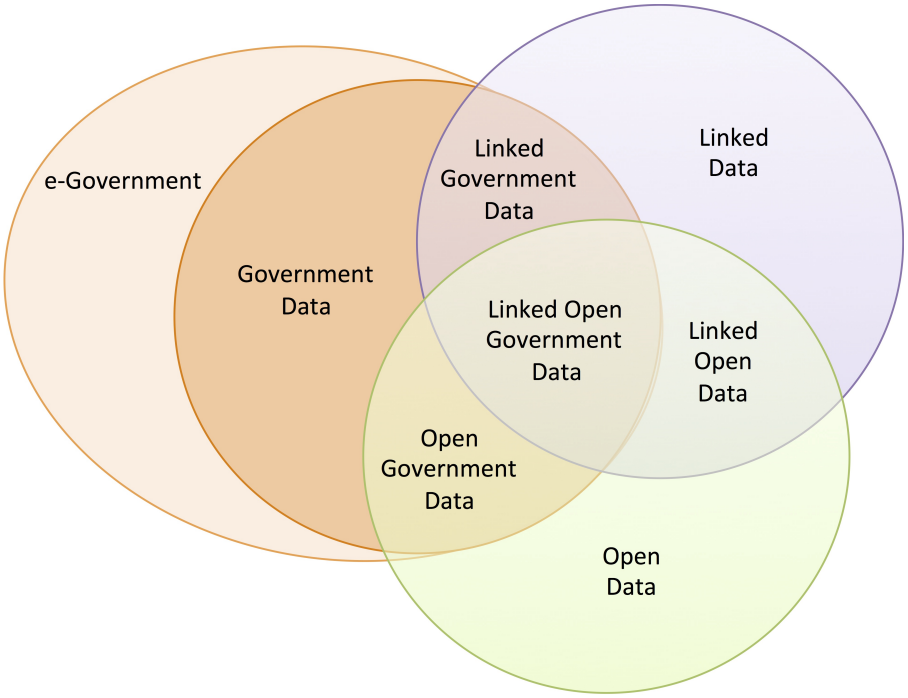

Figure 2.3: Relationship between open data, government data, and Linked Data.

In order to give some context to our discussion, we here define the most important concepts used within this thesis, as identified within the relevant literature. Figure 2.3 visually represents the relationships between open data, government data, and Linked Data.

Open Data - The 'Open' [12] definition sets out eleven requirements that Open Data should conform to. The latter requirements basically indicate how to enable the free use, re-use, and re-distribution of data. Moreover, open data should not discriminate against any person and must not restrict the use of the data to a specific field or venture. Thus, data published in an open data format would be "platform independent, machine readable, and made available to the public without restrictions that would impede

[12] http://opendefinition.org/od/ (Date accessed: 2 August 2016)

the re-use of that information"[13]. Hence open data only refers to data that is available free of charge for the general public without any limitations [116]. Open data is considered to be a key enabler of Open Government [71].

Public Data - It is important to note the distinction between public data and open data. While public data is made freely available to the general public, it is not necessarily open. An extreme example of public data which is not open is an archive of legal documents. While they are freely accessible, imagine the effort required to identify and locate a specific document. On the other hand, if such data is digitalised and made available online in a standardised format (also indexed), then this public data is also open.

Open Government Data - Open Government Data is a subset of Open Data, and is simply government-related data that is made open to the public [71]. Government data might contain multiple datasets, including budget and spending, population, census, geographical, parliament minutes, etc. It also includes data that is indirectly 'owned' by public administrations (e.g. through subsidiaries or agencies), such as data related to climate/pollution, public transportation, congestion/traffic, child care/education. This is known as public sector information (PSI). Several countries have already demonstrated their commitment to opening government data by joining the Open Government Partnership (OGP)[14].

E-Government - While many different definitions of e-government exist in the literature, we here stick to the government's use of technology to enhance the services it offers to other entities, including citizens, business partners, employees, and other government agencies [73]. Technologies used for this purpose are most often web applications. Thus, by aiding the interaction between citizens and their government, an e-government has the potential of building better relationships and also deliver information and services more efficiently. While initially e-government just referred the simple presence of government on the Internet, mostly in the form of an informative website, the concept has since evolved. With the introduction of the 'Open Government' concept, we now consider open government data initiatives to be a subset or an extension of e-government [63].

Linked Data - Linking data is the process of following a set of best practices for publishing and connecting structured data on the Web [17]. It is the final step in the Five Star Scheme for Linked Open Data[15]. The term 'Linked Data' thus refers to data which is published on the Web and, apart from being machine readable, it is also linked to other external datasets. The increased rate of adoption of Linked Data best practices has lead the Web to evolve into a global information space containing billions of assertions, where both documents and data are linked. The evolution of the Web enables the exploration of new relationships between data and the ensuing development of new applications.

Data Portal - The open data movement aims at opening public sector information with the purpose of maximising its re-use. A typical implementation is to collect and publish datasets into central data portals or data catalogues in order to provide a "one-stop-shop" for data consumers. While a data catalogue would most commonly act as a registry of data sources [1] and provide the relevant links, a portal is more commonly a single entry point hosting the actual data, where end users can search and access the published data and explore or interact with it in some manner. A key function of a data portal is the management of metadata for the datasets, possibly including metadata harmonisation. Various tools are provided on government data portals, such as data format conversion, visualisations, query endpoints, etc.

Publishing - Publishing data on the Web enables data creators to add their data to the global data space. This allows data consumers to discover and use this data in various applications. By following Linked Data best practices, published data is made more accessible and eases its re-use. A large number of Linked Data publishing tools exist; they either serve the content of RDF stores as Linked Data on the

[13] http://www.whitehouse.gov/open/documents/open-government-directive (Date accessed: 2 August 2016)

[14] http://www.opengovpartnership.org/countries (Date accessed: 2 August 2016)

[15] http://5stardata.info (Date accessed: 2 August 2016)

Web or otherwise provide Linked Data views over non-RDF data sources [17]. The majority of these tools allow publishers to avoid dealing with the technical details behind data publishing.

Consuming - The aim of publishing data on the Web is to enable its use, re-use, and distribution. Such data is made more discoverable and accessible if the data publishers follow Linked Data best practices. For example, if published data has good quality metadata [116], then consumers would more easily discover the contents of the published dataset, and decide whether it is fit for the intended use. While the role of data consumers and data publishers is distinct, it is also interchangeable in that a publisher can also be a consumer and vice versa. To describe this, the authors of [2] coin the term *prosumers*. Data consumption can be either data *exploration*, where a user visualises or scrutinises open data, or data *exploitation*, where a user adds value to the open data by creating mashups, leading analysis, or innovating upon the data itself. This is also known as *knowledge economy*.

Data Quality - Since the concept of quality is cross-disciplinary, there is no single agreed-upon definition of quality [71]. However, data quality is commonly perceived to be *fitness for use* [65]. Fitness for use is, however, a multi-dimensional concept that has both subjective perceptions and objective measurements based on the dataset in question [109]. Subjective data quality assessments reflect the requirements and experiences of the consumers of the data. Let us take an example using restaurant reviews. What one person might consider to be a tasty dish, another might find bland. These different perceptions result in varying reviews of the same dish. Objective assessments can be task-dependent or task-independent. Task-independent quality assessment metrics reflect the properties of the data without contextual knowledge of how it will be consumed. Continuing on the previous example, if a restaurant uses fresh ingredients in its food, then it is considered to be a good restaurant. Task-dependent metrics, on the other hand, reflect the requirements of the application at hand. For example, if a person who does not like fish is served a fish dish, then of course he will not like it. Thus, albeit a public entity publishes governmental data, if this data does not have good quality standards with regard to its consumers, then the data will not be exploited to its full potential.

2.3 Open Government Data Life Cycle

In this section we propose and explain the open government data life cycle. Albeit a number of open data life cycles exist[16], most of them are not tailored to reflect the specific features of open government data. Other publications, such as [163], do explore government-focused processes, however some vital steps are missing, and only the most common procedures for opening data are discussed. Therefore, using the existing data life cycles as a basis, as well as other open government data life cycles identified in the primary studies, we here attempt to cover all the processes in the life cycle of open government data, in order to provide a standard process that government open data stakeholders can follow.

The proposed life cycle, shown in Figure 2.4, is made up of three sections, namely a *pre-processing* section (rectangle), an *exploitation* section (oval), and a *maintenance* section (hexagon). The latter sections, in order, take care of: (i) preparing the data to be published, (ii) using the published data, and (iii) maintaining the published data in order for it to be sustainable. We proceed to give an overview of each independent step in the life cycle.

- **Data Creation** - The open government data life cycle typically starts with the creation of data. In public or governmental entities, the creation of data is usually part of daily procedures, however, it is also possible to collect data for the specific purpose of publishing it.

[16] http://www.w3.org/2011/gld/wiki/GLD_Life_cycle (Date accessed: 2 August 2016)

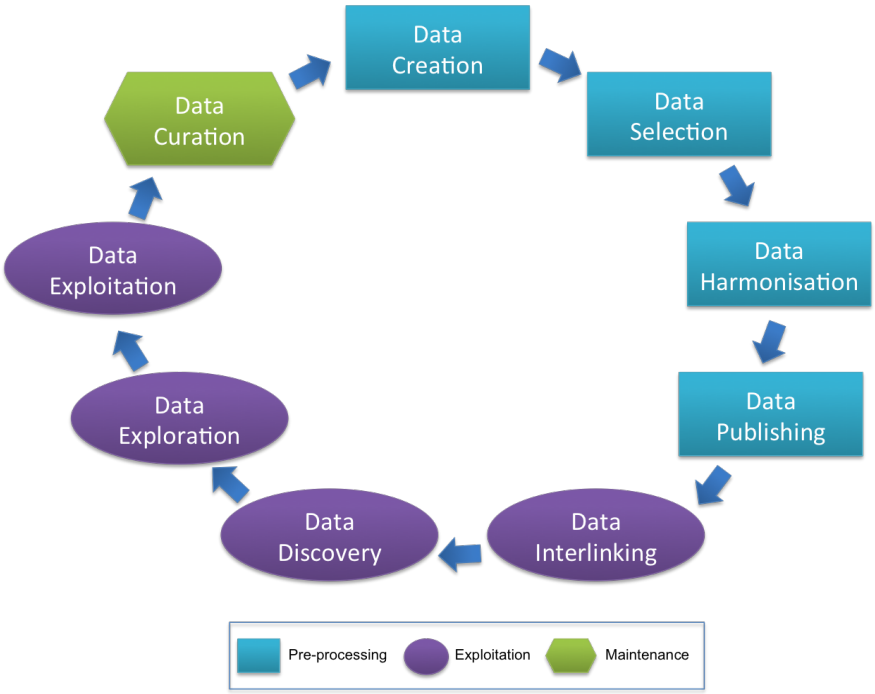

Figure 2.4: Open government data life cycle.

- **Data Selection** - This is the process involving selecting the data to be published. This requires removing any private data or personal data, as well as identifying under which conditions will this data be published (potentially through the specification of open government data policies) [165].
- **Data Harmonisation** - This step involves preparing the data to be published in order to conform to publishing standards, such as the Eight Open Government Data Principles (explained further in Section 4.1.2).
- **Data Publishing** - This is the actual act of opening up the data by publishing it on government portals.
- **Data Interlinking** - Data Interlinking is the final step in the Five Star Scheme for Linked Open Data. This allows published data to have additional value, as the linking of data gives context to its interpretation.
- **Data Discovery** - The publishing of data is not enough to enable its re-use. Data consumers must discover the existence of open data in order to be able to consume it. Data discovery can be enhanced by actively raising awareness on its existence (e.g. through organising hackathons).
- **Data Exploration** - This step is the most trivial way of consuming data. Here, a user *passively* examines open data by visualising or scrutinising it.
- **Data Exploitation** - This step is a more advanced way of consuming data. Data Exploitation enables a user to *pro-actively* use, re-use or distribute the open data by leading out analysis, creating mashups, or innovating upon the open data.

- **Data Curation** - While not necessarily occurring at a fixed stage, data curation is vital in ensuring the published data is sustainable. This involves a number of processes, including updating stale data, data and metadata enrichment, data cleansing, etc.

In this part of the thesis we focus on the processes of publishing and consuming open data, as these processes are essential to open data. Basically, a data life cycle such as the one we define will not exist without the initial publishing of data which makes it available for use, and its consumption. Whilst the consumption of data can be taken to be any time a stakeholder manipulates data, such as to curate it, in this part of the thesis we focus on the more literal meaning of consuming, whereby a user, either passively through Data Exploration, or more actively through Data Exploitation, makes use of data to achieve a particular goal. Such a goal is also usually the achievement of information or knowledge, as opposed to the simple manipulation of the data. The rest of the steps in the life cycle, whilst certainly also important, are not as crucial and a life cycle can exist in their absence. For this reason in the following chapters we direct our discussion on the processes of publishing and consuming open data.

CHAPTER 3

Open Government Initiatives

The Open Government movement aims to achieve a government that enables cooperation between public administrations and the general public, in order to become more transparent and democratic [98]. Open government data does not only enhance the transparency and accountability of a government, but can result in economic benefits, innovative solutions for community advancement, as well as supporting public administrations' functions [11, 16, 44, 45, 67, 71, 76, 82, 90, 92, 105, 119, 136, 139]. Furthermore, these benefits can be achieved simply by publishing and re-using data which has already been produced in the day-to-day administration of a governing entity. We can thus assume that the two major motivations which prompt governments to jump on the open data bandwagon are: (i) the spirit of democracy, and (ii) economics [85]. Regarding the first motivation, governments exploit open data initiatives in order to lift the veil of secrecy and become more transparent. The second motivation, on the other hand, enables the growth of the information marked by sharing government data. Whilst sensitive or personal data cannot be shared, other data can have economic value to businesses or individuals if exploited, and new uses for the particular data can also be discovered. The publishing of data, such as traffic, meteorological, budgetary, geo-spatial, and geographical data, provides consumers with opportunities to create new services, which, apart from being profitable, can also benefit the common good [16]. This, in turn, can potentially contribute to economic growth. Other important benefits resulting from open government initiatives include crowdsourcing for error reporting, increased public service employee motivation due to the re-use of published data, more informed citizens, enhanced citizen participation, and job creation [105].

To date, 64 countries have joined the Open Government Partnership[1] (OGP) to demonstrate their commitment to making data free to use, re-use and redistribute according to Open Data principles. The OGP initiative aspires to guarantee commitments from governments to promote transparency, accountability, empower citizens, and exploit technologies to strengthen governance. In order to be eligible to participate in the OGP, countries (and their respective governments) should meet the eligibility criteria and demonstrate their commitment to open government principles in four key areas:

1. Fiscal transparency;
2. Access to information;
3. Income and asset disclosures; and
4. Citizen engagement.

[1] http://www.opengovpartnership.org/ (Date accessed: 2 August 2016)

A typical implementation to opening government data is to collect relevant datasets and their respective metadata and publish them on an *Open Government Data Portal*. Open Government Data Portals can have different operators, i.e. either an official government entity or a citizen initiative. Another difference between open government implementations is the scope, where a portal or catalog may publish data relevant to a specific administrative region, for example, a city or a country. A large number of countries have created local or national government data portals in order to provide access to open government datasets [87]. Four major sites to date are in the US (http://data.gov), the UK (http://data.gov.uk), France (http://data.gouv.fr), and Singapore (http://data.gov.sg) [53]. Such portals act as one-stop-shops and facilitate consumers' access to government data, saving the trouble of collecting data from various authorities, offices, or websites.

While the main implementations of open government data initiatives are data portals, there exist a number of different implementations with various characteristics. *Government Data Catalogues* or *Metadata Portals/Repositories* are indexes which store structured descriptions (metadata) about the actual data (e.g. http://PublicData.eu). Such tools have the potential of improving the discoverability of published datasets, as the discoverability of data is directly dependent on the quality of the metadata [116]. An open government catalogue would contain a collection of metadata records that describe open government datasets and also have the corresponding links to the online resources [71, 86]. The implementation of a catalogue, however, raises an important question: *What metadata should be stored and how should it be represented?* This question is especially significant when automatic importing of metadata records (also known as *harvesting*) is performed, as metadata structure and meaning are not usually consistent or self-explanatory [86]. Open data portal software such as CKAN[2] or vocabularies such as DCAT [82] provide solutions for this problem. Furthermore, the authors of [116] propose the implementation of metadata quality metrics on CKAN-powered government data catalogues with the aim of determining the metadata's adequacy for a user's specific need.

Figure 3.1 shows the *2014 Global Open Data Index*[3] of a number of places (some places might not be officially recognised as countries). This index tracks whether published data is actually released in a way which is accessible to all stakeholders, and measures the openness level of data globally. The index represents the percentage of dataset entries that are deemed to be open, based on the Open Definition[4]. The technical and legal dimensions of each dataset available from the various places is assessed using the following nine questions:

1. Does the data exist?
2. Is the data in digital form?
3. Is the data available online?
4. Is the data machine-readable?
5. Is it available in bulk?
6. Is the data provided on a timely and up to date basis?
7. Is the data publicly available?
8. Is the data available for free?

[2] http://ckan.org (Date accessed: 2 August 2016)
[3] http://index.okfn.org/place/ (Date accessed: 2 August 2016)
[4] http://opendefinition.org/od/ (Date accessed: 2 August 2016)

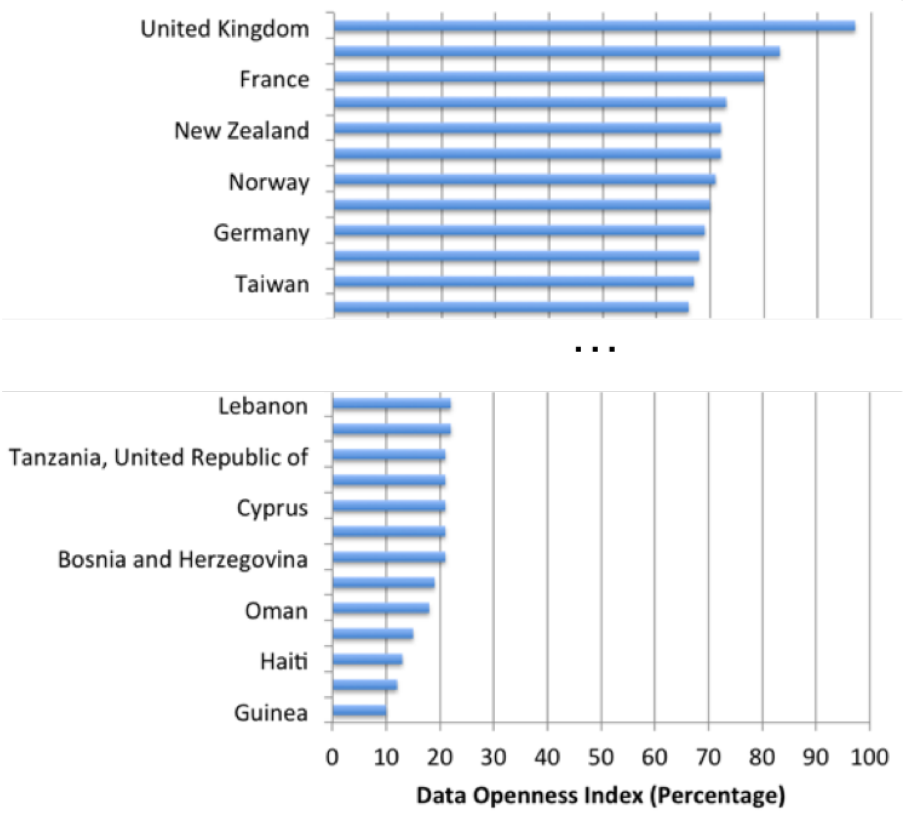

Figure 3.1: Global Open Data Index for a sample number of places for the year 2014 (Source: http://index.okfn.org).

9. Is the data openly licensed?

3.1 Assessment Frameworks

It is undeniable that with all the current open government initiatives, a large amount of data has been released to the public. This, however, does not mean that the targeted aims of promoting transparency and facilitating accountability have been achieved yet. For example, Arcelus points out that after interacting with transparency websites (such as data portals), consumers do not consider that transparency and access to information have been achieved [4]. In such cases, while these portals would be complying with the law and following the requirements for the publishing of information, they would not be promoting transparency in itself. Unfortunately, while being aware of such deficiencies, several governments do not tackle them. This is because government initiatives are evaluated according to whether they are complying with the law or not, and not based on the usefulness of the information provided. A very apt example in this case is the publishing of data in PDF format, which makes it pretty inconvenient for any intended use, re-use and re-distribution.

Another contributor to the afore-mentioned deficiencies is the lack of an agreed-upon framework to evaluate and assess the content provided on such data portals [4]. Whilst various authors propose and discuss recommendations and requirements for evaluating data portals or catalogues, the contribution varies

from publication to publication. For example, Susha et al. analyse a number of different benchmarks with the aim of identifying the strengths and weaknesses in assessing the application of open data in an open government context [140]. Yet, the majority of the proposed frameworks and recommendations reflect the *Five Star Scheme for Linked Open Data* by Tim Berners-Lee[5], such as [54], as well as the *Eight Open Government Data Principles*[6], such as in [79]. Table 3.1 gives an overview of the different aspects on which the authors of the specified literature focus on within the discussed assessment frameworks. In this table, by 'Nature of Data' we mean the assessment of various data aspects according to the Five Star Scheme for Linked Open Data, and the Eight Open Government Data Principles.

	Data		*Portal*			*External Factors*		*Public Engagement*	
	Nature of Data	**Accountability**	**Transparency**	**Access to Information**	**Openness**	**Legal Obligations**	**Institutional Arrangements**	**Participation**	**Collaboration**
[4]		✓	✓	✓		✓			
[18]	✓		✓		✓			✓	✓
[54]	✓								
[79]	✓	✓	✓	✓					
[124]		✓			✓	✓	✓		✓
[148]	✓		✓		✓			✓	✓

Table 3.1: Overview of aspects evaluated by assessment frameworks proposed in literature.

Arcelus [4] introduces a number of elements that should be considered in an assessment framework. Focusing on transparency, access to information, and accountability, the author also mentions the importance of the usability and accessibility of the portal in enabling stakeholders to access the required data. In context of existing regulations for government data portals, Arcelus also proposes to strive to raise the bar of what are considered to be the minimum requirements for data provision.

Bogdanović-Dinić et al. [18] propose a model for assessing data openness by relying on the Eight Open Government Data Principles. With the aim of of automatically evaluating openness, the model was implemented as a web tool, and it also aids in the process of building openness principles. The authors applied the model to seven data portals with the aim of demonstrating its capabilities and results.

In [54], Hoöchtl and Reichstaädter focus on the data within a government portal and define an architectural model. The authors here propose how the ideal data should be presented, basing their contribution on the Five Star Scheme for Linked Open Data.

In [79], Lourenço proposes an assessment framework. The author focuses on accountability, aiming to propose a set of requirements intended to assess whether portals are actually contributing to a higher degree of transparency. The author raises two essential questions regarding the effectiveness of portals in making data available for accountability purposes, and how this can be evaluated. Lourenço analyses the related literature on internet-based transparency and considers two dimensions, namely the type of public entities studied and the information types that consumers looked for. The author proceeds to extract a set of requirements from the led studies, and proposes them as part of an open government dataset portals assessment framework. Finally, the author concludes that while dataset portals were created with the intention of meeting open government strategies, as yet no evidence was found in open government literature that publishing a large amount of datasets actually contributes to promoting transparency and facilitating accountability.

Another assessment model is proposed by Sandoval-Almazan and Gil-Garcia in [124], whereby the authors analyse existing assessment models and then proceed to propose a new model catering for previous discrepancies in the older models. They proceed to test and analyse the proposed assessment model using actual open government data portals.

[5] http://www.w3.org/DesignIssues/LinkedData.html (Date accessed: 2 August 2016)
[6] https://public.resource.org/8_principles.html (Date accessed: 2 August 2016)

In [148], Veljković et al. propose a benchmark for open government data initiatives. The benchmark is based on data openness, transparency, participation, and collaboration. It assesses both the openness index as well as the maturity of relevant initiatives.

3.2 Open Government Initiative Evaluations

The number of evaluations carried on existing portals, catalogues, and other open data initiatives is nearly as varied as the number of initiatives itself. Furthermore, since there is no agreed-upon evaluation framework as yet [18], the authors of such literature employ different approaches. Table 3.2 shows an overview of the approaches undertaken within literature covered in the rest of this section. One should note that while all authors assess various aspects of an initiative, they base their evaluation on one or two main aspects. Most of the evaluations assess the published data properties of the initiative in question using the Five Star Scheme for Linked Open Data or the Eight Open Government Data Principles, however others also assess data availability, data content, and data accessibility. Two other popular assessment approaches consider the features and functions of an initiative (usually in the form of a portal). In contrast to the above approaches, some authors assess the maturity of an initiative as a whole, rather than based on specific aspects such as data, functions or features. In the latter cases, the maturity is assessed based on other aspects such as the amount of fulfilled objectives, compliance to existing laws and regulations, and the usability from a stakeholder's point of view. A number of approaches in literature also consider stakeholder participation in the initiative in question, as well as their feedback.

	Data	Functionality	Features	Stakeholder Participation	Initiative Maturity	Stakeholder Feedback	Evaluated Initiatives	Geographic Coverage
[1]	✓	✓	✓				Various Portals	Greece
[36]	✓	✓			✓		Rio Inteligente and Cidadão Recifense	Brazil
[37]			✓			✓	Meu Congresso Nacional Application	Brazil
[42]					✓		http://data.wien.gv.at	Vienna, Austria
[45]	✓		✓				Various Mobile Applications	Mexico
[46]	✓		✓				http://datosabiertos.df.gob.mx, http://labplc.mx/hackdf-2	Mexico
[63]					✓	✓	n/a	Stockholm and Skellefteå, Sweden
[76]	✓	✓					Various Portals	Taiwan
[77]			✓				Various Portals and Agencies	Australia
[86]	✓	✓					http://GovData.de	Germany
[87]	✓						http://PublicData.eu	Europe
[89]	✓		✓		✓		Various Entities and Portals	Brazil
[90]		✓					Mato Grosso, Paraíba, Piauí and Paraná	Brazil
[104]	✓				✓		Various Portals	Italy
[105]						✓	n/a	Vienna, Austria
[108]	✓	✓	✓				Various Portals	European Union
[114]					✓		http://www.datos.gov.co	Colombia
[118]	✓						http://datosabiertoscolombia.cloudapp.net	Colombia
[123]	✓			✓	✓		n/a	Colombia, Chile, Brazil
[125]			✓				Various Mobile Applications	Various Countries
[126]	✓		✓	✓			n/a	Worldwide
[146]	✓	✓				✓	RASOI Mobile Application	India
[153]	✓	✓					http://www.stat.gov.rs, http://PublicData.eu, INSIGOS	Europe, Serbia, Poland
[156]					✓		n/a	Taiwan
[162]	✓						EU Open Data Hub, Junar, ENGAGE 2.0	Various

Table 3.2: Overview of evaluated aspects in open government initiative evaluations.

Within the results of our systematic survey, one of the most popular approaches was to evaluate the functionality of portals or catalogues. The authors of [86, 90, 153, 162] all focus on this approach, for http://GovData.de, Brazilian anti-corruption and transparency portals, http://PublicData.eu, and the EU Open Data Hub, Junar, and ENGAGE 2.0 respectively. Through the four publications, the authors assess these portals by identifying the functions, limits, and challenges of the evaluated portals, and also give recommendations towards avoiding or solving the challenges.

In [86], Marienfeld et al. describe how http://GovData.de achieves its functionalities. They

analyse the metadata structure, how the data is harvested, and how this portal aligns with other EU activities. The authors then proceed to provide an insight on the portal's sustainability issues, such as an increasing amount of datasets and data providers, and the change of governmental institutions (resulting in a shifting of responsibilities). To conclude, the authors identify a number of challenges for portals such as http://GovData.de.

The aim of the authors of [90] was to find the limits and challenges of Brazilian anti-corruption and transparency portals through reviewing the existing literature. After assessing the existing literature, Matheus et al. compare the findings based on Web 2.0 usage and new Information and Communication Technologies (ICTs). The authors proceed to discuss the limits and challenges of the portals in question, and give recommendations towards avoiding and/or solving them. A specific limitation that the authors encountered is that most portals consider transparency to be limited to offering data to the consumers, without proceeding to provide help for its consumption.

In [153] a description of the key functionalities of open data portals is presented. Ubaldi uses http://PublicData.eu as an example throughout the publication. The author identifies two essential factors required for opening governmental data; namely discoverability and good quality data. Regarding the former factor, the author points out the importance of data catalogues, and also the lacking efforts with regard to cross-language searching of open datasets.

Zuiderwijk et al [162] propose a framework of 35 functionalities for comparing open data infrastructures and consequently compare three different open data infrastructures. The authors focus on the functionality of the infrastructures, rather than on the quality of the provided data or other requirements such as reliability and scalability. They concentrate on the requirements for the open data process, such as the creation, publication, finding, use, and linkage. The authors hence identify various differences within the assessed infrastructures, but also note that none of the infrastructures supports all aspects of the open data process.

Another popular approach was to evaluate the features provided in data portals and their usability, such as the number of data formats available and multilinguality. The authors of [46] and [77] both evaluate, for different use-cases, how portals and catalogues actually enable the use, re-use, and distribution of data. They identify shortcomings such as consumers' difficulty in identifying the required datasets and the use of different formats.

Liu et al. [77] analyse a number of Australian catalogues that provide sustainability-related data with the aim of investigating a Linked Data approach for supporting sustainability researchers in their efforts to create and investigate sustainability science hypotheses. The authors observe that different government entities may contribute to the same field of data from different perspectives. Thus it would be difficult for data consumers to quickly and easily identify the most useful dataset for their use-case unless they know the responsibility of each entity who published the dataset. Another observation is that different government entities publish their data in different formats. This results in extra efforts required to convert the datasets into one format, if even possible.

Similar to the above, González et al. [46] focus on Mexican open government data implementations. The authors focus on both the demand and supply side of open data and note that while the law mandating governments to open their data has been in force for ten years, there are still various challenges to achieve the true potential of open government data.

In [45], Fuentes-Enriquez and Rojas-Romero research the use of mobile applications in Mexico by all stakeholders of open government data. They report that the involved entities have an active participation in the creation of government mobile applications. The authors also explore a number of applications developed by different stakeholders, including the government, private organisations, and also citizens, and identify their contribution in allowing the stakeholders to actively participate in governance processes. A similar research is carried out by Sandoval-Almazan et al. in [125], where the authors analyse the use

of open data and mobile applications in a number of countries.

In [126], Sayogo et al. carry out a preliminary exploration of the worldwide status of open government. The authors analyse open government data portals from 35 countries, reviewing the published data, the provided features, and the level of stakeholder participation. They also provide a framework for assessing open government data initiatives. The authors of [114] and [118] also evaluate the status of open government initiatives, however, they directly focus on the Colombian government initiative as a whole, rather than for specific portals. Similarly, the authors of [36] and [89] both discuss the current state of Brazilian open data initiatives. The authors of [76] and [156] both lead out a study on the Taiwanese open data platforms with the aim of identifying their status. The Greek open data movement is analysed in [1], where the authors analyse the current state of open data from three different perspectives, namely the functionality, the semantics of the data, and the provided features.

The authors of [42], in a somewhat non-traditional evaluation, attempt to analyse the relationship between the open data ambitions at the European level and those at the Austrian federal level (focusing mostly on Vienna), both from the data consumer and producer side. The led study attempts to identify to what extent developments at the EU level influence Austrian practices with regard to open data initiatives. The authors, Egger-Peitler and Polzer, point out that the efforts within Vienna are mostly decoupled from EU strategies.

The authors of [87] analyse the `http://PublicData.eu` catalogue and assess the metadata recorded for each dataset within it. The objectives behind this study are twofold: firstly to identify the quality of a sample of metadata properties and secondly to study the stated level of data openness. The authors use Tim Berners-Lee's Five Star Scheme for Linked Open Data as an evaluation scale, and assess the data format, licences, and level of openness.

In publication [104] a number of Italian Municipalities' portals are evaluated with the aim of understanding the link between the Open Government Data legislation and a newly enacted Transparency Act. Palmirani et al. theorise that the latter does not enable and enhance open government data. After the led evaluation the authors conclude that the Transparency Act has affected badly the quality of the published datasets, as it is only oriented towards reducing corruption, rather then enabling open government data as a means to that end.

Publications [37, 63, 105, 146] all assess stakeholders' opinion to a certain degree. Through interviews and online polls, the authors of [105] identify a number of factors that enabled the success of the open government data strategy in Vienna. The authors of [37] and [146] discuss issues and challenges of developing applications that implement open government data. While the former extract these challenges from evaluating and analysing an application developed during an organised hackathon, the latter describe the challenges they faced during the development of their own application. Finally, the authors of [63] analyse the interpretations and perspectives of stakeholders with regard to opening municipalities' data in Sweden, and strive to identify how the stakeholders' opinion contributes to the implementation success of open data initiatives.

Petychakis et al. [108] do not focus on a single aspect for their evaluation. Rather, the authors carry out a comprehensive analysis of open government data initiatives in the European Union, focusing on functions, data semantics, and features. They collect and categorise a number of public data sources for each European Union member country, and they assess their characteristics and provided services. The authors identify the differences in content, licences, multilinguality, data accessibility, data provision, and data format. Finally, the authors point out that while the quality of open government infrastructures is improving, there are still great differences between national open data portals. Petychakis et al. also identify two important challenges which are still not catered for, namely multilinguality and open licences. In a similar but downscaled manner to [108], the authors of [123] compare three South American open government data initiatives (Brazil, Colombia, Chile), however, the authors rather focus on open

government policies, citizen involvement and the use of new technologies.

3.3 Stakeholders

Data, whether government data, public sector information, or other privately-owned data, is a resource holding great potential for a large number of stakeholders. Taking open government initiatives as a use case, governmental agencies, citizens, non-profit organisations, and businesses, are but a few of the potential stakeholders who, through the exploitation of open government data, can reap substantial benefits. Since the efforts of the latter stakeholders remain largely uncoordinated, their motivations, levels of expertise, and priorities differ. In this section we proceed to identify and explore the various stakeholders who can participate in an open government data initiative.

The most obvious role of **governments** in open government data initiatives is the role of a data publisher or creator. Yet, public entities are also the direct beneficiaries of their own published data. Through transparency as a motivation, the publishing of data can increase accountability, and moreover inhibits corruption. In turn this can increase citizens' trust in their government. The analysis of government data, such as budget data, has the potential of increasing efficiency and influencing decision-making. Innovations based upon such data can also be used to provide more personalised public services, thus increasing the quality of the interactions between governments and their citizens.

Through the publishing of government data, **citizens** are given the possibility of participating in governance processes. Apart from being able to make more informed decisions, citizens are sometimes given the opportunity to take part in participatory governance. For example, in a participatory budget effort citizens are given a say as to how, or for what, budget should be prioritised. Citizens can also participate in open government initiatives by being data *prosumers*. By this we mean citizens who both produce and consume data. For example, the Fix My Street[7] application provides a platform where anyone can submit an existing problem in a street, in order to indicate the problem areas to the government. In this crowdsourced co-production of value, we have geographical data consumption, and street issues data production. Open government data certainly has the potential of increasing citizens' quality of life.

Non-profit organisations, such as non-governmental organisations (NGOs) or Civil Society initiatives, can have huge differences in their goals, however they usually share the goals of demonstrating the benefits of opening governmental data both to the general public and to the governments themselves. They also play a vital role as intermediaries who can identify key datasets that have the potential of being very valuable if published as open data. Examples of such organisations include the Sunlight Foundation[8] and the Open Knowledge Foundation[9], present in various countries.

Private companies, small to medium enterprises (SMEs), entrepreneurs, and other **businesses**, have the potential of not only making an economic profit through using government data, but can also create more jobs, and (depending on the nature of the service) also provide innovative services that increase the beneficiaries' quality of life and indirectly impact job creation in this field. While the sole access to data does not provide competitive advantage, private entities can innovate upon the available data to provide value-added services.

[7] https://www.fixmystreet.com/ (Date accessed: 2 August 2016)
[8] http://sunlightfoundation.com/ (Date accessed: 2 August 2016)
[9] https://okfn.org/ (Date accessed: 2 August 2016)

3.4 Impacts

Open government data initiatives are based on *transparency*, *citizen participation*, and *collaboration* for strengthening democracy [4, 41, 85, 90, 98, 157]. Through these three pillars, the publishing of government datasets not only has the potential of improving accountability and decreasing corruption, but it also affects all the involved stakeholders in a number of ways. In this section we discuss the impacts of open government iniatiaves.

While there is an obvious niche in literature with regard to frameworks which assess the impact achieved through open government data initiatives, a number of authors discuss the different impacts that can be obtained through such initiatives. López-Ayllón and Arellano Gault [78] depict the different levels of impact that can be achieved by an open government data initiative. We adapt these levels in Figure 3.2 and portray, in context, how each impact builds upon or supports the other impacts. While each impact does not strictly require the previous one, each impact supports the next one to achieving a higher level of impact on the relevant stakeholders.

As shown in Figure 3.2, the most direct impact is **access to information**. Once data is published (made open to a given degree), this impact is immediately effective, since it provides the means for data to be re-used. Of course, the data's re-use is conditional on how the data is published (its level of openness), and the consumer's willingness to participate in such an effort. Through providing access to relevant information, an open government data initiative can be more transparent.

Transparency, the second level of impact for publishing government data, can result in a considerable increase in social control by citizens through enabling them to scrutinise the data. Subsequently, if provided with the relevant means, they can also provide relevant feedback to the data publisher, and monitor policies and government initiatives [37, 90, 156]. Consequently, stakeholders gain more responsibilities as they are able to interact with the government and other public entities more actively than in traditional governmental structures. For example, following the publishing of budget data, stakeholders such as citizens, NGOs and even other private entities can provide feedback on budget

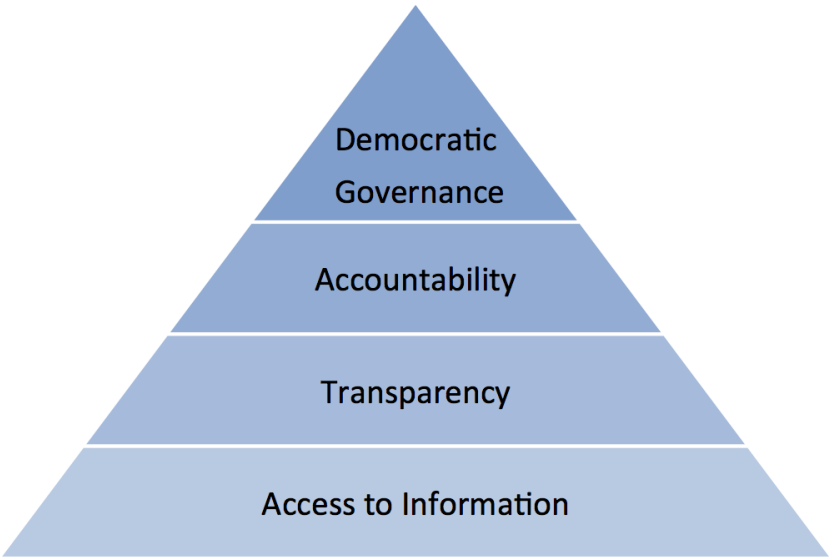

Figure 3.2: Relationship between different impacts of open government data initiatives.

priorities and specific transactions[10]. Therefore, by easing social control, open government data initiatives allow citizens to further exercise their duty and right of participation. Moreover, it helps citizens establish a trusting relationship with the government, which is able to prove legitimacy of the actions taken. The latter can however backfire if the media focuses only on issues based on the published data and portrays a negative image of the government.

The increased transparency resulting from publishing data will also impact public administrations in that there will be enhanced **accountability** within public sectors. In his publication, Bovens [20] defines accountability as the disclosure of data that provides stakeholders with the information required for assessing the propriety and effectiveness of the government's conduct, while the authors of [78], López-Ayllón and Arellano Gault, identify accountability as having a dimension of answerability. They separate the latter into two components, namely information and justification. The first implies that there should be an entity that is obliged to provide information to which the stakeholders should have access. Justification, on the other hand, is more challenging to achieve since it implies that the data-providing entity should justify their actions to the citizens. Yet, as Lourenço [79] points out, even if the published data is usable and adheres to good quality standards, the simple provision of data does not guarantee that the public entity or government is immediately enhancing transparency and/or accountability.

Through the long term interaction with an open government data platform, open data promotes not just transparency and accountability, but also **democracy** [98]. As mentioned in the example for the budget data, stakeholders can be enabled to provide feedback on the published data. Such feedback loops will not only inform the public entity of the public opinion, but also can improve service delivery through the repeated querying of the open data by all stakeholders, including citizens and government agencies. For example, the analysis of published budget data would enable the shift from a centralised government to a citizen-centric governance model.

While datasets are usually published in their raw form, and thus have little value on their own, public entities can leverage on other stakeholders, such as the private sector, community groups, and citizens, to innovate upon the published data and strive to achieve the utmost potential of open government data initiatives [41, 98, 157]. Benefits are plenty, including exploiting user participation (crowdsourcing) in order to enhance data quality through feedback [101]. Yet, active participation is not so simply achieved. While open data initiatives form the basis for citizen participation and collaboration, there is no guarantee that there is actually any resulting participation or collaboration [2, 41, 85, 137]. Moreover, as the authors of [92] and [98] point out, there is the need to bridge the gap between data providers and consumers by using *data intermediaries*. Thus, those who can make sense of the published data should interact with the software developers in such a way that the latter can develop innovative applications or services based on the published data. Even though this informal type of collaboration is facilitated by the existing technologies, it is not yet fully endorsed by public and governmental entities.

In order to achieve the impacts discussed in Section 3.4, the data provided in an open government initiative needs to be used by the involved stakeholders. Citizen participation is hence essential to promote the innovative potential of developers and other stakeholders. This is, however, easier said than done. A number of barriers hinder public participation, and mostly include challenges related to the cultural domain.

In [137], Solar et al. point out the need of an action plan for stimulating the consumption of open datasets between both the original data producers and the other consumers. User participation usually follows a "90-9-1 rule" [11] where:

[10] http://www.participatorybudgeting.org/about-participatory-budgeting/where-has-it-worked/ (Date accessed: 2 August 2016)

[11] http://www.nngroup.com/articles/participation-inequality/ (Date accessed: 2 August 2016)

- 90% of users are lurkers who follow by reading or observing but do not actively contribute;
- 9% of users contribute from time to time, but other priorities dominate their time;
- 1% of users participate a lot and account for most contributions.

'Lurking' tends to have a negative connotation, however lurking is also valuable in a democratic society where an informed citizen can take effective decisions [41]. In open government data initiatives, the aim is to achieve the highest number of active users as possible, keeping in mind that collaboration is not done for the sake of doing it, but to enable all stakeholders to participate in efficient and effective decisions.

Anonymity can be seen as an advantage in online participation. It allows anyone to be able to speak freely about his/her opinions and about any agendas they might be interested in, without the fear of being persecuted for them. This makes it easier for stakeholders to participate in efforts such as decision-making. Yet, anonymity also has its downside as it allows participants to contribute undesirable and useless information, as well as making participants more likely to insult or verbally attack others whilst hiding behind their anonymity [41]. Furthermore, a single user can use multiple online identities to manipulate the discussion in progress.

The participation of third parties in processes such as policy-making or decision-taking does not only potentially increase citizen satisfaction, but it also increases the potential of more innovative solutions or approaches to problems. Von Lucke and Große [81] term this participation as *open government collaboration*, which involves the collaboration of different entities during the implementation, monitoring, and evaluation of policies. Entities such as unions and political party associations were always traditionally included in the process of policy-making. Yet, these entities do not represent all members of society equally. By allowing all stakeholders to participate through eParticipation, a new collaboration approach that enables a many-to-many communication allows all individuals to participate in shaping the democracy they live in.

Albeit the benefits of open data outweigh the efforts required, it appears that there is a lack of public participation in open government data initiatives. In [157] the authors identify that the lack of research on the factors influencing external stakeholders' decision to participate and consume open data might be a factor in this problem. Chan [25], on the other hand, point out that governmental agencies do not have effective strategies to encourage participation from external stakeholders. Such public entities must come to the realisation that successful open data initiatives are based on the actual usage of the data rather than simply the creation of an open data portal. In [16], Bertot et al. carry out a case study with the aim of identifying how community data can be leveraged through public libraries. Amongst the authors' conclusions, they point out that stakeholders (i) not only need more data, but need it to be meaningful, (ii) need the identification of best practices for using the data, and (iii) request the collaboration of different stakeholder communities.

On the premise that the role of government agencies in open data initiatives is not only to publish the data, public agencies are starting to focus their efforts on motivating external stakeholders to use the published data. While there is no agreed-upon method to achieve public participation, there are a number of popular methods. Challenge competitions are a commonly-used approach [44], where the competition involves developing the best application, or finding an innovative use, based on the published data. Usually the winners are awarded a prize or recognition for their efforts. A disadvantage of such competitions is that most participants are usually novices rather than professionals [85]. This is of course somewhat reflected in the submitted entries, which tend to be amateurish. Moreover, the entries do not usually contribute to the development of sustainable services [44]. Professionals are usually deterred from participating in such competitions due to the minimal (if any) prize money. In any case, it is

evident that in such cases the governmental entity does not have any direct control on the output of the competition, and there is no assurance on the quality. For these reasons, challenge competitions are more suitable for just raising awareness about the open data initiative, and introducing stakeholders to public participation. Another approach towards encouraging participation are Calls for Collaboration, where companies are invited to submit proposals to create particular services. As opposed to challenge competitions, the governmental entity now has a say as to what will be developed as the output of the call, as well as the possibility to enforce the participants to meet specific requirements.

A number of publications in the literature attempt to identify the best method to achieve public participation. In [98] and [157], the authors propose their intentions in researching the best practices in increasing the consumption of open data. Kalampokis et al. [66] research the use of social media platforms in eParticipation and propose a two-phased approach for backing participatory decision-making, along with an architecture which supports its implementation. This approach is based on the integration of government and social data and attempts firstly to help the government identify public opinion and predict public reactions, and secondly to enable citizens and stakeholders to contribute to the decision-making process. With the similar aim of identifying what motivates stakeholders to participate and collaborate, in [25] Chan identifies a set of considerations for motivating stakeholders to innovate upon the published datasets.

3.5 Challenges

Even though there are numerous open government data initiatives, there still exist a number of setbacks which prevent them from reaching their full potential. There are also issues that hinder governments and other stakeholders from joining an open data initiative, or otherwise preventing data from being truly open. Through the evaluations led in the literature mentioned in the previous sections, and otherwise, we identified the most common challenges faced, and also propose possible solutions where appropriate. These challenges vary between technical, economic and financial, organisational, policy and legal, and cultural barriers.

3.5.1 What discourages entities from joining an open government data initiative?

Awareness - The concept of open data, while not new, might seem a daunting task for people unfamiliar with the term and what it involves [119, 149]. Public entities in the past would have only been concerned with delivering reports formatted to given templates. Recent requests to provide data in its raw format might not be understood clearly [27]. For this reason, the value and potential use of raw open data needs to be highlighted [159].

Motivation - The provision of raw data can be considered to be extra work without any purpose, especially to public entities such as those described above [149]. The value of the data generated during day-to-day administration needs to be pointed out. The re-use of open datasets can be a great motivator in portraying the unexpected use of the generated data, and can also help the data producers in understanding the true value of the data they create and publish [105].

Capacity - The use of open data should be targeted towards nobody in particular. Having said that, it should be available for the use, re-use, and distribution of all, whether machines or humans. Unfortunately, many entities are not so open-minded about the application of open data, and rather focus on the simple publishing of data rather than ensuring that it is of good quality in this aspect. Furthermore, public entities might focus on publishing data with no value, rather than other, more relevant, data [149, 161].

There is the urgent need for the application of standards and large-scale training in order to overcome these issues.

Budget Provision - Being a relatively new concept, there might not be any local budget allocation for open government data efforts [149]. Considering the required processes for publishing data are "extra" tasks, requiring effort, resources, and time, there is the new necessity of having a specific budget allocated for this purpose, otherwise there is the risk that open government initiatives are not given the priority they deserve. This is especially true if public entities do not grasp the true value of open data.

Technical Support - Most of the existing government data portals were not envisaged for large-scale open data publishing and consumption. Thus, these public entities now require technical support to update their websites or portals to enable their published data to achieve its highest re-use potential [27, 40, 119, 149, 161].

Institutionalisation - First and foremost, entities such as governments and large organisations try to manage new developments using established mechanisms of governance, however such mechanisms are not usually developed to adapt to changes [61]. This inflexibility directly hinders the new participation in an open government initiative. Moreover, being a relatively new effort, open data tasks are usually assigned to employees whose job was already predefined, with no institutional structure or public entity dedicated solely to this task [40, 119, 149, 161]. This issue results in no regular monitoring of the open data initiative performance. The establishment of open government initiative policies would help in this challenge by clearly defining required responsibilities.

3.5.2 What hinders an open government data initiative from reaching its full potential?

Data Formats - The whole point of opening and publishing data in portals is to enable its use, re-use, and re-distribution. Two of the Eight Open Government Data Principles, in fact, regard the format in which data is published, and state that such data should be made open to the public in a *machine processable* data format which is *non-proprietary*. Unfortunately, while this is a guideline, it is not legally required by many open government initiatives (which only require the publishing of data). Many governmental entities still publish data in a large variety of data formats which can also be proprietary. This has resulted in a number of data silos which appear to be available for use but which in reality require significant effort before being actually usable [32, 53, 64, 77, 82, 87, 131].

In an ideal world, in order to achieve economic growth, governmental entities (data publishers/providers) should take into account the requirements of the data end-users (data consumers) [158]. This should include the specific formats that are most convenient for the widest spectrum of consumers. W3C recommends the use of established open standards and tools, such as XML and RDF as a publishing format[12]. A feasible solution would then be to enforce data providers to publish their data in machine processable and non-proprietary formats through the open government initiatives in which they partake [137]. Thus the portal's 'success' would not only be evaluated on the amount of data published, but also on the usability of this data.

Data Ambiguity - While of course any machine-readable data format, such as CSV, is preferred over non-readable ones, such as PDF, more expressive data formats are generally preferred, simply because they are more descriptive of the actual data they represent. This decreases the risks of ambiguity and misinterpretations [87]. Consider the example of the concept of a year. While a calendar year would be the most common in our everyday lives, some financial agencies within the public sector might use a financial year to describe their data [77]. This leads to difficulties when attempting to find relationships

[12] http://www.w3.org/TR/gov-data/#formats (Date accessed: 2 August 2016)

between two datasets due to this difference in temporal representation. Semantic ambiguity therefore would require extra efforts in order to link and understand the data in question [27]. Similar to [32] we can thus conclude that although data is available in a machine readable format, such data is not really useful unless it is easily *understandable*; maybe by requiring just minimal background knowledge on the subject.

A simple enough solution for this issue is to publish data with descriptive titles, or otherwise provide a key to code names, if the latter are used [101]. This would help data consumers to clearly and easily understand what the data is about, and if it is actually useful for them. The use of RDF as a data format is also encouraged as it is a highly descriptive data format.

Data Discoverability - Publishing data and making it accessible qualifies as 'open data', however open data also needs to be discoverable. The discoverability of open data is bound to the quality of the metadata describing the data itself, which is not always complete or accurate [27, 71, 87, 116]. In addition, other factors lead to difficulties in finding useful data quickly [77]. For instance, some portals support only simple search functions which do not return only relevant data, but also related policies and documents such as research papers [2]. This may result in the user being overloaded with information and having to go through all the results to potentially identify the relevant datasets [161]. Moreover, most portals only allow users to simply download the available data, with no possibility of exploring it directly through the portal (for example through visualisation). These issues are particularly evident when the data consumers do not know the responsibilities of the government entity in question or the data structures that they implement, making it even harder to locate the relevant data they need. The fact that even most of the datasets are spread over a number of decentralised data sources further aggravates the problem [27, 38, 159].

A number of efforts in the literature focus on metrics which assess metadata quality. The authors of [116], for example, tackle the problem of metadata quality by applying five quality metrics, namely: completeness, weighted completeness, accuracy, richness of information, and accessibility, to three public government data repositories. This evaluation is carried out with the aim of measuring the metadata's efficiency, identifying low-quality metadata records, and also understanding the reasons behind the origin of the low quality. Evaluated metadata is then assigned a quality score which enables the uniform comparison of the metadata quality across different repositories or catalogues. Evaluated metadata can consequently be improved in order to achieve better searchability, and subsequently better discoverability.

Data Representation - The heterogeneity of the published datasets and their representation is quite an obvious setback for open government data initiatives. Data as varied as traffic, budget, geographical, and environmental data, etc., is published onto portals in a non-standardised manner, meaning that there exists a large heterogeneity in terms of semantics, standards, and most importantly in this case: schema. This leads to interoperability issues and challenges to aggregate existing metadata in a way that would be useful for data consumers [19, 53, 86, 87]. Additionally, such heterogeneous data would potentially even require to be mapped to a global schema. A further aspect to this issue is versioning. An ideal representation of a dataset would also capture how it evolves over time.

A number of efforts in the literature approach this challenge by proposing a generic schema. For example, in [86], Marienfeld et al. propose a minimal schema that is compatible with the predominant data catalogue vocabulary and software. The schema supports the description of datasets as well as documents and applications, and most importantly includes a list of resources containing pointers to the actual data, documents, or applications. In contrast, Maali et al. [82] propose a standardised interchange format which enables machine-readable representations of data catalogues. Thus, for catalogues differing widely in scope, terminology, structure, and metadata fields, this contribution acts as an interoperability format. With regard to versioning, a solution to the issue is the use of Named Graphs [24], where the metadata represents the temporal validity of the annotated RDF data. However, this solution is only

available with the use of RDF.

Overlapping Scope - Provenance, whilst not a challenge in itself, is also an issue. Provenance refers to details about the origins of data, or, in other words, who created or generated the data. The issue with provenance occurs when there is the assumption that data strictly travels in a vertical direction, for example from local, to regional, national, European and international level. There are numerous parallel entities which collect data, and then pass it on to another relevant entity. For example, budget datasets from a city can be published on the city's portal, but also transferred to the entity taking care of cities within a specific region. This results in an overlapping scope, where data may have duplicates, but also new or modified data [86]. Hence, provenance does not only regard the source of the data, but also how the data was modified or manipulated during the publishing process.

Here again, named graphs can be a solution to provenance issues, as different provenance metadata can be attached to datasets with varying provenance [131]. Using a somewhat different approach, the authors of [82] propose a standard interchange format which enables federated search over catalogues or portals with overlapping scope, providing a way around this problem. Using a more concrete approach, the W3C Provenance Incubator group[13], on the other hand, strives to provide a roadmap in the area of provenance for Semantic Web technologies.

Public Participation - A very relevant challenge to achieving the full potential of published datasets in portals is their use, or lack thereof. The increasing number of open data initiatives, where government entities are opening up their data, ideally would result in increased transparency, participation, and innovation [116]. Yet, as the authors of [41, 44, 45, 85, 90, 157, 164] point out, the full potential of consumer participation and collaboration for achieving innovation in government services has yet to be reached. Participation, as defined by [126], means the extent to which stakeholders can participate in the governance of an open government data portal, such as suggesting what data to publish, or rating datasets or features on the portal itself. Collaboration, an extension to participation, refers to features on a portal that enable cooperation and collaboration amongst different stakeholders.

Public access to government data also remains challenging due to the heterogeneous and dispersed nature of the data. The lack of consumers exploiting existing open data portals indicates that there is the need to understand what factors influence participation in open data, and the requirement to engage stakeholders in participating and collaborating. If the projected consumers of the data do not use it, then the objective of open government initiatives is futile. For a portal to be successful, consumers (including citizens, end users, and beneficiaries) must be made aware of the published data, and its relevance and usefulness [98]. Considered to be a core pillar of democratic society, the collaboration between a government and its citizens has the potential of enabling open data consumption, policy making, service delivery, and also political opinions and decisions [147]. This interaction would allow the government to provide more citizen-centred services and data.

In literature such as [157] the authors attempt to identify what influences the participation of stakeholders in consuming open data, with the aim of mitigating the barriers they face. Furthermore, Marie and Gandon [85] establish strategies to ensure that open data initiatives reach the desired participation rate. Similarly, in [136], Solar et al. tackle the question of what kind of services should governmental entities provide in order to increase stakeholder participation. In contrast, the authors of [16] focus on issues that smaller communities face when attempting to consume open data. The authors analyse these issues with the aim of enhancing public participation with the purpose of creating local data infrastructures. In [134] Sheffer et al. attempt to give structure to unstructured documents (such as PDF) and store them in repositories compliant with open government data principles, with the aim of providing stakeholders

[13] http://www.w3.org/2005/Incubator/prov/wiki/W3C_Provenance_Incubator_Group_Wiki (Date accessed: 2 August 2016)

with analysis functionality and unrestricted data access.

3.5.3 What hinders data from being truly open?

Conflicting Regulations - Whilst there is a lack of open government data policies, many open government data initiatives still belong to existing legal frameworks concerning freedom of information, re-use of public sector information, and the exchange of data between public entities. The issue lies in the unclear task of how such initiatives can interact, resulting in uncertainty on the possible use of the relevant data. This issue does not only concern data consumers, but also data producers who end up being sceptical of fully opening up their institutions' data, even if it is covered by a clear legal framework [119].

Privacy and Data Protection - There is a considerable conflict between open data and the aims of transparency and accountability, and data protection and the right to privacy [60, 91, 119, 159, 161]. Even though data is anonymised before publishing, the merging of different datasets can still possibly result in the discovery of data of a personal nature [165]. For example, if garbage collecting routes are published, along with the personnel timetable, a data consumer would be able to identify the location of a particular employee. This issue requires more research in order to come up with guidelines that can provide a solution to this conflict, however a plausible approach would be to employ access control mechanisms which regulate data access. Unfortunately this restricts the openness level of such data.

Copyright and Licensing - The licensing of published data is one of the Eight Open Government Data Principles. The first aspect of this issue is the incompatibility of licences [119]. Data publishers should provide efforts towards publishing their data in an open format, allowing the free and unrestricted use, re-use and distribution of data. Since there are no agreed-upon standards, this can result in a number of incompatible open licences. While they all, in different grades, allow the re-use of data, they might contain restrictions which prevent data with different licences from being merged for a specific use. The definition of clear data policies is a means to provide a solution to this challenge. The second aspect of this issue is copyright inconsistencies that arise from unclear dataset ownership resulting from data sharing, for example between public entities [27, 40, 161]. This hinders data from being published.

Competition - While open data can be considered as unfair competition for private entities, public entities might consider the commercial appropriation of public open data unfair [40, 119]. In the first case, consider companies who invested in creating their own data stores (e.g. database of streets and locations for navigation purposes). If the same data they created is made public through government open data initiatives, these companies will obviously deem it to be unfair competition as there is the possibility of new competitors who did not need to invest anything but could get the freely available open data. Thus, management mechanisms need to be applied in order to ensure that private companies do not suffer financial consequences due to opening up their data. On the other hand, public entities might be reluctant to publish their data openly due to not wanting data belonging to the public (and paid by taxes) to be used for commercial gain. A possible approach for the latter issue is to provide the data for a nominal fee. Yet, this limits the openness of the data in question.

Liability - This issue is limited to data publishers or providers. Public entities fear being held liable for damage caused by the use of the provided data, due to it being stale, incorrect, or wrongly interpreted [40, 119]. To cater for this fear, many public entities either do not publish their data or otherwise impose restrictions on its use, resulting in data which is not truly open. In the worst case, due to fears of data being used against the publishing entity, such data might not even be collected/generated any longer [161]. A possible solution for these issues is to enable social interaction with regard to the data in question. A community of stakeholders within the data platform where the data is published can aid data consumers to better interpret and exploit the published data.

CHAPTER 4

Publishing and Consuming Open Government Data

As essential parts of the data life cycle, publishing and consuming are vital for the existence of an open government data initiative. Without the existence of data and its re-use, an open data initiative is deemed to fail. In this chapter we therefore focus on these two processes with the aim of identifying their specific characteristics within open government data initiatives.

The act of publishing data is the very basis of open government data initiatives. Government and public entities are sharing data on the Internet at an astonishing pace. Yet, there is a lack of agreed-upon standards for data publishing [38], and as discussed in detail in Section 3.5, there are many challenges to be overcome in order for the published data to be exploited to its full potential. While not all challenges are directly related to publishing issues, tackling these issues at the root could prevent subsequent issues related to data consumption. For example, if data is published in a machine-readable format with good metadata descriptions, then usability issues will most probably be avoided when it is consumed.

The publishing of data enables it to be available for use by the public, in an attempt to achieve the main aim of open government data initiatives; namely to use, re-use, and distribute the published data. This is only achievable through the consumption of the data by stakeholders. Data consumption is possible through a number of means. The most direct example is to obtain a copy of the actual published data, generally with the aim of using it for a specific use-case. Certain portals might also provide exploration tools, where a data consumer can simply look through the published data. Other tools, such as analysis tools, enable a consumer to actually identify potential patterns in the published data. Usually analysis tools also provide for visualisations, which aid data consumers to view the data in a pictorial manner. An even more hands-on way of consuming the data is to create mashups, where different datasets are merged in order to create new knowledge using existing data.

4.1 Publishing Data

In this section we provide a classification of different data publishing approaches, and proceed to discuss guidelines and best practices for publishing data in any data publishing effort.

4.1.1 Data Publishing Approach Classification

There are countless methods towards publishing data. Following the contribution within [67], we here classify open government data publishing initiatives into two:

1. The **technological approach** - followed by the data publisher in the actual act of publishing data, i.e. making the data available on the Web. Publishing initiatives are classified within this approach depending on the variation of technologies implemented for publishing the data. These include:

   a) The format of the published data (proprietary, machine readable, descriptive);

   b) The access method (RESTful APIs, custom APIs, search interfaces);

   c) The use of Linked Open Data principles (HTTP, URIs, RDF); and

   d) The level of linkage to different datasets (Linked Open Data Cloud).

   As is evident, the above reflect most of the existing guidelines for publishing data, particularly the Five Star Scheme for Linked Open Data.

2. The **organisational approach** - followed by the data provider, i.e. the manner in which the data is provided to the data consumers. This second dimension for open government data publishing initiatives focuses on the *provision* of data, rather than the actual act of publishing. The authors of [67] identify two different methods of providing Linked Open Data, the epitome of an open government initiative, each with their own advantages and disadvantages.

   a) **Direct Data Provision** - Direct data provision involves a one-stop portal aggregating all processed and value-added data provided by a public entity. In this case, the data publisher is not necessarily the same as the data provider. In the case that the latter are two different entities, the maintainability is limited unless an effective data synchronisation process is in place. For example, if the original data from the public entity changes over time, this change must be reflected in the data provided on the data portal, otherwise the data provided here will be obsolete [38]. An advantage of having direct data provision, however, is the consumers' direct access to data through a single entry point.

   b) **Indirect Data Provision** - Data Catalogues are a good example of indirect data provision, where the data cannot be directly accessed through the catalogue. Catalogues contain links (metadata) to the actual data provided by the public entity. To access data, a consumer has to search for the relevant data through the catalogue, then follow the provided links to the public entity that provides the actual data. In contrast to direct data provision, indirect data provision has the advantage of being up to date and unique, since the actual data is provided by the data producers, and the catalogue simply provides links to it. On the other hand, processed and value-added data has to be performed by the data consumer, as it cannot be provided by the data catalogue.

4.1.2 Publishing Guidelines

In order to tackle the previously-mentioned issues in Section 3.5, and other publishing-related problems, a number of publications in literature, such as [54, 77, 136], propose guidelines for publishing data on the Web. The basis of most of these guidelines are the Eight Open Government Data Principles:

1. **Complete** - All available public data that is not subject to privacy, security or privilege limitations is made available.

2. **Primary** - Data is made available as it is available at the source, and not aggregated or modified.

3. **Timely** - Data is made available to the public as soon as possible after the actual data is created, in order to preserve the value of the data.

4. **Accessible** - Data is made available to all consumers possible, and with no limitations on its use.
5. **Machine Processable** - Data is published in a structured manner, to allow automated processing.
6. **Non-Discriminatory** - Data is available for all to use, without requiring any registration.
7. **Non-Proprietary** - Data is published in a format which is not controlled exclusively by a single entity.
8. **Licence-Free** - Other than allowing for reasonable privacy, security and privilege restrictions, data is not subject to any limitations on its use due to copyright, patent, trademark or trade secret regulations.

The above principles provide a roadmap for the data publisher and help result in good open government data with the best potential for being consumed by the stakeholders. Further to these principles, the Five Star Scheme for Linked Open Data, listed below, provides a more technical guide towards publishing Linked Open Data:

1. Available on the Web in any format but with an open licence (Open Data);
2. Available as machine-readable structured data (e.g. Microsoft Excel table instead of image scan of a table);
3. Available as machine-readable structured data in a non-proprietary format (e.g. CSV instead of Microsoft Excel);
4. All of the above as well as using open standards from W3C (RDF and SPARQL) to identify things;
5. All of the above as well as linking the published data to other existing data to provide context.

In order to provide official guidelines, the W3C eGov Interest Group has also developed the following set of steps for publishing open government data[1], which emphasise standards and methodologies to encourage the publishing of government data, with the aim of enabling easier use by the public:

1. **Identify** - The use of permanent, patterned and/or discoverable URI/URLs enables processes and people to find and consume the data more easily.
2. **Document** - Documentation helps the data to be more understandable and less ambiguous, as well as enabling easier data discovery. The use of formats such as XML/RDF would be self-documenting.
3. **Link** - Linked data contains links to other data and documentation, providing context.
4. **Preserve** - The use of versioning of datasets enables data consumers to cite and link to present and past versions, where new and upgraded datasets can refer back to original datasets. Versioning also allows the documentation of changes between versions.
5. **Expose interfaces** - To make it easier for published data to be discovered and explored, published data should be both human-readable and machine-readable. Preferably, data should be published separate from the interface, and external parties should have direct access to raw data. This enables them to build their own interfaces if needed.

[1] http://www.w3.org/TR/gov-data/ (Date accessed: 2 August 2016)

6. **Create standard names/URIs for all government objects** - The use of a unique identifier for each object is as important as having information about the object itself. This aids in discoverability, improves metadata, and ensures authenticity.

Along with the above, the W3C eGov Interest Group also discusses the importance of *choosing what data to publish*, the *right format* to publish it in, and the *restrictions on its use*. Data which is to be shared with the public should be published in compliance with applicable laws and regulations, and only after addressing issues of security and privacy. Such data is usually already available in other formats, and may already have been shared with the public in other ways. The best format to publish this data is in its raw form serialised as XML and RDF, to allow for easy manipulation. The use of established open standards is also recommended. Finally, the published data should have clear documentation on any legal or regulatory restrictions on the use of that data.

Liu et al. [77] present some recommendations for data publishing and analysis based on a survey on the sustainability related datasets published by the Australian government, with the aim of identifying underlying opportunities and issues. While not entirely reflecting the above-mentioned guidelines, the proposed recommendations complement the essential aspects. The authors tackle commonalities amongst data published by different public entities, the ideal formats for publishing data as Linked Data, its discoverability, and its re-usability.

Similarly, the authors of [119] identify common issues and challenges to the accessibility and re-usability aspects of public sector information. De Rosnay and Janssen point out that such obstacles can be of legal, institutional, technical or cognitive nature. They proceed by providing common solutions that can be implemented to overcome these issues.

In [136], Solar et al. propose a maturity model for open data, with the aim of assessing the commitment and capabilities of public agencies in pursuing the principles and practices of open data. The authors extend the discussed guidelines and principles by considering other aspects towards publishing data, including an Establishment and Legal Perspective, a Technological Perspective, and finally a Citizen and Entrepreneurial Perspective.

Another maturity model was defined in [80]. Here Lourenço and Serra aim towards identifying essential contextual aspects which affect the way data is published by public entities on their portals. The latter aspects are then organised into an online transparency for an accountability maturity model, which has the purpose of assessing the level of advancement of a governing region. In other words, researchers requiring to assess an entity should start by analysing the context using the proposed maturity model, and then proceed to define the assessment model depending on the identified maturity level.

4.1.3 Publishing Tools and Standards

While there exist a huge number of government data portals that enable data producers to publish their data, there are not many tools aiding data publishers in this task. Yet, efforts are currently being focused on providing portals and other open government data initiatives which allow stakeholders to publish (and consume) datasets without requiring background knowledge on the open data life cycle. An example of such efforts is the LinDA project[2]. A contribution within this project enables a stakeholder to publish data in any format, which is then converted to RDF to enable easy linking with other open datasets.

In [55], Hofman and Rajagopal propose a technical framework for data sharing between data providers and consumers, based on an analysis of a number of data platforms. They aim to identify, from the relevant literature, the required functionality for data sharing, considering challenges such as different published formats, data ambiguity, and privacy issues.

[2] http://linda-project.eu/ (Date accessed: 2 August 2016)

Meijer et al. [91] present two case studies involving two different public sector entities, with the aim of demonstrating the use of pre-commitment to resolve conflicts during a data request procedure. Pre-commitment involves applying restrictions on the type and content of the data that is available for request, ensuring the data conforms to the legal requirements (e.g. removing privacy sensitive data), and deciding on whether to open the data publicly or restrict its access to specific user groups.

The authors of [2] propose a second generation platform that offers both the basic functionality of a government data portal, but also additional functionality (based on Web 2.0 technologies) aiming to stimulate and aid value generation from open government data. This additional functionality includes the capability of performing a number of processing techniques, information and knowledge exchange, and collaboration between stakeholders.

In [64], Jiříček and Di Massimo introduce the European Open Government Data Initiative, which is a free, open-source, cloud-based collection of datasets that public entities can exploit. In this case, public data can be uploaded and stored into the Microsoft Cloud through the Windows Azure Platform and environment. This tool is aimed at experts, and allows developers to use a variety of programming languages. This initiative strives to keep in line with the open government data principles and thus enables data to be openly published in a re-usable format, enables stakeholders to develop new applications based on the published data, allows developers to use the free and customisable source code, and has the aim of enhancing transparency through increased visibility of a governments' services.

In contrast to the above, Maali et al. [82] propose a standardised interchange format, the *dcat* vocabulary, for machine-readable representations of government data catalogues, with the aim of bringing all published datasets into the Web of Linked Data, resulting in higher interoperability. The use of this interchange format results in a number of advantages:

1. The embedding of machine-readable metadata in Web pages increases discoverability;
2. The decentralised publishing by individual agencies could be aggregated into national or supra-national (e.g. EU-wide) catalogues;
3. Catalogues with overlapping scope (e.g. Bonn, Germany and EU) can be searched in a federated manner;
4. One-click download and installation of data packages is available for application developers;
5. Priority is given towards archiving and digital preservation of valuable government datasets through the use of manifest files with accurate metadata; and
6. Software tools and applications, such as improved search and data visualisation interfaces, can be built to work with multiple, or even across, catalogues.

The *dcat* vocabulary has since been proposed as a W3C recommendation by the Government Linked Data Working Group[3].

4.2 Consuming Data

The provision of data does not only enable stakeholders (whether individuals, businesses, NGOs, or otherwise) to scrutinise the published data, but also to stimulates them to create, deliver, and use new services that are coupled with the published data [41]. Services can be as simple as offering exploration

[3] http://www.w3.org/TR/vocab-dcat/ (Date accessed: 2 August 2016)

of the published datasets, but may also include visualisation and data discovery services such as data mining and comparative analysis. The latter enables stakeholders to explore the data and identify patterns. Furthermore, if the published data is linked with other data on the Web, the services can be enhanced with mashups, and further increase the knowledge that can potentially be discovered through the available data. This opportunity can then result in an improvement in e-government service provision, increasing work opportunities and finally also contribute to economic growth [67, 85].

Unfortunately, few open government data portals provide consumption functionalities other than simple data downloads [2]. In an attempt to enhance the consumption experience, Janev et al. [57] explore the challenges and issues related to the integration and analysis of open data. Amongst other challenges the authors identified:

- The lack of standard procedures for querying government portals;
- The low quality of metadata;
- Low reliability and non-completeness of public datasets; and
- The heterogeneity of formats used to publish open data.

They proceed to propose a Linked Open Data approach to modelling, merging and analysing specific data; namely spatio-temporal and statistical data.

Jetzek et al. [63] tackle the question of how open data can encourage the creation of sustainable value. They discuss that new methods of generating value can be brought about by the sharing and re-use of open data. The authors proceed to propose a model describing how various processes within an open data system can generate sustainable value, based on a number of contextual factors that provide stakeholders with the motivation, the opportunity, and the ability to create it.

4.3 Data Quality

As defined in Section 2.2, data quality has no agreed-upon definition, and apart from being cross-disciplinary, it is also subjective [101]. Also, the publishing of data on portals does not guarantee that it is of good or high quality [35, 116]. For these reasons, we hereby do not define how published data can be of good quality, but we discuss the different aspects which influence the quality of the data, whether positively or negatively, and ultimately affect the (re-)use of the published data.

Ochoa and Duval [100] propose a set of metrics to identify metadata quality, based on parameters used for human reviewing. The authors of [116] build upon these metrics, adapting them for assessing the quality of the actual data, rather than the metadata. Similarly, in [71, 79], the authors discuss a number of quality dimensions, as found in the majority of related literature. We here establish the following criteria which are considered by most efforts in the literature for calculating data quality.

Usability - This is the most "generic" quality criterion. By usability we mean *how easily can the published data be used*. It is the most generic as it depends on other quality dimensions whether the published data is usable or otherwise. For example, it is directly related to what degree the data is accessible, open, interoperable, complete, and discoverable [77, 88]. The more the published data is usable, the more potential data consumers are encouraged to re-use and exploit the data.

Accuracy - By accuracy we mean *the extent to which a data/metadata record correctly describes the respective information* [71, 87, 116]. With respect to metadata, this quality dimension directly affects the discoverability of datasets, as good quality metadata enables the dataset to be easily discovered by data consumers.

Completeness - This quality dimension deals with *the number of completed fields in a data/metadata record* [100, 116, 136]. Thus, a record is considered complete only when the record contains all the information required to have the ideal representation of the described data. The completeness of the metadata, like accuracy, also directly affects the discoverability of datasets.

Consistency - The consistency of record fields depends on whether they *follow a consistent syntactical format, without contradiction or discrepancy* within the entire catalogue of metadata [71, 82]. Apart from the syntactical format, a field is considered to be consistent if the respective values are selected from a fixed set of options. An example of inconsistency is if within two records the use of "U.S" and "United States" is interchangeable. Another example is the representation of dates, where the date, month and year follow an arbitrary order.

Timeliness - By this quality dimension we mean *the extent to which the data or metadata is up to date*. As pointed out in Section 4.1, the organisational approach affects the timeliness of the published data, which depends on whether the data is directly or indirectly provided by the data publisher.

Accessibility - As identified by the authors of [100], the accessibility quality dimension has two measures. The *cognitive accessibility* defines *how easy it is for a data consumer to understand the published information*. Several aspects of the data affect the cognitive accessibility, such as the ambiguity of the data, discussed in Section 3.5. The second measure is the *psychological or logical accessibility*, which can be defined as *the ease with which the relevant dataset is discovered* through a data catalogue or repository. This quality dimension is affected by the format in which the data is published, the search tool used, and the discoverability of the dataset [82].

Openness - The openness of a dataset directly influences the use, re-use, and re-distribution of data. Tim Berners-Lee's Five Star Scheme for Linked Open Data (Figure 4.1) can be seen as a mix of the accessibility and usability quality dimensions. As the authors of [71] point out, open data can be technically defined to be open if it is *available as a complete set in an open, machine readable format, at a reasonable price which is not more than the cost of reproduction*.

The authors of [71], Kučera et al., identify two types of strategies for improving data quality; namely *data-driven* and *process driven*. The first involves directly modifying the values of data, such as correcting invalid data values or normalising data. The second involves the redesign of the data creation and

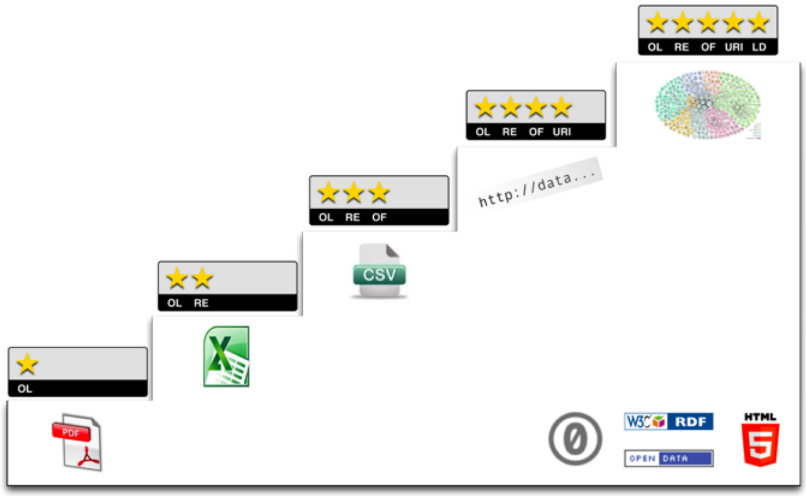

Figure 4.1: Five Star Scheme for Linked Open Data (Source: 5stardata.info).

modification processes in order to identify and correct the cause of quality issues, such as implementing a data validation step in the data acquisition process.

Efforts in publications such as [33, 71, 116] take a number of quality dimensions and implement them, with the aim of assessing the quality of published data. Debattista et al. [33] evaluate and assess the datasets' quality in such a way that consumers can then identify the ideal quality for the intended use, attaching the results of the evaluation to the actual dataset graph. In [71], Kučera et al. focus on the quality of catalogue records within initiatives in the Czech Republic. They proceed to propose some techniques and tools to improve the quality of the data catalogue records. Similarly, Reiche and Höfig [116] propose quality assessment metrics and implement them in three public government data repositories.

CHAPTER 5

Budget Data: A Use Case and an Assessment Model

Budgetary or financial data is collected and maintained by all governments and public administrations as part of their day-to-day administration. Whilst mostly related to transparency efforts by governments, budget data can also improve democratic participation, allow comparative analysis of governments, and boost data-driven businesses. The importance of publishing government budgetary data can therefore be summarised in four key elements as follows:

1. **Transparency** - Opening budget data unveils public funds' management. This increases accountability and can augment citizens' trust in public administration, whilst having a potential of uncovering hidden transactions and thus preventing corruption. An important factor which can stimulate corruption is the fact that funding goes through the hands of public officials without further scrutiny. In European Union Member States, this is particularly evident within public procurement, which is prone to corruption owing to deficient control mechanisms [43].

2. **Participation** - Opaque regimes may compel citizens to engage against the government. A transparent public administration, on the contrary, can stimulate social participation in community enhancement. Open budget initiatives can not only enable meaningful civil societal scrutiny of transnational financial flows, but they can also provide platforms for stakeholders to develop benchmarks that in turn create pressure on public authorities to provide data in a timely, comparable, re-useable and well-structured manner. These platforms can also involve local citizens in the budget planning and auditing phases, by allowing them to interact with the process, providing opinions and suggestions on setting budget priorities, and providing feedback on the published transactions. A virtuous circle can be created, in which both public officials and civil society will realise the value of data and analysis tools, in a collaborative environment open to contributions and engagement.

3. **Comparative Analysis** - Well organised budget data facilitates researchers and policy makers to compare spending strategies between cities, states, and countries, and also among different administration levels. Visualisation, analytics, and exploration tools can offer different stakeholders an opportunity to scrutinise and interpret financial data related to a region of interest. It also allows comparing allocations and transactions between multiple regions, to visualise detected trends and budget projections, and to investigate anomalies and activities that have been flagged as suspicious. A necessary condition for comparative analysis is the compatibility and consistency of data from different data sources.

4. **Business Value** - Publishing budget data can stimulate the creation, delivery, and use of new services on a variety of devices that utilise new web technologies, coupled with open public data. In fact, Manyika et al. [83] estimate that open data can help unlock between 3 to 5 trillion U.S. Dollars in economic value annually. Budget data can also generate value by empowering journalists when they report on spending items, and accurate information on public funds' usage may enable content producers to create better articles.

As a sub-set of the broader open government data, budget data is also being published in open government data initiatives, and unfortunately also suffers from similar challenges and issues. A core issue is the large number of diverse data structures that make the comparison and aggregate analysis of transnational financial flows practically impossible. The tools to present, search, download and visualise this financial data are also nearly as diverse as the number of existing portals. This heterogeneity may even prevent an analysis of the quality of the data for the same funds administered by different funding authorities [145]. Moreover, most of the budget publishing efforts results in simple data catalogues, fragmented and dispersed, because they do not share standards and methodologies [145]. This absence of standards can lead to data misuse [161], or even to results opposed to the initial aims [48].

In this chapter we propose a *structured analysis framework* to analyse open government initiatives that publish budget data. Our aim is to identify problems generated by the lack of standards and help policy makers to understand the importance of various aspects of publishing budget data. We also envision the framework as a tool to design more adequate budget publishing systems. Together with other ongoing initiatives [102, 152], we believe that the development of a solid standard can help governments to make their budget data more usable, and thus enable citizen participation in the democratic process.

5.1 Terminology

Although budget-specific terms are already defined in the Economics or Accountancy fields[1], some of the basic concepts used in this chapter still miss a specific definition in the context of open data. We here provide a very short description of the most relevant concepts as used in this chapter.

Budget - The description of the amount of money planned to be spent in a specified time period. Budget descriptions can refer to several levels of specificity, from general (total amount to be spent) to specific (amount by area, or category). There are different types of budget, such as proposed, planned, and certified, which is presented after the budget term.

Spending - Also known as expenditure, spending refers to the amount of money actually spent by the public administration. It can also be seen as the realisation of the budget. There also exist different types of spending, such as planned (according to the budget), authorised (payment order) and executed (money transferred from government to the recipient).

Revenue - The amount of money received by a government administration. Revenues can have several types of origins, such as taxes (revenue, commercialisation), service fees (transportation), royalties (oil and mine exploration), concessions (roads), or financial operations. Predicted revenues, used to specify the budget, may differ from the actual revenues.

Open Budget Data - Any electronic file or set of files containing structured data related to *Budget*, *Revenue* or *Spending*. In order to be strictly designated as "open", data should follow the Eight Open Government Data Principles, however open data can have varying degrees of openness.

Open Budget Initiative - This refers to any effort that aims to publish budget data. This can take shape as a portal or application which publishes open budget data and allows stakeholders to access it.

[1]The http://financial-dictionary.thefreedictionary.com/ compiles several of these definitions.

5.2 Related Work

To the best of our knowledge, there are no works in literature that propose a reproducible framework for the evaluation of open budget initiatives, however, a number of publications propose frameworks, impact measures, and comparison criteria on the general open data domain. Certain publications, such as [148] and [160], aim to compare e-government and open data policies. In [144] Ubaldi proposes a framework to evaluate open government data initiatives whilst also contributing to the discussion on the development of impact metrics. In [47], Granickas provides a theoretical background to analyse the impact of open government data. The author divides impacts into three, namely economic, political, and social. For each of the latter impacts, the author discusses possible implementation issues and impact metrics. Recently, a working group was created to develop methods for assessing open data. In their first report [22], a draft of a framework is proposed. Unfortunately, none of the above cited works focus on budget data.

Even though structured analysis and comparison of open budget initiatives have not received much attention from the literature as yet, three works must be highlighted. The Open Budget Survey [56] is a research project that, every two years, analyses the current situation of budget transparency, participation, and oversight, in a number of countries worldwide. It generates the Open Budget Index[2], which is updated monthly, and is based on the publication of eight key budget documents. Despite being a very useful comparison tool, this methodology does not evaluate information systems used to publish budget data, which are the means by which the information reaches the society. In [13], an evaluation and comparison between thirty Brazilian fiscal transparency portals on several administration levels is presented. The Eight Open Government Data Principles were used by experts to evaluate various portals. Despite being a well defined and widely accepted model, these principles are quite general, and do not refer to specific characteristics of budget data. Moreover, they mostly focus on the publishers' perspective. Finally, the authors of [117] provide a report describing a thorough review of evidences of impacts of fiscal openness. Whilst recognising that there is a literature gap on testing causal effects, the most rigorous studies found a relation between open budget initiatives and the desired outcomes.

5.3 Structured Analysis Model

With the aim of developing a model that enables stakeholders to analyse open budget initiatives, we used a research approach similar to the one used by Zuiderwijk in [160]. After analysing the related literature and observing some randomly selected open budget initiatives, we used inductive reasoning to build a first approach of the model. The model is a set of *Dimensions*, which represent different *Aspects* to be assessed in an open budget initiative. Dimensions are grouped in *Parts*, according to their general functions. The same basis is used to define the *Use Perspectives*, as shown below, which represent different ways of using budget data. From the use perspectives we then extracted the related requirements.

1. **Transparency Use Perspective** - Stakeholders such as journalists, software developers, and NGOs use budget data to audit the government and to translate data into more accessible formats for the society. For this use case, detailed data (i.e. data at the transaction level), consistent classification levels, and machine readable formats are some important requirements. Discussion and feedback on the provided data are also requirements in this case, for example, for suggesting different priorities for budgeting, or discussing a particular transaction. Both citizens and public administrations can benefit from this feature since the citizens (or other stakeholders) can show their perspectives

[2] http://www.obstracker.org/ (Date accessed: 2 August 2016)

and the public administration entity would check the current priorities to see if they need to be amended.

2. **Participation Use Perspective** - For the last two decades, cities from all over the world have been implementing participatory budgeting experiences with different systems and procedures. Research shows how developing and promoting participatory budgeting digital solutions can increase civic engagement up to seven times [135]. In Europe, digital solutions to promote citizen engagement in budget creation include, for example, sending proposals by email, participating in online forums and discussion, subscription to SMS updates and video streaming [106]. A comparison between offline, online and hybrid models of participatory budgeting showed that the use of ICTs added public value in process, enhancing efficacy, effectiveness, efficiency and transparency [94]. As an example of a participation use case, we here use a participatory budgeting case, where participatory budgeting stakeholders must have access to accurate and easily understandable budget data. Through this perspective, design, usability, and human readable formats are the most important requirements. Hierarchically aggregated categories also play an important role.

3. **Policy Making Use Perspective** - If adequately published, budget data can be used to compare the way each government manages public funds. Researchers and policy makers should be able to compare the budgets and spending data between (i) different public administrations (e.g. Cologne v.s. Berlin); or (ii) different periods (e.g. year 2013 v.s. year 2014), and thus relate spending strategies to political, economical and social outcomes. Comparing spending profiles among governments requires the use of common classifications, vocabularies, and ontologies, and the possibility of linking data with other databases. In order to enable the integration of the corresponding budget data on the different public administration contexts, a semantic data model for budgets and spending has to be defined. The use of a standard format facilitates the comparison of data from different municipalities or regions. More importantly, it allows all the stakeholders involved or interested in budget planning or spending to manipulate data using the same tools and methods, thus supporting financial transparency in public budgeting and spending. This may allow the creation of visualisations and comparative data analyses for the discovery of trends. Stakeholders will therefore be able to view and compare allocated budgets and transactions, and give feedback on each item. This feedback can then be shared through social media and also be directly exploited by governments and public administrations to achieve better budget management. The latter two stakeholders will thus benefit from receiving targeted suggestions, comparative benchmarks and scenarios.

In order to verify the fitness of the dimensions and the coverage of the use perspectives we propose, we used deductive reasoning and applied both the model and use perspectives to other open budget initiatives. Missing items were added to the model and to the use perspectives, and the feedback loop was run until no significant changes were found. Finally, use perspectives were checked against the model, in order to verify the correspondence between model dimensions and use perspectives.

The resulting model we propose is depicted in Figure 5.1. The main objective for building this model is the evident lack of existing mechanisms to assess different strategies for publishing budget data. Our aim is hence to organise open budget initiatives in order to assess their fitness for specific use perspectives. The model consists of the three parts within an open budget initiative:

1. **General Aspects** - the overall characterisation of the initiative;

2. **Data Publishing** - the aspects specific to the data publishing process; and

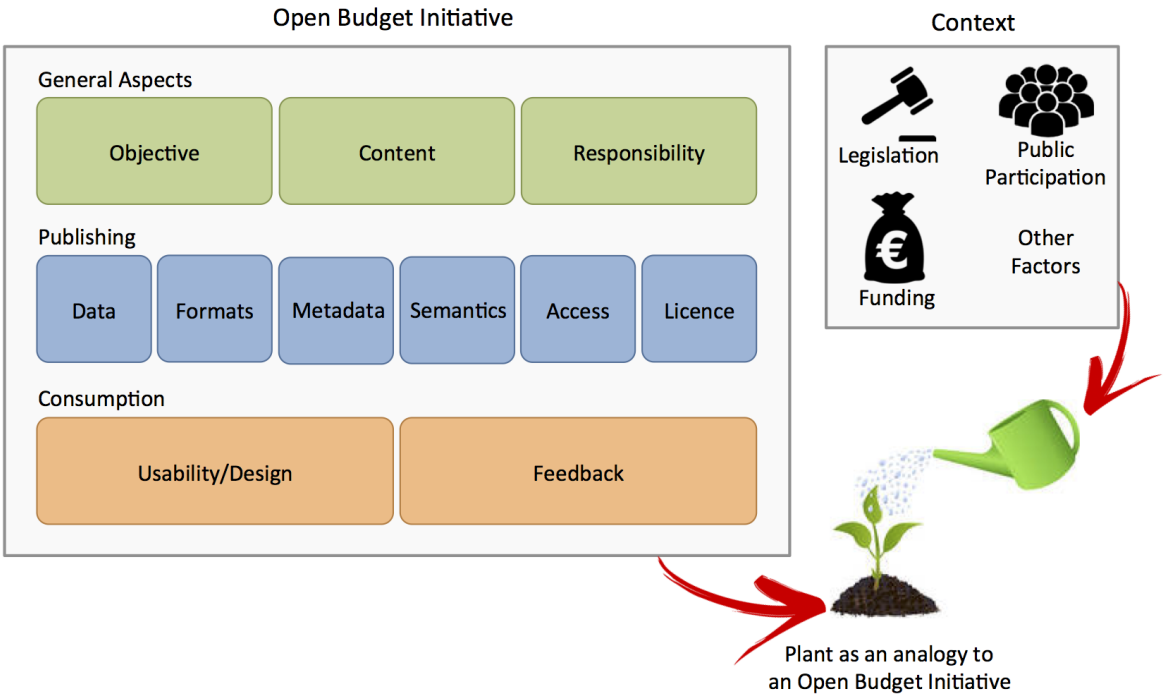

Figure 5.1: Model to analyse open budget initiatives.

3. **Data Consumption** - the aspects specific to the data consumption process.

Due to the nature of an open budget initiative, there is a strong coupling between these parts. The way data is published directly affects its consumption. With the same reasoning, the feedback generated by users (should) affect data publishing.

In Figure 5.1 we also represent the context of a budget initiative, which portrays the external aspects related to the initiative in question. These aspects are usually of varying natures, including policy/legal, economic/financial, organisational, and cultural (see Table 5.3 for a summary). Examples include public participation, budget provision, licences, etc. as discussed in Section 3.5. Such aspects have the capacity to influence the success, or otherwise, of a budget initiative. Let us consider the analogy of a plant, as shown in Figure 5.1. If a plant is watered then it will grow. Similarly, if for instance good policies and laws are established which encourage and enable an open budget initiative, then the latter will have a better chance of success. The context particularly impacts the general aspects, but also influences the other parts, as the context represents the environment in which the open budget initiative is involved. Whilst we recognise that the context is a key element for the success of an open budget initiative, we will not consider it in the scope of this chapter because its complexity would make this first approach towards an objective model unfeasible. Thus, we here focus on aspects within the initiative itself; the general aspects directly related to the initiative, and on the issues related to the publishing and consuming of data.

The characterisation of an an open budget initiative is the first step in order to be able to assess quality. In this chapter we consider quality to be dependent on the conformance to requirements, which in our case are extracted from the above-mentioned use perspectives. We hence proceed to describe our characterisation approach. Each dimension in the different parts of the model will be assessed through a specific *Characterisation Attribute*, defined as follows:

Definition 3: *Characterisation Attributes* are features of open budget initiatives that: (i) are objectively assessable; (ii) expect qualitative values; and (iii) have direct impact on the realisation of use perspectives.

After identifying the three use perspectives and our characterisation approach, we here define the various dimensions in the different parts of the model we propose. For each dimension we explain why it is relevant, the requirements as extracted from the use perspectives, and the characterisation attributes (summarised in Table 5.1) that are used to assess the fitness of the dimension in question in regard to use perspective.

5.3.1 General Aspects

- **Objective** - Motivations to publish budget data, or open data in general, can be very diverse. In the introduction section of this chapter we listed four common reasons for publishing budget data; transparency, participation, comparative analysis, and generating business value. Defining the target audience is also important, since different user profiles require different approaches. For example, in the Transparency use perspective detailed data is desirable in machine readable formats, while for the Participation use perspective, human readable charts and tables are most suitable. A SPARQL endpoint would probably be a better fit for the needs of stakeholders in a Policy Making use perspective.

  Characterisation attributes: We define as characterisation attributes: (i) whether an initiative clearly states its objective (CA1), and (ii) whether the intended audience is explicitly defined (CA2).

- **Content** - Open budget initiatives are very heterogeneous regarding to the presented content. Data can refer to several administration levels (local, regional, national), and also to the different power instances (Executive, Legislative or Judiciary), according to the political system of each country.

  Characterisation attributes: The first important distinction we require to make is whether the initiative is exclusively for publishing budget data, or if it contains other kind of information (CA3). We also make the distinction between primary sources of data and secondary data, i.e. applications working over data published by other initiatives (CA4). Finally, we assess the scope of the initiative (CA5), classifying it into local, regional, national or transnational range. When an initiative allows publishers to display different datasets having different scopes, we consider the scope to be generic. For this dimension we also identify initiatives focused only on the legislative power.

- **Responsibility** - Governments, as suppliers of primary data, may define specific sectors to be responsible for publishing budget data. For example, in the US, responsibility is under the General Services Administration, while in UK there is a Transparency and Open Data team under the Cabinet Office. Since budget data is sensitive, mistakes can lead to severe consequences, and hence the publishing of budget data implies a great responsibility to the entities in charge. Civil society organisations also play an important role by building applications over primary data, especially regarding the Participation use perspective. In this case, responsibility lies in making the context clear and simplifying as much as possible for data to be understood, but as little as possible to avoid misinterpretations.

  Characterisation attributes: We define, as a characterisation attribute, the distinction between data provided by governments and by society (CA6). We also consider the possibility of a joint government/society partnership.

Model Part	Dimension	Characterisation Attribute	Possible Values
General	Objective	CA1: Is the objective clearly stated?	Yes/No
		CA2: Is the intended audience defined?	Yes/No
	Content	CA3: Is data exclusively on budget?	Yes/No
		CA4: What is the source of data?	Primary Source/Secondary Source
		CA5: What is the scope covered by the strategy?	Local/Regional/National/Transnational/Generic, Legislative
	Responsibility	CA6: Who is responsible for the strategy?	Government/Society/Both
Publishing	Data	CA7: What categories are available?	Budget/Spending/Revenues/Generic
		CA8: What measures are available?	Time/Place/Payer/Payee/Category/Generic
		CA9: What is the finest data granularity?	Transaction/Aggregate/Generic
	Formats	CA10: Which formats are available?	Five Stars of Open Data
	Metadata	CA11: Is metadata available?	Yes/No
	Semantics	CA12: Is any ontology or vocabulary used?	Yes/No
	Access	CA13: How is data made available?	Catalogue/Raw Data/Querying System/Stories/Infographics
	License	CA14: Is the data licensed?	Yes/No
Consumption	Usability	CA15: What software tool is used?	CKAN/OpenSpending/Other
	Feedback	CA16: Is it possible to give feedback over data?	Comments/Data Request/Issue Reporting

Table 5.1: Model Parts, Dimensions and Characterisation Attributes defined to characterise an open budget initiative.

5.3.2 Publishing

- **Data** - This dimension focuses on specific aspects of the data content, and determines what kind of information is possible to be extracted from an open budget initiatives.

  Characterisation attributes: In order to determine the data content, we define three characterisation attributes: (i) *Category* - the types of represented quantities, which can be budget, spending and/or revenue (CA7); (ii) *Measures* - how the categories are quantified, which can be time, space and/or other categories (CA8); and (iii) *Granularity* - the finest level of detail available: transaction, aggregate, or generic for when the options are not predefined and several datasets in the same initiative present different settings (CA9).

- **Formats** - When data is offered for download, the format in which it is encoded plays a very important role. Especially for the Transparency use perspective, data in machine readable formats is crucial. For the Policy Making use perspective, the unique identification of entities and relations is also very important. The semantic resources generated by open budget initiatives can be instantly ready for re-use when resources follow Linked Open Data principles and guidelines [51]. The resulting data, usually available in a standard interoperable format such as RDF, is then fully compliant with the statement for best practices given by the G8 Science Ministers [120]: "Data should be easily discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards".

  Characterisation attributes: Here, we adopt the well-established Five Star Scheme for Linked Open Data as a characterisation attribute (CA10).

- **Metadata** - Adequate metadata is fundamental for providing complementary information about the context of the data in question, as well as for enabling the data to be discoverable by search engines. Information such as dataset author, published date and last update, formats, and licence, are usually the basic metadata. Another useful class of metadata is provenance. Provenance metadata describes the transformations applied to the dataset, and can also explain the process through which each data item was generated.

  Characterisation attributes: As a characterisation attribute, we here check for the existence of metadata in an open budget initiative (CA11).

- **Semantics** - In order to be correctly interpreted, data must be contextualised in order to avoid problems that emerge from the ambiguity in the used terminology or lack of agreement. Without post-hoc unification the data may be difficult to understand, as the users may need to familiarise themselves with different terminologies for each dataset. Having a single data format may solve *structural heterogeneity*, at the cost of introducing yet another format bridging the others. A more complex issue refers to *semantic heterogeneity*, which may be addressed by simpler solutions based on vocabularies, or more comprehensive approaches based on ontologies. Arguably, the most important benefit of Linked Data is the improvement of data interpretation. The key to such improvement comes from the recognition that measures in public budget and spending data are relative. If there is no way to compare them and put them into context, it is difficult to make sense of the data. Putting money into a wider context, such as how it was spent, helps to perform meaningful analyses and find comprehensible *stories* in data. The context may be provided by linked datasets, such as population statistics. For the Policy Making use perspective, it is vital to follow semantic standards, and even though budget data tends to be very heterogeneous, especially between different countries, some common points can be found.

Characterisation attributes: We define the *Semantics* characterisation attribute as a Boolean value that indicates the presence of standardised vocabularies or ontologies (CA12) in the open budget initiative.

- **Access** - The simplest way of publishing budget information is by offering data for download, which can be done in several formats. However, in the Participation use perspective, interactive charts, maps, or infographics are more useful than downloadable datasets, even if this might not be considered open data in the strict sense. Thus, this dimension aims to check the adequacy between the desired audience and the way data is offered.

  Characterisation attributes: Data Access is a characterisation attribute (CA13) which can be assigned as:

  - Downloadable data;
  - Data and metadata catalogue;
  - Exploration through tables;
  - Visualisation through charts, maps, comparison; and/or
  - Stories.

- **Licence** - Licensing is a fundamental issue for data (re-)use. In the Transparency use perspective some kinds of use can be hindered by the absence of adequate licensing. Currently, three types of general licences for open data are available[3]: Public Domain Dedication and License (PDDL), Attribution License (ODC-By), and Open Database License (ODC-ODbL). Some governments developed their own open data licences, for example, Germany[4] and the United Kingdom[5].

  Characterisation attributes: We define a Boolean characterisation attribute to describe the existence of a licence (CA14) on data published by an open budget initiative.

5.3.3 Consumption

- **Usability/Design** - A good set of visualisations, which are self-explanatory and easy to understand, can certainly improve the usage of an open budget initiative. Interactive visualisations and infographics can also enable a stakeholder to focus on a particular aspect of the data. In [154], Walker discusses the impacts of usability and design issues. The author leads out experiments that show how improvements on design led to better results with users. Several aspects of this dimension overlap with dimensions of the Publishing part of the model. Particularly, different ways of accessing data (*Access Dimension*) heavily impact usability, and exporting data in different formats (*Formats Dimension*), such as CSV, XML, or RDB, is also important to encourage the re-use of data. The way data is published can enable stakeholders to get the most out of the open data.

  Characterisation attributes: The complexity of analysing user interfaces surpasses the scope of this section. Nevertheless, we define a characterisation attribute related to the software tool used by the initiative (CA15), with the understanding that the tool behind the initiative plays an important

[3] http://opendatacommons.org/licenses/ (Date accessed: 2 August 2016)

[4] https://www.govdata.de/dl-de/by-1-0 (Date accessed: 2 August 2016)

[5] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/ (Date accessed: 2 August 2016)

role on the usability. Possible values are the two major open source software tools available for publishing open data: Open Spending[6] and CKAN[7].

- **Feedback** - In order to enable the collaboration between the public sector administration and the other stakeholders, open budget initiatives have to provide means to discuss and give feedback on the provided data. This feedback might be provided to the public administrators either as comments or as a set of recommendations. Ideally, this communication process should be transparent, that is, feedback and recommendations given to public administrators should be publicly available and any changes resulting from the feedback should be recorded. The importance of stimulating user engagement in open data initiatives through feedback and collaboration has been stressed by the Five Stars of Open Data Engagement model [31]. This model justifies the necessity of data being demand driven, contextualised, and collaborative. The lack of collaboration has been listed by Zuiderwijk et al. [160] as one of the main factors hindering the development of open data policies.

  Characterisation attributes: Although this point requires a deeper analysis, we noticed that many open budget initiatives do not present any feedback support. We hence define one basic binary characterisation attribute which is the existence of a feedback mechanism (CA16). We check if it is possible to: (i) comment on data; (ii) submit a new data request; and (iii) report issues noticed in data analysis.

5.4 Analysis of Open Budget Data Initiatives

In this section we apply the proposed model to a number of open budget initiatives with the aim to test its validity. The goal of this evaluation is not to be extensive or to achieve statistical significance, but rather to discover the model's strengths and limitations. The complete results are shown in Table 5.2.

For this evaluation we selected 23 initiatives that provided a balance between primary (11) and secondary (12) sources (CA4). The chosen sample also contains at least five initiatives strongly related to each use perspective, and considers initiatives from 6 countries in addition to the European Union. Some of the analysed initiatives are listed on the Map of Spending Projects[8].

All primary sources in the initiatives we evaluate are maintained by the government, and most of the secondary ones are society driven. Two initiatives are maintained in a partnership between a governmental entity and societal organisations (CA6). Twenty-two open budget initiatives display their objectives (CA1), but only eleven explicitly mention their intended audience (CA2). Also, almost all initiatives offer data for download (18), which favours the Transparency use perspective, and more than half of them (13) make visualisation available, favouring the Participation use perspective. Commenting on data is only allowed in three initiatives, whilst three other initiatives offer a data request form. No reporting issues mechanisms were found, revealing a strong absence of feedback possibilities (CA16). The lack of semantics support (CA12) and linkable data (CA10), which were catered for by only three initiatives each, also may indicate that the Policy Making use perspective is still a long way from being fully enabled. Ten initiatives use categories for the datasets, which facilitate some form of comparison. With regard to the use perspectives, through the evaluation we can conclude the following:

1. **Transparency** - The main requirements for this use perspective, namely data portrayed in transaction level, and machine readable formats, were accomplished by most of the open budget initiatives.

[6] http://github.com/openspending/ (Date accessed: 10 October 2016)

[7] http://ckan.org (Date accessed: 10 October 2016)

[8] http://community.openspending.org/resources/map-of-spending-projects/ (Date accessed: 2 August 2016)

However, feedback handling still requires to be given the appropriate attention. For most of the analysed initiatives we can conclude that stakeholders interested in auditing the government and in translating data into more accessible formats are partially satisfied.

2. **Participation** - This use perspective requires human readable formats that allow citizens without extensive budget knowledge to understand the data and to participate in discussions. Slightly more than half of the initiatives present graphics that can help by providing quick insights over data. Only three initiatives offer maps to visualise budget data, which is coherent to the low number of initiatives that include the Location Dimension (8). Another requirement in this use perspective is the usability and design. Considering the already mentioned limitations on assessing this issue, ten initiatives we evaluated use standard open source software tools. Although this is not the most relevant factor regarding usability, the use of standard tools favours users dealing with several open budget initiatives. Moreover, as open source tools, the more initiatives using these tools, the better they can be developed.

3. **Policy Making** - The main requirements in this perspective are the use of common classifications, vocabularies, and ontologies, and the possibility of linking data with other databases. As already mentioned, semantics support was mostly absent. Comparison tools, also important in this case, were found only in three of the initiatives. Thus, this use perspective is still far from being enabled in most of the analysed initiatives. These results indicate that working on standard terminologies and common conceptualisations, as suggested by OpenSpending [102], is highly desirable.

The above testing of the model and analysis of 23 open budget initiatives has provided us with a number of insights. The particular weak performance on the Feedback Dimension directs us to focus our efforts on further exploring the Consumption part of the model, in order to propose solutions that can contribute to solving the existing issues. The data formats available (CA10) and the semantics (CA12) also need more attention. The Usability Dimension also yields unsatisfactory results.

With regard to the use perspectives, we can conclude that the transparency requisites were mostly accomplished by the analysed initiatives. Participation, on the other hand, is still not heavily supported, while tools for comparing budget data by policy makers are not fulfilling the existing requirements. One has to note, however, that transparency is not easy to achieve. The simple publishing of budget data is not enough. An essential issue here is the *data divide*, a parallel concept to the digital divide, distinguishing people "who have access to data which could have significance in their daily lives and those who don't" [48]. Thus, transparency policies cannot be implemented without actions to foster digital inclusion, and possibly also "data inclusion".

This model we propose has therefore provided us with more information on the current state of existing open budget initiatives. Whilst there is ongoing progress and the basic requisites are somewhat catered for, there is a lot of room for improvement. Through assessing the dimensions specified in the model and identifying existing problems, stakeholders can improve upon any issues and ultimately increase the success potential of the open budget initiative in question.

ID	General Aspects						Publishing																							Consumption				Use Perspective	Language
	Objective			Content		Responsibility	Data																			Formats	Metadata	Semantics	Acces							License	Usability	Feedback				
	CA1	CA2	CA3	CA4	CA5	CA6	CA7				CA8						CA9	CA10	CA11	CA12	CA13									CA14	CA15	CA16				
							Generic	Budget	Spending	Revenue	Generic	Time	Place	Payer	Payee	Category					Downloadable Data	Catalogue	Table	Graphics	Map	Comparison	Stories					Comments	Data Request	Issue Reporting		
1	Yes	Yes	Yes	Sec	1	Both		x		x		x	x	x		x	Ag	3	No	No	x		x	x	x			No	OS				UP1, UP2	DE
2	Yes	Yes	No	Sec	5	Soc			x		x	x	x	x	x		Ag	1	No	No						x	x	No					UP1, UP2	EN
3	Yes	No	Yes	Sec	5	Soc	x	x	x		x	x	x	x	x	x	Tr	3	Yes	No	x	x	x	x				Yes	OS				UP3	EN
4	Yes	Yes	No	Sec	1,2,3	Soc	x	x	x	x	x	x	x	x	x	x	Ge	1–3	Yes	No	x	x						Yes	CK	x			UP1, UP2	DE
5	Yes	No	No	Prim	1	Gov	x				x						Ge	1–3	Yes	No	x	x						Yes	CK	x			UP1, UP2	DE
6	Yes	No	No	Prim	1	Gov	x		x			x	x	x	x	x	Ge	3	Yes	No	x	x	x		x			Yes	CK	x			UP1, UP2	EN
7	Yes	No	No	Prim	1	Gov	x				x						Ge	3	Yes	No	x	x						No	CK		x		UP1	EN
8	Yes	Yes	No	Prim	1	Gov	x	x			x						Ge	1–5	Yes	Yes	x	x	x	x				Yes	CK				UP1, UP2	PT
9	Yes	No	Yes	Sec	1,2,3 L	Soc			x			x		x	x	x	Tr	N/A	No	No			x	x				Yes					UP2	PT
10	Yes	Yes	Yes	Sec	5	Soc			x								Tr	5	Yes	Yes	x		x	x		x		No					UP3	EN
11	Yes	No	Yes	Prim	3	Gov		x	x	x		x		x	x	x	Tr	3	No	No	x		x	x				No			x		UP1, UP2	PT
12	Yes	No	Yes	Sec	1,2,3	Gov			x			x	x	x	x	x	Tr	3	Yes	No	x		x	x	x			No					UP1, UP2	EN
13	Yes	Yes	Yes	Prim	1	Gov			x	x		x		x	x	x	Tr	3	No	No	x		x					No					UP1, UP2	PT
14	Yes	No	Yes	Prim	2	Gov		x	x	x		x		x	x	x	Tr	3	No	No	x		x	x				No					UP1, UP2	PT
15	Yes	Yes	Yes	Sec	1	Soc		x				x		x		x	Ag	N/A	No	No			x	x				No	OS				UP2	PT
16	Yes	Yes	Yes	Prim	1,2,3	Gov		x	x	x		x	x	x	x	x	Tr	3	Yes	No	x		x					No					UP1	PT
17	Yes	Yes	Yes	Sec	3	Soc		x	x	x		x	x	x		x	Ag	N/A	Yes	No			x	x		x	x	Yes					UP3	IT
18	Yes	Yes	Yes	Sec	4	Both			x			x	x	x	x	x	Tr	3	Yes	No	x							Yes					UP1, UP3	EN
19	Yes	Yes	Yes	Sec	4	Soc			x			x	x	x	x	x	Tr	N/A	Yes	No	x		x			x		No	OS				UP3	EN
20	No	No	Yes	Prim	1	Gov		x	x			x	x	x	x	x	Tr	5	Yes	Yes	x							No					UP1	PT
21	Yes	No	No	Prim	3	Gov	x	x	x	x	x	x		x	x	x	Tr	3	Yes	No	x		x	x				Yes					UP1, UP2	RU
22	Yes	Yes	No	Prim	3	Gov	x	x	x	x	x	x		x	x	x	Tr	3	Yes	No	x		x	x				Yes			x		UP1, UP2	RU
23	Yes	No	Yes	Sec	2	Soc	x	x	x		x	x		x	x	x	Tr	3	Yes	No	x		x	x				Yes	OS				UP1, UP2	RU

Table 5.2: Results of the application of the open budget initiatives assessment model on 23 open budget initiatives. In CA5: (1) local; (2) regional; (3) national; (4) transnational; (5) generic; and (L) legislative budget. In CA6: (Gov) Government; and (Soc) Society. In CA9: (Tr) transaction; (Ag) aggregate; and (Ge) generic. In CA10, N/A means not applicable, when there is no data for download. In CA15: (OS) Open Spending; (CK) CKAN. Use Perspectives: UP1 - Transparency, UP2 - Participation, UP3 - Policy Making.

Concluding Remarks for Part II: Open Data in the Government Domain

In the chapters in Part II we gave an overview of open government data initiatives in order to answer the following research question:

Research Question 1:

What are existing approaches and techniques that enable the publishing and consumption of open data?

We therefore implemented a systematic research with the aim of obtaining a clear picture of the current situation of open data initiatives. We answer a set of questions, mainly concerning open government data initiatives and their impact on stakeholders, existing approaches for publishing and consuming open government data, existing guidelines, and challenges (see Table 5.3) for the discussed approaches. We identify corruption to be the major problem which triggered open government data initiatives, and we point out the various motivations for opening government data. One major motivation is transparency, which however should not be an end in itself. It should rather be a means to enhance an open government initiative. This perspective will avoid governments who publish their data for the sake of it, rather than striving to provide useful data which stakeholders can use, re-use and distribute, and ideally even innovate upon.

Nature of Challenge	Challenge	Possible Solution
Technical	Formats	Using a Machine-processable, non-proprietary format
	Ambiguity	Using a descriptive format; Adding documentation/metadata
	Discoverability	Using good quality metadata; More advanced search tools on portals
	Representation	Defining and using standardised representation; Using named graphs for versioning
	Capacity	Applying standards; Large-scale training
Policy/Legal	Copyright/Licensing	Defining standard data policies
	Conflicting Regulations	Defining open government data initiative policies and legal frameworks
	Privacy/Data Protection	Defining privacy regulations; Implementing access control mechanisms (this limits the openness of the data)
	Liability	Social interaction; Raising awareness; Defining legal frameworks
Economic/Financial	Budget Provision	Providing budget specifically for open data initiatives
Organisational	Institutionalisation	Re-organising the current organisational structure; Defining open government initiative policies
	Overlapping Scope	Using provenance metadata
	Technical Support	Providing support to public entities with the executing of an open data initiative
Cultural	Motivation	Raising awareness on the re-use of open data and its benefits
	Awareness	Highlighting the value and potential of open data
	Public Participation	Raising awareness; Providing incentives
	Competition	Providing specific data at a nominal fee (this limits the openness of the data)

Table 5.3: Overview of challenges in open government data initiatives.

Based on existing open data life cycles and on existing open data initiatives, in Chapter 2 we define the open government data life cycle, which is provided as the depiction of the processes and their ideal order required during the lifetime of open government data. The definition of this life cycle is not meant to be an extensive description of the processes; rather we propose it to act as a guideline for stakeholders to follow during their participation in an open government data initiative.

One of our main contributions in this part of the thesis is the discussion about open government data initiatives in Chapter 3. We first discuss different assessment frameworks for evaluating various aspects of open government initiatives. We follow by providing a summary of open government initiative evaluations found in our primary studies. The various publications covered evaluate different aspects of the initiatives, such as the features provided, the openness level of the available data, and the impact on relevant stakeholders. Many of them also evaluate the current status for specific administrative regions. Based on the results of our evaluations, we proceed to point out challenges and issues which hinder open government initiatives from reaching their full potential, data from being truly open, or factors that influence public entities from jumping on the open data bandwagon in the first place. We also direct our efforts towards identifying the different stakeholders who participate in open government initiatives, and discuss public participation. We also explore the different levels of achievable impacts through open government initiatives; namely access to information, transparency, accountability, and finally democratic governance.

In Chapter 4 we focus on the publishing and consumption processes of open government data, which are the most essential processes within the life cycle. We classify different publishing and consumption approaches, and identify different data quality aspects which influence or are influenced by the approaches undertaken for consuming or publishing the data. Based on the literature covered in the survey, the Eight Open Government Data Principles, and the Five Star Scheme for Linked Open Data, we extract and integrate various guidelines for publishing open government data. Adhering to these guidelines will improve the end usability of the data (for consumption), and the resulting success of the initiative in question.

In the final chapter of this part, Chapter 5, we focus our efforts on open budget initiatives, as a subset of open government data initiatives. We provide a model that enables stakeholders to analyse open budget initiatives. This model is provided with the aim of targeting a niche in existing approaches and mechanisms to assess the various strategies for publishing budget data. We define three use perspectives and hence assess the open budget initiatives' fitness for their use in the use perspectives. We thereafter validate the model upon 23 existing initiatives and provide the relevant discussions with regard to the defined use perspectives.

To conclude, we revisit the research questions posed in Section 2.1 and summarise the discussions in this part with the following observations:

- *What are the characteristics of existing implementations of open government initiatives?*
  Open government data initiatives vary in nature, and the implemented approaches reflect this heterogeneity. However, the most common approaches include data portals, data catalogues, and services. Whatever the implementation, open government initiatives are in essence very similar to any open data initiative, and the aims and motivations are therefore also somewhat homogeneous. The aims and motivations are usually focused on transparency, access to information, and stakeholder engagement.

- *What are the supported technical aspects, features and functions in existing approaches?*
  The aim behind most open government data initiatives is to publish data in order to make it available for re-use. The most commonly available feature is therefore the availability of data. This basic feature is then complemented through other technical aspects, together with features and functions, such as multilinguality, different data formats, data accessibility, data content, and visualisation tools. Many assessment frameworks are defined in literature with the aim of analysing existing open government initiatives, however most of them are based on the Five Star Scheme for Linked Open Data or the Eight Open Government Data Principles.

- *Are there any defined guidelines for the publishing or consumption of open government data?*
  While a number of different guidelines are defined in literature, there are no agreed upon standards for the publishing or consumption of open government data. Yet, by following the integrated overview of guidelines we propose, we attempt to provide a higher possibility for an open government data initiative to succeed. We also focus on the quality of the data, which directly impacts the eventual publishing or consumption of the data itself.

- *What are existing challenges within open government initiatives?*
  We identified and explored a number of challenges, including technical, policy and legal, economic and financial, organisational, and cultural barriers. These challenges impact a number of aspects of an open data initiative, such as the potential that can be exploited within an initiative, whether stakeholders decide to participate in an initiative, and whether the data in the initiative is truly open or otherwise. In combination these challenges ultimately affect the success of an open government initiative.

- *What are possible impacts of open government initiatives on the relevant stakeholders?*
  Transparency was identified to be one main aim of opening government data, however it is not the only impact. There are varying impacts of open government data initiatives, including the direct impact of access to information that results in more informed citizens, as well as an increase in accountability and a higher opportunity for citizens to actively participate in governance processes.

Part III

Lowering Barriers to Open Data Re-Use

The research and contributions in this part are focused on exploring existing approaches that aid stakeholders in consuming open data. Through considering data consumption as the initial process required for value creation, our aim here is to further enable stakeholders, especially non-experts, to easily and efficiently consume open data. This part is divided in two chapters. In Chapter 6 we provide insight into our motivation, as well as an overview of related literature, while in the next chapter, Chapter 7, we describe our contribution of the ExConQuer Framework as an approach towards consuming open data.

The chapters in this part are based on the following publications:

- **Judie Attard**, Fabrizio Orlandi, Sören Auer. *ExConQuer Framework - Softening RDF Data to Enhance Linked Data Reuse*. In Proceedings of the ISWC 2015 Posters & Demonstrations Track co-located with the 14th International Semantic Web Conference (ISWC-2015), Bethlehem, Pennsylvania, USA, October 11, 2015.
- Spiros Mouzakitis, Dimitris Papaspyros, Michael Petychakis, Sotiris Koussouris, Anastasios Zafeiropoulos, Eleni Fotopoulou, Lena Farid, Fabrizio Orlandi, **Judie Attard**, John Psarras. *Challenges and Opportunities in renovating Public Sector Information by enabling Linked Data and Analytics*. In Proceedings of the Information Systems Frontiers Journal, 2016.
- **Judie Attard**, Fabrizio Orlandi, Sören Auer. *ExConQuer: Lowering barriers to RDF and Linked Data re-use*. To appear in Proceedings of the Semantic Web Journal, accepted on 12 October 2016.

CHAPTER 6

Open Data and its Re-Use

The radical advances in technology, particularly through the advancement of the World Wide Web, have created new means to share knowledge. However, although barriers to information access have been lowered through various means (e.g. hypertext links, web search engines, REST APIs), accessibility to raw data was only afforded the same importance in recent years [19]. The relatively recent open data movement, through motivations such as transparency, accountability, and other societal goals, has prompted the release of a huge number of datasets from a large number of different domains to the public. Moreover, Linked Open Data, as a subset of generic open data, shows an evident increase of data release. The increasing adoption of Linked Data practices, as indicated by the extraordinary growth in the Linked Open Data Cloud's[1] volume over the past eight years, as well as the number of triples continuously crawled by the LOD Laundromat[2], act as an affirmation. Through the implementation of Linked Data practices, open data is published with a more meaningful representation, as opposed to raw data that used to be published in formats such as CSV, which need metadata to be interpretable. Yet, this does not mean such data is easier for the average stakeholder to locate, access, or most importantly, re-use. Individuals facing these hurdles are typically more acquainted with file formats such as generic JSON, XML, basic CSV, or other legacy formats such as XML-based Keyhole Markup Language (KML) or GPS Exchange Format (GPX), therefore finding the sophisticated nature of the RDF format overwhelming.

Unfortunately, the emergence of a wide number of tools supporting people to publish their data as (Linked) Open Data[3], has not been complemented by approaches supporting non-experts to consume existing Linked Data in formats other than RDF [19]. Such tools and approaches would be vital to aid non-experts to exploit Linked Open Data even though either they are not able to understand and interpret it, or have a system that understands a different format.

The contributions in this part of the thesis form part of the research and results within the LinDA Project[4]. This project has the objective of enabling stakeholders to better and more easily exploit Linked Data. Whilst attempting to target the niche in existing tools that provide such services, the LinDA Project provides the relevant tools that enable data providers to provide re-usable, machine-processable Linked Data, and that enable consumers to easily consume Linked Data without requiring any expertise. In the following chapters we hence strive to answer the following research question, as defined in Section 1.2:

Research Question 2:

[1] http://lod-cloud.net/
[2] http://lodlaundromat.org/
[3] http://www.w3.org/wiki/LinkedData (Accessed on 21 August 2016)
[4] http://linda-project.eu/ (Date accessed: 2 August 2016)

How can we enhance the consumption process of a data product in order to enable further value creation?

With the aim of identifying strengths and weaknesses in existing approaches, we here investigate existing open data consumption approaches and related Linked Data technologies. We identify vital aspects that enable and encourage stakeholders in exploiting open data. In this chapter we therefore proceed to provide an overview of related work that forms the basis to the contributions in the following chapter, where we propose the *ExConQuer Framework* (Explore, Convert, and Query Framework); a set of open source tools[5] whose aim is (i) to facilitate the publication and consumption of RDF data in a wide variety of generic, legacy or domain-specific formats[6], as well as (ii) to enable stakeholders to easily re-use persisted transformations. The ExConQuer Framework thus provides a query builder tool that allows stakeholders to easily construct queries over RDF using the SPARQL querying language, a converter tool that provides the conversion of data from RDF to various formats, and a provenance-aware management system that enables stakeholders to re-use previous queries and conversions.

6.1 Preliminaries on Linked Data

The term *Linked Data* is used to refer to a set of best practices for publishing and connecting structured data on the Web [17]. In order to establish some guidelines, Berners-Lee defined four principles for Linked Data[7]:

1. Use URIs as names for things;
2. Use HTTP URIs so that people can look up those names;
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL); and
4. Include links to other URIs so that they can discover more things.

Therefore, Linked Data is published on the Web in a machine-readable format, where its meaning is explicitly defined. It is also linked to and from external datasets. This has the potential of creating the *Web of Data* (also known as Semantic Web); a huge distributed dataset that aims to replace decentralised and isolated data sources [50]. The benefits of applying Linked Data principles to government data as covered in literature include [30, 67]:

- Simpler data access through a unified data model;
- Rich representation of data enabling the documentation of data semantics;
- Re-use of existing vocabularies;
- Use of URIs allow fine-grained referencing of any information; and
- Related information is linked, allowing its unified access.

[5] Source code on Github: https://github.com/LinDA-tools/QueryBuilder (Date accessed: 2 August 2016)

[6] While hundreds are in existence: http://en.wikipedia.org/wiki/List_of_file_formats, we here focus on the more popular ones such as JSON, CSV and RDB. (Date accessed: 2 August 2016)

[7] http://www.w3.org/DesignIssues/LinkedData.html (Date accessed: 2 August 2016)

Widely established as a standard, the RDF (Resource Description Framework) data model is commonly used to represent Linked Data. RDF enables the machine-readable and interoperable modelling of concepts (whether real-world or abstract) as resources on the Web. Moreover, RDF is domain independent, and is therefore flexible enough to enable the representation of concepts from varying domains. RDF is a graph-based data model whose structure is made up of a *triple* consisting of a *subject*, a *predicate* and an *object*. Triples can be portrayed as two nodes, where the subject and object are connected by the predicate acting as an arc. Figure 6.1 shows an example of a triple, where `ex:John` is the subject, `foaf:name` is the predicate, and `"John Doe"` is the object, where `ex` and `foaf` are the prefixes (shorthand notation) for the namespace of the vocabularies used. Listing 6.1 shows a number of triples serialised in RDF/XML as an example. This small knowledge base consists of two resources representing Mary (`<http://example.org/Mary>`) and John (`<http://example.org/John>`). These two resources are defined as `Person` using the FOAF vocabulary[8], where each resource has a name (`foaf:name`) represented as a string literal. John's resource is linked to Mary's resource via the `foaf:knows` predicate. A number of triples such as the ones in Listing 6.1 then make up a dataset, and are usually stored in a triple store and accessed through a SPARQL endpoint.

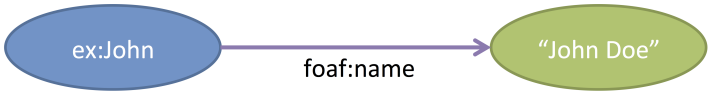

Figure 6.1: Triple structure: Subject - *ex:John*, Predicate - *foaf:Name*, Object - *"John Doe"*.

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
    xmlns:foaf="http://xmlns.com/foaf/0.1/"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
>
  <rdf:Description rdf:about="http://example.org/Mary">
    <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
    <foaf:name>Mary Doe</foaf:name>
    <foaf:age rdf:datatype="http://www.w3.org/2001/XMLSchema#int">28</foaf:
        ↪ age>
  </rdf:Description>
  <rdf:Description rdf:about="http://example.org/John">
    <foaf:name>John Doe</foaf:name>
    <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
    <foaf:knows rdf:resource="http://example.org/Mary"/>
  </rdf:Description>
</rdf:RDF>
```

Listing 6.1: A small knowledge base in RDF/XML portraying the serialisation of two resources

The SPARQL query language is a protocol that allows the querying and manipulation of RDF data. Through implementing a graph-based pattern matching approach, SPARQL can be used to retrieve or modify a subset within a dataset that matches the specified query. In fact, SPARQL supports four different queries, namely SELECT queries (data retrieval), ASK queries (boolean "yes/no" queries), CONSTRUCT queries (allowing the creation of new RDF graphs from query results), and DESCRIBE

[8] `http://xmlns.com/foaf/spec/` (Date accessed: 2 August 2016)

queries (providing as a result a graph describing a queried resource). Listing 6.2 shows an example of a SELECT SPARQL query that retrieves all friends of John as well as their age, and sorts the results by the age in ascending order.

```
SELECT ?friend_names ?friends_age WHERE {
<http://example.org/John> <http://xmlns.com/foaf/0.1/knows> ?friends .
?friends a <http://xmlns.com/foaf/0.1/Person> .
?friends <http://xmlns.com/foaf/0.1/name> ?friend_names .
?friends <http://xmlns.com/foaf/0.1/age> ?friend_age .
} ORDER BY ASC(?friends_age)
```

Listing 6.2: A SPARQL SELECT query example that can be executed on the knowledge base portrayed in Listing 6.1

6.2 Related Work

Our approach is varied in nature, comprising data exploration, query generation, data views, and a transformation explorer. To the best of our knowledge, there is no Linked Data consumption framework that includes all the functions as the one we propose. Yet, there are a number of tools that tackle the different approaches separately.

6.2.1 Linked Data Exploration Systems

In the ExConQuer Framework we enable users to explore datasources in order to identify if and how the data they require is represented in existing open datasets. Therefore we here explore various data exploration systems.

In [84], Marchionini distinguishes between *lookup* and *exploratory search* activities. Lookup activities are done to satisfy specific information needs, such as searching for a known item, where the user has defined keywords to use. On the other hand, exploratory search refers to cognitive consuming search tasks, such as learning or investigation. Here, the information need is less well-defined than in a lookup activity and the keywords are not known in advance, therefore also evolving during the activity. In our approach we cater for both activities, where users are given both results that exactly match the specified keyword, and also results that are related to that keyword, as well as being given the option to freely explore the datasource in question by viewing all contained classes and their subclasses.

Tvarozek and Bieliková [142] attempt to facilitate exploratory search by extending their own base browser through the implementation of three search paradigms; keyword-based, view-based, and content-based. The browser also enables dataset exploration through adaptive result overviews and incremental graph-based resource exploration. A drawback for using this approach is the possibility of information overload, since a huge dataset might result in an enormous amount of facets or nodes.

The authors of [52] use Facet Graphs in their approach to build semantically unique queries. Users are given the option to choose the result set they need, as well as the facets to filter it. Both are represented as nodes in a graph visualisation and enable them to produce a personalised interface to build search queries. Compared to the previous approach in [142], by enabling users to enter keywords the authors reduce the risk of information overload.

In [122], Araújo et al. present Explorator, a tool for exploring RDF data through direct manipulation. Users are enabled to explore a semi-structured RDF database through browsing and searching. While the led experiments and studies indicated that users with a basic knowledge of RDF were able to use the tool,

the authors also point out that the Explorator is better suited to advanced users who have solid knowledge about RDF, further motivating our approach.

Popov et al. [110] propose Visor, a multi-pivot approach that allows users to explore datasets from multiple points in the graph. Visor consists of a generic data explorer tool that can be configured on any SPARQL endpoint. Here, a user is able to explore existing classes in the dataset at hand, the related properties and classes, and individual instances. A graph is then rendered in order to show the user selection and the relations between them (if any). Visor enables users to query a user's selection by creating custom spreadsheets, and then convert them to CSV or JSON.

While numerous tools that enable users to explore Linked Data exist, most of them are targeted for more experienced users who have some knowledge of either RDF or the data's underlying schema. Therefore, such tools are unsuitable to fit our aim of lowering the entry barrier towards consuming and re-using Linked Open Data for value creation.

6.2.2 SPARQL Query Builders

The first process towards achieving re-usability is data access. Linked Open Data is usually accessible on data portals or catalogues through SPARQL endpoints or data dumps. The latter method for accessing data has the disadvantage of generally resulting in a large bulk of data, with the user having no control to get specific data (such as a subset) from the data the provider made available as a dump. Moreover, data might also be outdated. While SPARQL endpoints allow thorough control over what data to access, then there is the disadvantage of having to use SPARQL, and using SPARQL to search through data stores is a tedious process and limits data access to Semantic Web practitioners [21, 26]. This is mainly due to two reasons; (i) because of the syntax barrier, and (ii) due to the heterogeneity of the data and its schema. As yet, there are few tools that help inexperienced users with respect to the creation and editing of SPARQL queries.

Russell and Smart [121] present NITELIGHT, a tool that enables users to create SPARQL queries using a set of graphical notations and GUI-based editing actions. NITELIGHT uses a visual query language, vSPARQL, to provide graphical formalisms for SPARQL query specification. Users can construct a query through dragging and dropping ontology elements. This approach, while suitable for users with at least a minimal understanding of the SPARQL query language, is not suitable for users who do not know SPARQL or the underlying schema of the dataset to be queried.

Similar to NITELIGHT, Haag et al. [49] also implement a visual approach. The authors define it to be a novel approach for visual SPARQL querying based on the filter/flow model. Thus, no structured text input is required, rather, queries can be generated entirely through the use of graphical elements, and filter restrictions are shown, rather than a representation of the complete query. While this approach does not require knowledge of the SPARQL query language, users are expected to be familiar with the Semantic Web and the filter/flow concepts. Moreover, while this approach allows users to query a dataset, they need to know if and how the information they need is available in the dataset in question.

In contrast to the above, in [113] Pradel et al. present an approach where users can enter a natural language query that is then translated into a formal graph query through the use of query patterns. The aim behind this approach is to hide the complexity of formulating a query expressed in graph query languages such as SPARQL, thus enabling end users to use natural language queries to query ontology-based knowledge bases. The approach described here still has some usability issues. For instance, only English and French can be used as natural languages for the input query. Besides, users who might know the data they need, but not exactly how it is represented in the dataset, will find difficulty in expressing the correct query even if a natural language is used.

QueryMed [129] is the tool that is most similar to our approach for query generation. Focused on the

medical domain, this tool enables users with no knowledge of SPARQL to run queries across SPARQL endpoints. The tool requires users to input specific search terms. Users are then given the possibility to filter the results and restrict the query further. A key difference in QueryMed when compared to our approach is that the authors base their search on properties. Thus, when a user selects one or more data stores, the tool displays all the properties within these stores. Apart from resulting in an information overload, this approach is not particularly useful when there many domains involved (e.g. DBpedia), specifically due to the heterogeneity of the data.

6.2.3 Data Transformation and Exploration Systems

There are a myriad of tools available for converting between data formats, such as Any23[9], Datalift[10] [127], Db2triples[11], and METAmorphoses[12] [141]. However, there are very few tools that enable the conversion of RDF to other, less semantically rich formats (such as [138]). Considering RDF is much more expressive than most other formats, it is understandable that efforts and interest are focused in that direction, however we need to cater for users who require the conversion of Linked Open Data (which is generally available in RDF) to a format they understand which is also compatible to their native systems, such as Microsoft Excel, Open Office, R, or Tableau. Albeit this might result in some loss of information, the advantages outweigh this shortcoming since it will encourage users to exploit such data, rather than being deterred due to unfamiliarity with Linked Data or RDF.

The Transformation Explorer, a provenance-aware management system, is a core contribution within this part of the thesis. The aim behind this tool is to provide a means for users to explore and re-use what we call *Linked Data Publications*. A Linked Data Publication consists of all the information generated in the transformation of data, including the SPARQL query used, its description, the datasource(s) queried, the initial and target data formats, and the user generating the Linked Data Publication instance.

In [85], Marie and Gandon survey existing Linked Data-based exploration systems, however all the systems they review are based on exploring data, rather than Linked Data Publications which represent the data, as well as the transformations made on it. SPARQLpedia[13] is more similar to what we propose, in that it is a service that allows users to submit SPARQL queries in a searchable repository. The Transformation Explorer follows the same concept, however through retaining provenance information we enable users to not only browse existing queries, but also re-execute them to get updated results or even edit them to refine their query.

[9] https://any23.apache.org/download.html (Date accessed: 2 August 2016)
[10] http://datalift.org/ (Date accessed: 2 August 2016)
[11] http://www.w3.org/2001/sw/wiki/Db2triples (Date accessed: 2 August 2016)
[12] http://metamorphoses.sourceforge.net/ (Date accessed: 2 August 2016)
[13] http://composing-the-semantic-web.blogspot.nl/2009/01/sparqlpedia-sharing-semantic-web.html (Date accessed: 23/05/2016)

CHAPTER 7

The ExConQuer Framework

The ExConQuer Framework assists data publishers and consumers in exploiting and re-using Linked Data by providing tools that enable them to easily and simply explore, query, transform, and publish Linked Data. For these reasons, the ExConQuer Framework is also ideal to introduce Linked Data (and the SPARQL querying language) to new users. The framework is based on the concept of *RDF softening*. In contrast to the semantic *lifting* of data into RDF, which addresses the enrichment, mapping, and transformation of semantically shallow formats, the softening process is then *the generation of domain-specific RDF data views in semantically-shallow representation formalisms*. This will enable stakeholders to more easily obtain, interpret and re-use existing Linked Data in conventional formats. Moreover, any transformations executed on the data are persisted to enable their re-use. Initiatives such as the one undertaken by the W3C CSV on the Web working group[1], which aims to standardise JSON-LD serialisation, promise to lower the entry barrier to Linked Data re-use. Yet to the best of our knowledge, very few approaches address the need for the provision of semantically-rich RDF data in shallower formats. Although this might appear to be counter-productive, it is favourable to offer the reduction of a degree of semantics in favour of an increase in the *degree of (re-)usability* by stakeholders who would otherwise refrain from using the data. Through retaining provenance information we also ensure that the softening process does not result in the loss of the richness of RDF representation, and users are also given the option to lift back the results to RDF.

Based on the motivation of providing stakeholders with a tool that enables them to consume Linked Open Data easily without requiring previous knowledge of RDF, SPARQL, or the datasets' underlying schema, we provide the following contributions as part of the ExConQuer Framework[2]:

- **Query Builder Tool** - enables users to explore, query, and convert datasources (datasets or subsets) through endpoints, available online at http://butterbur22.iai.uni-bonn.de:3000/query/builder;
- **RDF2Any API** - provides the functionality to query and convert RDF datasources into a number of different formats through RDF softening;
- **ConQuer Ontology**: used to represent transformations carried out in the Query Builder, available online at http://purl.org/eis/vocab/cqo;

[1] http://www.w3.org/2013/csvw/wiki/Main_Page (Date accessed: 2 August 2016)

[2] More information on the framework, including source code and evaluation results, can be found here: http://eis.iai.uni-bonn.de/Projects/ExConQuer.html (Date accessed: 2 August 2016)

- **Transformation Explorer**: a faceted browser that enables users to explore and re-use Linked Data Publications (all information generated during the use of the Query Builder Tool, such as the query used, the datasource queried, the data formats, etc.) available online at http://butterbur22.iai.uni-bonn.de/pam/;

- **Evaluation** - a usability evaluation on the tools within the ExConQuer Framework, as well as a further effort evaluation that analyses the time and effort required with or without the ExConQuer Framework. Surveys are listed in Appendices A and B, and the respective results available online at http://eis.iai.uni-bonn.de/Projects/ExConQuer.html.

In order to have a clearer idea about the use of the ExConQuer Framework, we here explore a simple use case scenario with German Tours; an SME that provides various tours for tourists visiting Germany. A number of languages are used for the tours, including English, Italian and German. Alice, the manager, thinks the tours provided by German Tours need to be updated to reflect current tourist trends. She therefore starts looking for any relevant information on the web and discovers statistical data on incoming tourists published by the German National Statistics Office (NSO). Alice does not know SPARQL, so she accesses the *Query Builder Tool*, found within the ExConQuer Framework. This tool enables Alice to easily access the required data through an easy-to use and intuitive interface that creates a SPARQL query for the user in order to query the desired data. She then queries the NSO dataset to get data such as the yearly number of visiting tourists, the most popular tourist attractions, the country where the tourists came from, etc.

Once Alice is happy with the query results, Alice proceeds to export the results in RDB (relational database file format), in order to be able to easily integrate the data in the SMEs native database. The *RDF2Any Converter* converts the data from RDF to RDB and prompts Alice to store the data on her system. After linking her data with the data published by the NSO, Alice realises that she forgot to export the language spoken in the country of origin of tourists in Germany. She proceeds to access the *Transformation Explorer* (a provenance-aware management system) to explore the available views. She uses the filters in the faceted browser to find the query she had just executed, then loads it back into the Query Builder. Alice then adds the required property, and exports the data once again.

Through the integration of the newly-extracted data with the enterprise data, Alice starts discovering details that help her adapt the services provided by the SME. For instance, Alice discovers that while a large number of Spanish tourists visits Germany, very few book the provided tours. This indicated that Spanish-speaking tour leaders are required. Thus, by re-using the data published by the NSO, and including the SMEs private data, German Tours is given the opportunity to adapt according to current tourist trends, and retain a competitive edge.

Figure 7.1 shows an abstract overview of the processes within the framework. Keeping in line with the above use case, through the first stage (*Dataset Exploration*), the user can explore the available datasource, for example the NSO portal. The user then generates a SPARQL query in the *Query Building* step, adding filters in order to obtain the required data. The user then has the option to *Transform* the query results into various formats. The querying and transformation processes are then represented as a *Linked Data Publication*. Through the Transformation Explorer, the user can explore Linked Data Publications and proceed to re-use, share, or edit them by executing further transformations. The abstract overview in Figure 7.1 is implemented through the tools provided within the ExConQuer Framework; namely the the *Query Builder Tool* (Section 7.1), the *RDF2Any API* (Section 7.1.1), the *Transformation Explorer* (Section 7.2), and the *ConQuer Ontology* (Section 7.2.1).

Figure 7.2 shows an overview of the architecture within the framework, and how the various tools interact with each other. The user can create a SPARQL query through the Query Builder Tool, then

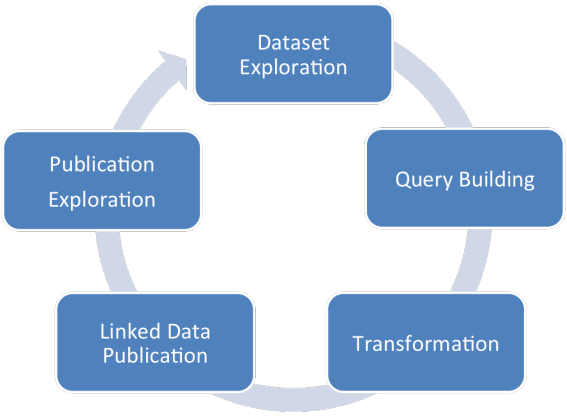

Figure 7.1: Abstraction of the processes within the ExConQuer Framework.

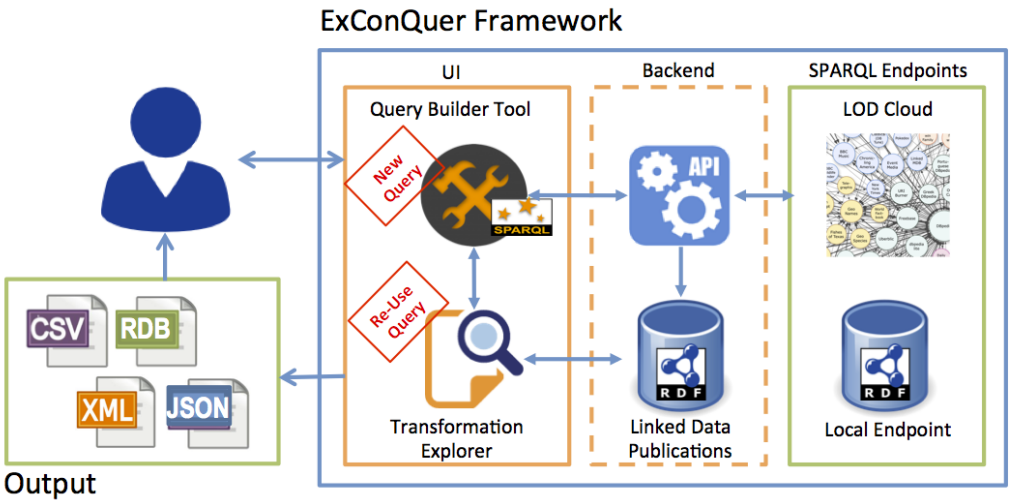

Figure 7.2: The architecture of the ExConQuer Framework.

query a datastore (through a SPARQL endpoint) through API calls. Once happy with the results, the user can export them in a number of different formats, and re-use them accordingly in his or her native system. Information pertinent to the executed processes is then persisted in a triple store as Linked Data Publications. The latter are represented using the ConQuer Ontology which we propose for recording the provenance information of the transformation. The represented data includes the queried datasource, the SPARQL query, the format conversion, etc. A user can access all this relevant information through the Transformation Explorer, which allows a user to re-use existing resultsets or modify them through the Query Builder.

7.1 Query Builder Tool

In the ExConQuer Framework we enable users to explore existing open datasets. We target users who either do not know the content of the datasource in question, or otherwise do not know how specific data is represented in this datasource. Our approach is intended to be particularly user friendly and simple, to allow non-experts to easily use the tool to achieve the goal of re-using open data. An additional

advantage of this simplicity is that the tools can be used to introduce Linked Data to new users, as well as helping them to learn the SPARQL query language. Through the RDF2Any RESTful API, and by using the datasets' schema, the Query Builder Tool (shown in Figure 7.3), enables users to navigate through classes, subclasses, instances, and properties in a somewhat similar manner to a faceted browser, without requiring them to know the structure of RDF data. The API calls concerned with this exploration task are made up of a number of actions that essentially hide the RDF data model and help in the exploration of RDF data and the underlying structure (e.g. to get class labels). This API hence encapsulates the functionality of this tool, and can also be re-used in other frameworks.

7.1.1 Dataset Exploration

Figure 7.3 shows the different parts of the UI of the Query Builder Tool. The provided exploration functions are particularly useful for users who do not know exactly what data from the available linked datasets is useful for their purpose, or for those who do not know the underlying schema behind the dataset in question.

In Step 1, the user can select any datasource (usually a SPARQL endpoint containing one or more datasets) from the auto-complete drop down list or otherwise add a new endpoint. In Step 2 the user can then proceed to explore the classes contained in the selected datasource. Here the user can either view all classes, or view the classes which match a given keyword. The user also has the option (by clicking on the plus button) to expand the view and show the subclasses of the selected class, if any are available.

Consider the API call required for this process. This call abstracts the complexity required to get all classes matching a given keyword. After the user has selected the datasource to explore (Step 1), the user enters a keyword and the API call containing the SPARQL query shown in Listing 7.1 is executed, where the `%%Search-String%%` variable will be replaced by the keyword entered by the user. This query is used to search within the selected datasource to look for resources of type `owl:Class` or `rdfs:Class`. Here the user also has a choice (through the UI) to search for the latter resources either through the resources' labels only, or otherwise extend the search to also include resources' URIs. In the case of the former, the query would be a little different, in that both the OPTIONAL clause and the REGEX component for the class would be removed.

```
SELECT distinct ?class ?label WHERE {
        { ?class rdf:type owl:Class }
        UNION
        { ?class rdf:type rdfs:Class }.
        OPTIONAL { ?class rdfs:label ?label . }
        FILTER(
                ( bound(?label) && REGEX(?label, "\\b%%Search-String%%","
                    ↪ i") ) ||
                REGEX(str(?class), "\\b%%Search-String%%","i")
                )
} ORDER BY ?class
```

Listing 7.1: The SPARQL query to obtain classes matching a given keyword

In Step 2, along with the classes and subclasses, a number of example instances are displayed, ordered by the amount of local backlinks each instance has. The SPARQL query used to obtain this data is as shown in Listing 7.2.

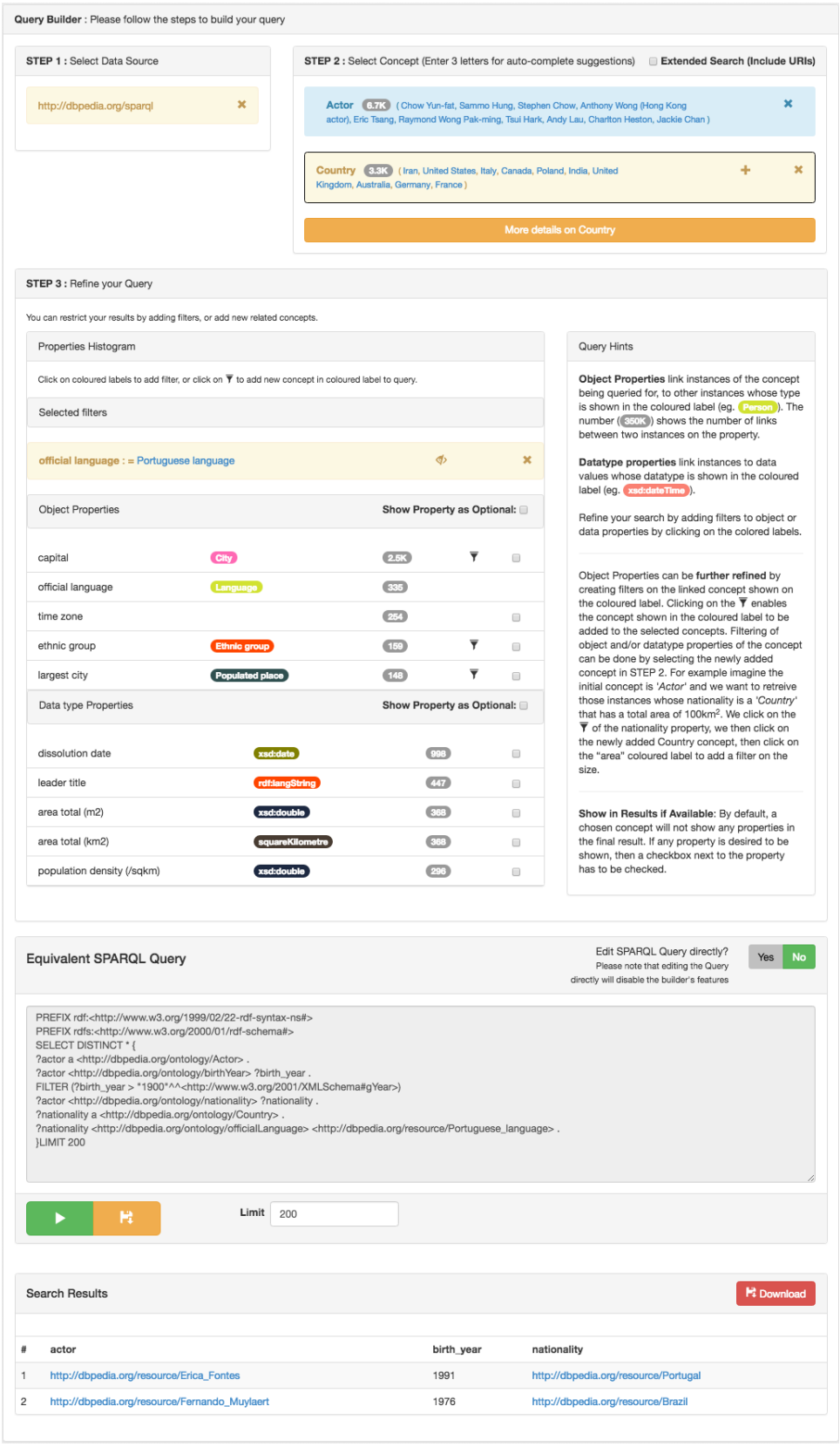

Figure 7.3: Query Builder Tool: Enables the exploration of linked open datasets and the generation of SPARQL queries.

```
SELECT ?label ?instance {
        ?instance rdfs:label ?label .
        {
                SELECT DISTINCT ?instance (COUNT(?x) AS ?cnt) WHERE {
                        ?instance a <%%Concept-URI%%> .
                        ?x ?p ?instance .
                }
                GROUP BY ?instance
                ORDER BY DESC(?cnt)
                LIMIT %%limit%%
        }
        FILTER(langMatches(lang(?label), "EN"))
}
```

Listing 7.2: The SPARQL query used to obtain a number of sample instances ordered by number of backlinks

Once a class is selected, the user can proceed to Step 3; the Properties Histogram view, where all the properties of the selected class are shown. The view is divided into *Object Properties* and *Data Type Properties*. The former is when a property is defined as an `owl:ObjectProperty` and thus expects an object URI resource as its range, whilst the latter is for properties defined as an `owl:DatatypeProperty`, and hence are expecting a data literal as their range. Listing 7.3 shows the query used to retrieve both Datatype and Object properties of the selected class, as well as for its superclass.

```
SELECT DISTINCT ?property ?label WHERE {
        ?property rdf:type owl:%%Type%%.
        ?property rdfs:label ?label .
        { ?property rdfs:domain <%%Concept-URI%%> . }
        UNION { ?property rdfs:domain ?superClass . }
        {
                SELECT DISTINCT ?superClass { <%%Concept-URI%%> rdfs:
                    ↪ subClassOf* ?superClass . }
        }
    FILTER(langMatches(lang(?label), 'EN'))
}
GROUP BY ?property ?label
```

Listing 7.3: The SPARQL query used to obtain Datatype and Object properties for a specified class

In Step 3, the user can add filters to restrict the results as well as add optionals to refine the results (shown as a preview). Furthermore, users can refine their search by adding other related classes (resulting in a multiple class query).

7.1.2 Query Generation

Apart from enabling users to explore datasources, the main task of the Query Builder tool is to aid users to generate a SELECT SPARQL query, without requiring prior knowledge of SPARQL or the dataset's underlying schema. This tool enables users to generate a SPARQL query through a user-friendly interface equipped with auto-complete features. Similar to the exploration function of the Query Builder, the query building function is also enabled through the consumption of the RDF2Any API, where the user selection

for the classes and properties is converted into a SPARQL query that is then executed on the selected datasource.

In order to generate a SPARQL query, the user can follow exactly the same procedure as explained in Section 7.1.1. After selecting the datasource and class to query, the user can proceed to select the properties to be included in the resultset. Properties can be freely selected to be included or excluded from the results, and a click on a property allows the user to define a filter. Additionally, at this stage the user can also select to add the object type class of the relevant property as a new concept, hence obtaining a *multiple class query* which enables the user to add further filters on the additionally selected classes. An example of a multiple class query is when the user would like to get as results any actors born after 1900 whose nationality is a country where the official language is Portuguese (as shown in Figure 7.3). This is done by selecting Actor as the first class and adding a filter on the birth year. Then the class of the actor's nationality (Country class) is added as a second class. Finally a filter on the official language property is set to only return countries where the official language is Portuguese. Throughout the query building process, the query, which is generated on the fly, is displayed. Thus, once the user has made the preferred selections, the generated query is previewed and can be edited if this is required. Finally, the user has the option to first preview a subset of the results, and then proceed to export the full result set.

7.1.3 Data Transformation

Provided within the Query Builder Tool, the Transformation function is aimed towards users who need the resultset in a format other than RDF. This might be because their native system understands other formats, or simply because they find results in another format more easily readable and interpretable. This *softening* does indeed result in a certain degree of loss in semantics. Yet, this is compensated through retaining links with the original RDF data and other relevant information through the ConQuer Ontology, as discussed further ahead in Section 7.2.1. This means that it is always possible to obtain the original data in RDF through exploiting the provenance information recorded for every transformation and resultset.

The transformation process consists in converting the results in RDF to a number of different formats through the consumption of the RDF2Any API. Currently, the conversions provided are from RDF to CSV, JSON, and RDB, as well as a more advanced configurable conversion. The latter allows a user to convert RDF into potentially any output format, such as XML, KML, TSV (tab separated values), etc. The exception are formats which require memory storage, such as RDBMS serialisation, which requires the storing of foreign key values. The use of the Generic Conversion requires some knowledge about the dataset(s) to be converted, and the user is required to pass required parameters through a template. Apart from being easily extendible with further converters, the transformation process provides the additional advantage that a user can directly convert the required subset of the datasource in question, rather than converting a bulky data dump. We manually validated the correctness of the various conversions for various queries on different datasources. While we confirm there is a loss from the rich representation of RDF, the essence of the data is retained and the provenance information allows us to retain the link to the original data and the transformations for reproducibility.

7.2 Transformation Explorer

All the processes executed through the ExConQuer Framework generate what we call a Linked Data Publication, which is basically what users can share, re-use, explore, and edit. Thus, a Linked Data Publication consists of all the generated information, including the SPARQL query used, its description, the datasource(s) queried, the initial and target data formats, and the user generating the Linked Data

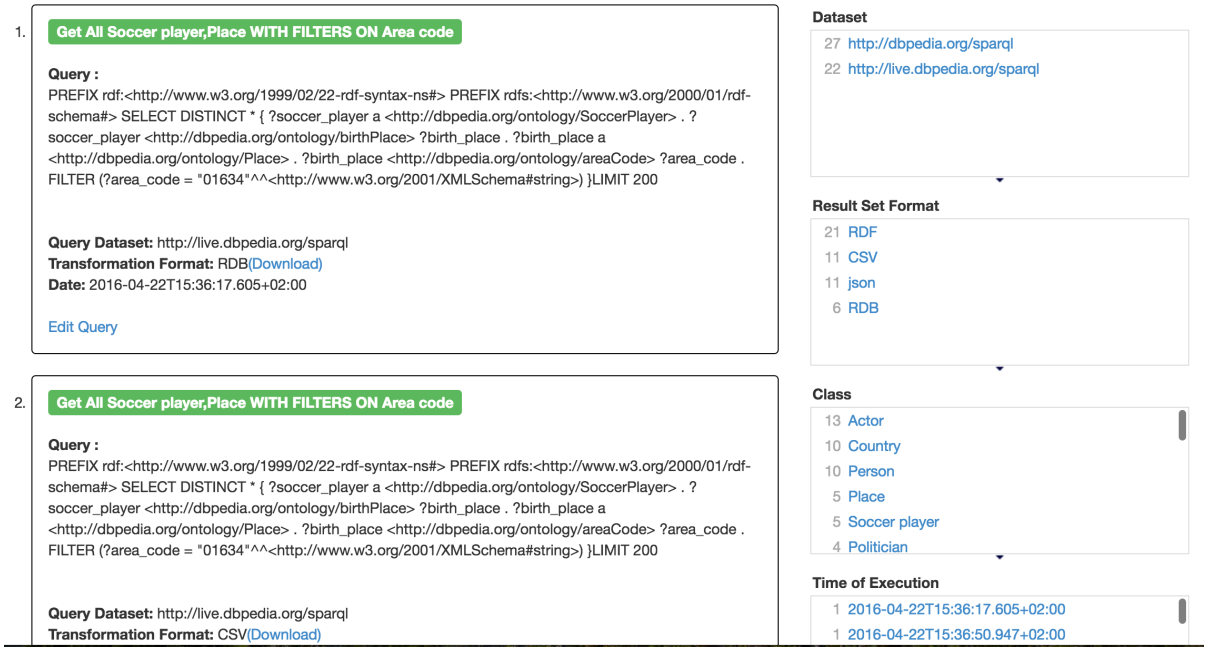

Figure 7.4: Transformation Explorer: Enables the exploration and re-use of Linked Data Publications generated through the use of the Query Builder Tool.

Publication instance[3]. We represent all this data using the *ConQuer Ontology* (see Section 7.2.1). All generated Linked Data Publications can then be explored using the *Transformation Explorer* (shown in Figure 7.4), which furthermore enables users to re-execute or edit existing queries.

The main aim of the Transformation Explorer is to provide stakeholders with the potential to explore all existing queries and transformations executed on different datasources. In this way, a user is given the opportunity to find any results that match the given requirements. Moreover, if the results are not exactly as the user requires, for example if they are in a different format, or the resulting data is too generic/specific, the user can proceed to edit or update the results with minimal effort, through re-loading the Linked Data Publication in the Query Builder Tool.

7.2.1 ConQuer Ontology

The ConQuer ontology (shown in Figure 7.5), through the represented information, not only allows us to represent all possible transformations on an entity through querying and converting, but it also allows us to replicate the resulting Linked Data Publications and edit them to achieve different results. Figure 7.6 shows how, starting from a transformation on a specific datasource (Original Transformation), a user can re-use the query but execute a different conversion on the resultset, or otherwise edit the original SPARQL query in order to obtain different (more generic, more specific, or otherwise) results. Thus, using the ConQuer ontology to represent our transformations allows us to *soften* RDF into semantically shallower formats without actually compromising on the the richness of RDF representation, as any resultsets in formats other than RDF are linked back to the original data in RDF. Additionally, through the provenance information, the ConQuer ontology allows us to track the changes to each entity, and also assign a reputation or a rating for the different agents generating the Linked Data Publications.

[3]This is not implemented in the online demo as yet, since we wanted to avoid forcing users to register and log in, in order to use the tool.

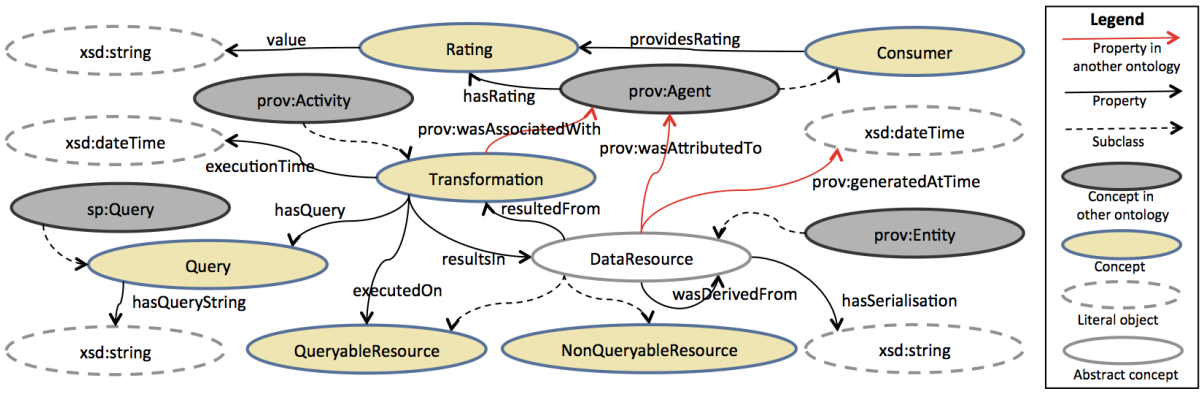

Figure 7.5: ConQuer Ontology for modelling Linked Data Publications.

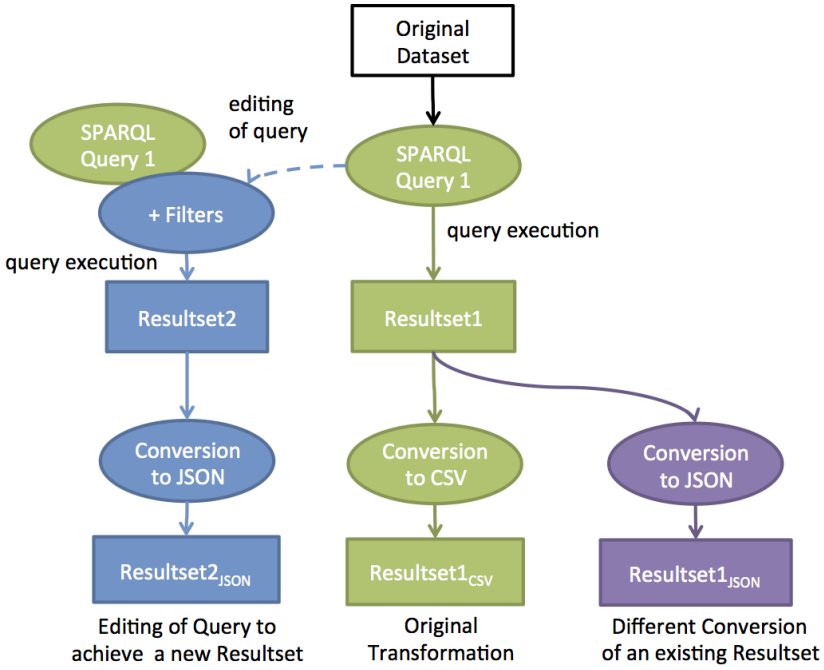

Figure 7.6: Example of possible Linked Data re-use scenarios enabled by the ExConQuer Framework and the underlying provenance-aware ConQuer Ontology.

The main concepts in the ontology are the following:

- **Transformation** - A Transformation represents all the information required to achieve a Linked Data Publication, as described above.
- **Query** - A Query represents a set of statements forming a SPARQL query.
- **Data Resource** - A Data Resource is used to represent a data store. This can be anything from a linked open dataset with a SPARQL endpoint such as DBpedia, to a database or a CSV document.
- **Agent** - An Agent is any entity, whether machine or human, that has some sort of control or authority over the generation of a Transformation instance.

To describe the ontology in an informal manner, a *Transformation* has a *Query* that is executed on one or more instances of a *DataResource* (enabling the representation of federated queries). The latter must be a *QueryableResource*, or, in other words, it should be expressed in one of the serialisations of the RDF data format (RDF/XML, NTriple, Turtle, etc.). The resulting *DataResource*, on the other hand, can be either a *QueryableResource* or a *NonQueryableResource* (formats such as CSV, PDF, etc). Finally, each *Transformation* and *DataResource* are linked through the relevant properties.

Since the ConQuer Ontology is representative of Transformations, thus making the latter class the main concept within the ontology, we define a *Transformation* T as follows:

Definition 1: $$T = \{q, d, f_d, r, f_r, a, t\}$$

where q is a *Query*, d and r are *DataResource* instances (original resource(s) and resultset), f_d and f_r are the serialisation formats of d and r respectively, a is an *Agent*, and t is the time the transformation was executed. Hence, a generates T, which represents a Linked Data Publication instance. The latter results from applying q to d and then obtaining the final Linked Data Publication by converting f_d to f_r. This means that $r \subseteq d$, as the user can query to get all, or part of resource d.

In the ConQuer Ontology we re-use concepts from the SPIN vocabulary [70], which is used to represent re-usable SPARQL queries as templates, and the PROV-O ontology [14], used to represent provenance information. The use of SPIN to represent SPARQL queries not only enables the direct querying of the queries themselves, but also allows the represented knowledge to be re-used in any frameworks or tools using the SPIN vocabulary. The re-used concepts are:

- **sp:Query** - A SPIN concept which represents a SPARQL query. This concept enables us to search within the persisted *Query* instances.
- **prov:Activity** - A PROV-O concept representing something that occurs over a period of time and either interacts with or acts upon *prov:Entity* instances. *prov:Activity* instances can include transforming, consuming, using, or generating entities.
- **prov:Entity** - An *Entity* can be physical, digital, conceptual, or any other thing with a fixed set of aspects.
- **prov:Agent** - This concept represents something or someone who bears some sort of responsibility for an *Activity* taking place or for the existence of an *Entity*.

7.2.2 Linked Data Publication Exploration and Management

We implemented the Transformation Explorer as a management tool that enables the exploration of Linked Data Publications with the aim of encouraging their re-use. The motivation behind providing such a tool is that queries are re-usable, and a single query might be the answer to many users' requirements. Besides, the Transformation Explorer also enables users to persist and re-use complex SPARQL queries. The re-use of queries is particularly useful when a dataset is frequently updated, as a user can simply re-run the query in question to get the updated results. We query the persisted instances of the Linked Data Publications and publish them through a faceted browser (Exhibit[4]). Through the use of the ConQuer ontology, the Linked Data Publications have queryable metadata that enables users to search for specific instances using various criteria, such as by the datasources used and the classes queried for. Moreover, a user would be able to search by Agent if the user is required to log in before using the Query

[4] http://www.simile-widgets.org/exhibit/ (Date accessed: 2 August 2016)

Builder Tool[5]. Through the persistence of such provenance information, users could query Linked Data Publications according to Agents who have the reputation of providing the best data for the intended use. This tool therefore allows users to share, explore, and directly edit (through the Query Builder or otherwise) and re-use Linked Data Publications, whilst keeping data lineage intact.

7.3 Evaluation

The purpose of this section is to discuss the evaluation led on the ExConQuer Framework. Our framework is intended for the use of stakeholders who are not familiar with RDF or the Linked Data paradigm. This does not exclude stakeholders who already use Linked Data in one way or another, simply because users who are not familiar with the underlying complexity are unable to exploit Linked Data to its fullest potential. For example, a user downloading a data dump of a linked dataset is hardly exploiting the potential of the data in question. For the above reasons, the framework requires to be very user-friendly and provide simple access to the required functionality, whilst also abstracting the underlying complexity. This evaluation uses two different sets of evaluators and is divided in two parts as follows:

1. A comprehensive survey intended to assess the usability of the tools (See Appendix A for complete survey); and
2. A shorter survey concerned with analysing the time and effort required to re-use open data with and without the ExConQuer Framework (See Appendix B for complete survey).

7.3.1 Usability Evaluation

Since the aim of this evaluation is to identify whether the ExConQuer framework helps or encourages people in the re-use of Linked Data, we shared this evaluation with relevant partners or colleagues who, to some extent or another, had contact with Linked Data. In total we had 27 evaluators, who, considering research such as Nielsen's [99], should be able to point out even more than the most relevant usability issues in the evaluated tools. Their domains differ in nature (such as education, healthcare, research, consulting, industry and marketing). 6 of them do not use Linked Data at all, whilst 21 use it either personally, in their work, or both (mostly for analysis, visualisation and integration). All the evaluators specified more or less the same processes while interacting with Linked Data, namely searching for existing data, accessing and gathering it, cleaning it, integrating it, leading out analyses, and consuming it by visualising it or in other ways such as data mashups. Apart from other issues, nearly all evaluators pointed out that the format of the data hindered them from re-using it, and very commonly data is also incomplete or invalid. Moreover, data might not be accessible at all. 11 out of the 27 evaluators are not familiar with the SPARQL query language, so we were able to interpret the results considering the two different target users.

For this usability evaluation we constructed a survey consisting of 25 questions and split it into three sections, namely questions on the current means and methods of accessing and using Linked Open Data (if any), questions on the Query Builder Tool, and finally questions on the Transformatin Explorer. Where relevant, we used the Likert scale [75] to assess the evaluators' perception of the tools.

Query Builder Tool

In order to have a better insight, the evaluators were asked to describe their current process of querying Linked Data. 7 of the evaluators directly specified they use SPARQL to query Linked Data. The rest

[5]This is not currently implemented in our online demo.

		Strongly Agree		Agree		Neither Agree Nor Disagree		Disagree		Strongly Disagree	
	Users of Linked Data?	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
	Query Builder										
1.	Do you agree that it was easy to execute this task?	6	1	9	6	2	2	1	0	0	0
2.	Do you agree that this tool would be useful in your SME/Company to access, explore and query open datasets?	4	2	11	5	2	2	1	0	0	0
	Transformation Explorer										
3.	Do you agree that it was easy to execute this task?	3	1	9	5	2	2	2	1	2	0
4.	Do you agree that this tool would be useful in your SME/Company to re-use saved data access queries?	6	1	6	6	3	2	2	0	1	0

Table 7.1: Four sample questions from the Usability Evaluation.

either do not use Linked Data (6), or use other methods for querying such as running test queries, using query designers, or exporting to Microsoft Excel (15). The evaluators were then asked to access the Query Builder Tool, explore a dataset, formulate a SPARQL query including filters, and download and convert the results in the preferred format. Questions 1 and 2 in Table 7.1 show the results for the evaluators' impression of the Query Builder tool. The evaluator who replied with 'disagree' in both question 1 and 2 was of the opinion that tutorials or demo videos would have been helpful with executing the given task. For question 2 in Table 7.1, almost all the evaluators (22) agreed that they would find this tool useful (to some degree or another) in their SME/Company/Academic Entity. When asked if the Query Builder is a better approach than their current way of consuming Linked Data (question is available in complete survey online), only 5 replied 'not sure' while the others all agreed that it would be better. From this part of the evaluation we can conclude that while this tool still requires some improvements with regard to usability, it is however generally deemed to be useful by the target stakeholders (both experts/non-experts, and users/non-users of Linked Data) and is an improvement on their current methods of exploiting Linked Data (if any).

Transformation Explorer

For this part of the evaluation, the evaluators were asked to use the Transformation Explorer to search for the Linked Data Publication they just created in the previous section of the survey, then re-load and edit it on the Query Builder Tool. The users were able to use a number of facets to filter the results. For this tool, the responses to question 3 in Table 7.1 were somewhat varied, however the majority of the evaluators still agreed that the tool is quite easy to use, and that it would be useful to their company. Most of the comments from the negative replies pointed out that the tool took quite long to load, and one evaluator who selected 'strongly disagree' commented that we show too many details (such as the SPARQL query). On the other hand, the other evaluator who selected 'strongly disagree' for question 3 still thought that the tool would be very useful in his context. When asked question 4, the evaluators' replies were mostly positive. Yet again, this indicates that while the tool needs improvement, mostly efficiency-wise, the majority of the evaluators still consider the tool to be useful.

7.3.2 Effort Evaluation

In this evaluation we required to analyse if the ExConQuer framework makes the open data re-use process more easy or efficient for the users. This evaluation consisted in asking the evaluators (different from

the evaluators in the usability evaluation) to execute a simple task that required obtaining some data from DBpedia, with and without the ExConQuer tools. In total we had 20 evaluators who, similar to the previous evaluation, have some contact with Linked Data but do not necessarily know SPARQL, RDF, or the datasets' underlying schema. In fact 9 of the evaluators stated they did not use SPARQL queries on a frequent basis, and two of whom did not even know anything about the querying language.

In order to determine whether the Query Builder improved on the time and effort required to obtain open data, we defined a simple task that required the users to get some data from DBpedia as follows:

Task: Get all actors whose nationality is a country where the national language is English.

The evaluators were thus required to execute this task using their usual method for accessing open data. In this evaluation, all the users attempted to use the DBpedia SPARQL endpoint, albeit showing different levels of ease and efficiency. Then, the users were required to get the same data using the Query Builder tool. As part of the evaluation, the users entered the ease with which they managed to execute the task as well as the time taken to do both tasks (separately). The results are shown in Figures 7.7 and 7.8.

For the task using the usual method for accessing open data, the users who were not so familiar with SPARQL ended up using a search engine to obtain the required SPARQL query. Figure 7.7 shows the ratings given (for both methods) to the effort required to do the given task, where 1 means 'not easy' and 5 means 'very easy'. For the users' usual method, 4 users rated the difficulty to be quite easy (rating = 4), the results of all the other 16 users ranged from neutral (rating = 3) to not easy (rating = 1). On the other hand, using the Query Builder Tool, 4 users rated the ease-of-use of the tool to be neutral (rating = 3), 10 rated it to be quite easy (rating = 4), whilst 6 rated the tool to be very easy to use (rating = 5).

With regard to the time taken, as shown in Table 7.8, only 1 evaluator (User 4) took the same time in both methods, finding both approaches equally easy to execute, whilst 2 evaluators (Users 9 and 17) took more time using the Query Builder, where one rated the ease-of-use to be equal, and the other said it is less easy to use the Query Builder. All the rest of the evaluators took less time to execute the task using the Query Builder Tool as opposed to using their usual method, namely 12 minutes and 1 second less on average (Table 7.2). One user (User 19) did not even manage to execute the given task without using the Query Builder, having no idea how to access the DBpedia datasource. Two users (Users 2 and 20) took a particularly long time in executing the task using their preferred method; about 2 hours each. Given that the users specified that they are not very familiar with SPARQL queries, we can safely assume that it is quite reasonable that they took two hours to do the task. First they required to figure out how to do a SPARQL query, which has quite a steep learning curve. Possibly, they did this through learning by example, since they only needed to do one task in this case. Then they needed to understand the DBpedia schema to identify how the required concepts are represented, before finally producing the SPARQL query which provides the required results. Being unexperienced in SPARQL, it is most probable that the users needed to do various corrections to the query before managing to obtain the correct one.

Taking into consideration the results of the effort evaluation, we can conclude that the tool enables users to more easily and more efficiently execute a data gathering task from a datasource with a SPARQL endpoint. Whilst it is not as useful for users who are very familiar with SPARQL queries, the Query Builder Tool was considered to be quite useful to introduce and teach SPARQL to users who are not familiar with the querying language (see Figure 7.9), therefore reaching the aim of lowering existing barriers to re-using Linked Data.

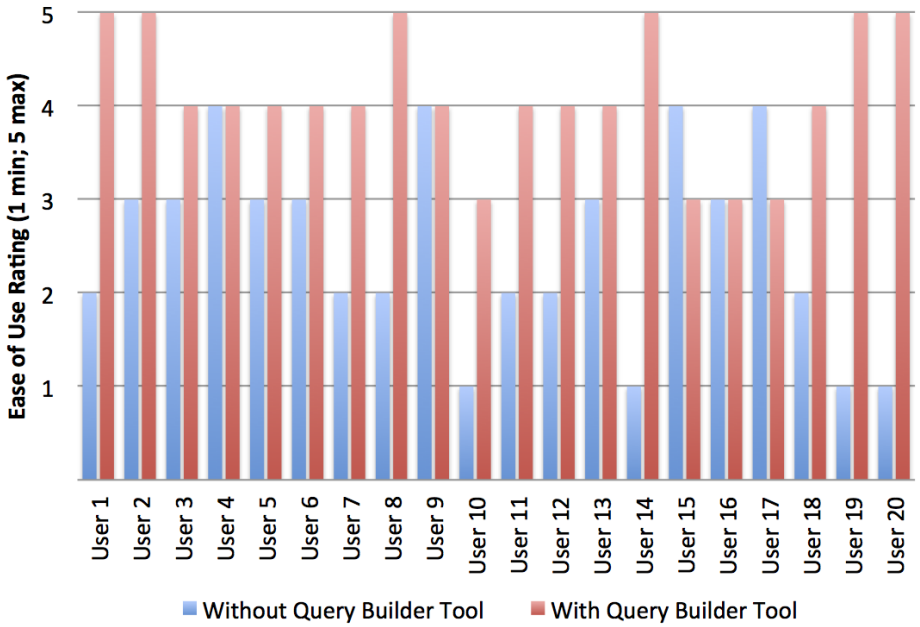

Figure 7.7: Comparison of ease-of-use rating for executing the task, with and without the Query Builder Tool (where 1 is not easy, 5 is very easy).

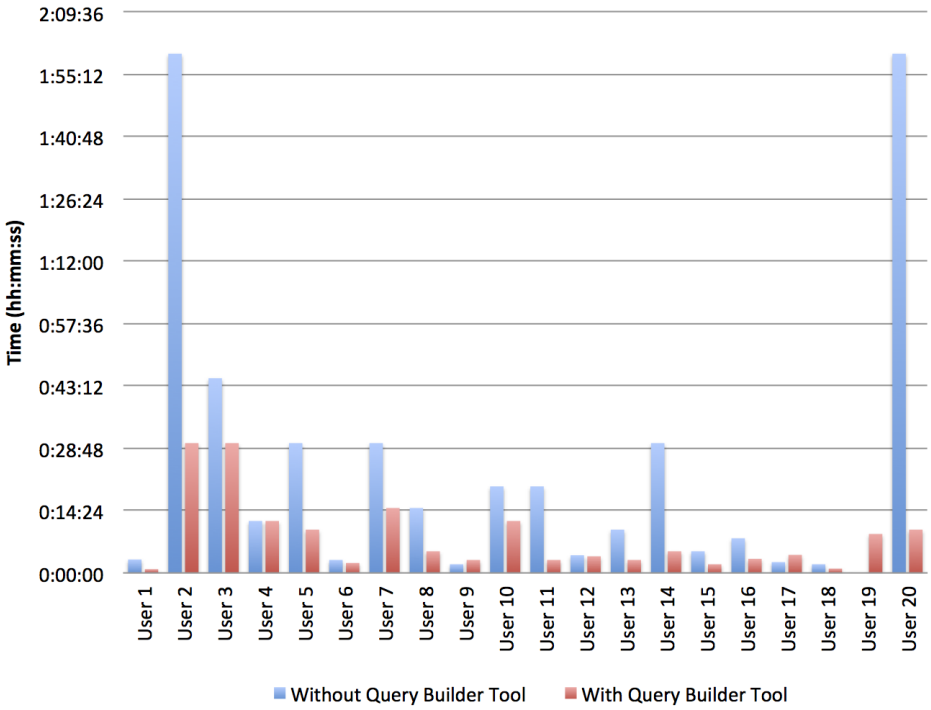

Figure 7.8: Comparison of time taken to execute the task, with and without the Query Builder Tool.

	Without Query Builder	With Query Builder
Average time	0:24:11	0:08:13
Maximum Time	2:00:00	0:30:00
Minimum Time	0:02:00	0:00:52

Table 7.2: Average, maximum, and minimum time taken to execute the task, with and without the Query Builder Tool.

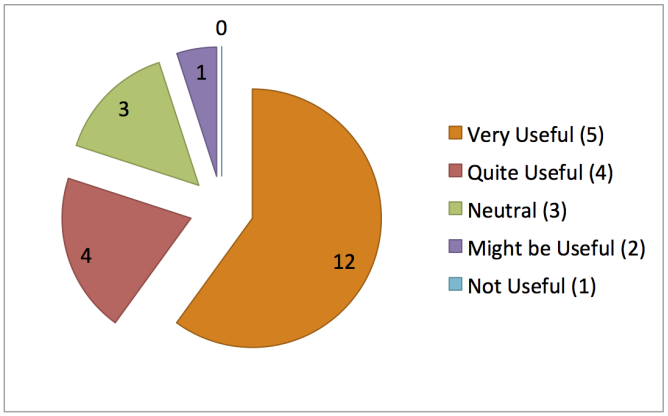

Figure 7.9: Results for rating whether the Query Builder Tool is useful to learn SPARQL.

7.4 ExConQuer in Use

The ExConQuer Framework, created as part of the LinDA Project[6], was used by a number of SMEs who participated within the project consortium, as pilot partners or otherwise. The ExConQuer tools were used in the following scenarios, using datasets that vary between open data, government data, and private data.

- A Business Intelligence scenario at Critical Publics[7], an SME headquartered in London that implements strategies to manage relationships with important stakeholders;
- A Water Management scenario, run by Hyperborea[8], an Italian Company specialising in ICT solutions for the environmental management sector;
- A Media Industry pilot, at an Italian broadcaster, TTNEWS24[9]; and
- A Media Industry pilot led by Piksel[10] (Italian Branch), a company specialised in providing holistic solutions for management of post production scripts and providing advanced media analytics.

Along with other LinDA tools, the ExConQuer framework is also being endorsed in a number of other initiatives, projects, or SMEs, including but not limited to the following. Other collaborations are listed on the LinDA website[11].

[6] http://linda-project.eu/ (Date accessed: 2 August 2016)
[7] http://www.criticalpublics.com/ (Date accessed: 2 August 2016)
[8] http://www.hyperborea.com/ (Date accessed: 2 August 2016)
[9] https://www.facebook.com/ttnews24/ (Date accessed: 2 August 2016)
[10] http://www.piksel.com/ (Date accessed: 2 August 2016)
[11] http://linda-project.eu/linked-projects/ (Date accessed: 2 August 2016)

- ODINE[12] - An open data incubator where more than 500 SMEs have applied to date;
- Your Data Stories[13] - A project that deals with finding, analysing, and visualising open data;
- Infamous Labs[14] - A software development company that provides high quality technical services related to Smart TVs;
- Open Aire[15] - A large-scale initiative that aims to promote open scholarship and improve the discoverability and re-usability of research publications and data;
- Suite5[16] - An SME working on transforming data streams from multiple sources to analytics and intelligence; and
- Weather ex Machina[17] - An SME providing a weather forecasting service based on data aggregation.

Apart from the above initiatives, the ExConQuer Framework is also being exploited directly on DBpedia[18] as a query builder tool and SPARQL query interface.

[12] https://opendataincubator.eu/ (Date accessed: 2 August 2016)
[13] http://yourdatastories.eu/ (Date accessed: 2 August 2016)
[14] http://www.infamouslabs.net/ (Date accessed: 2 August 2016)
[15] https://www.openaire.eu/ (Date accessed: 2 August 2016)
[16] http://www.suite5.uk/ (Date accessed: 2 August 2016)
[17] http://weatherxm.com/ (Date accessed: 2 August 2016)
[18] http://wiki.dbpedia.org/projects/exconquer (Date accessed: 2 August 2016)

Concluding Remarks for Part III: Lowering Barriers to Open Data Re-Use

In Part III we explore current methodologies for exploiting open data with the aim of answering the following research question:

Research Question 2:

How can we enhance the consumption process of a data product in order to enable further value creation?

Stimulated by the open data movement, open data use is becoming more and more prevalent in all dimensions of society. Linked Open Data, a subset of open data, has become a popular way to publish data, and as indicated through the exponential growth of the Linked Open Data Cloud, it is evident that the use of Linked Data principles to publish data is increasing at a fast rate. This increase is also reflected in tools aiding users in the publishing process, yet, tools aiding users to consume and re-use Linked Data are still not that prevalent. This has the consequence that users who are not familiar with Linked Data are hindered from exploiting open data published using Linked Data principles. After researching current approaches, we therefore propose the ExConQuer Framework. This framework targets existing challenges that hinder non-experts in consuming Linked Data and provides an easy-to-use solution.

In Chapter 6 we provide an overview of our research motivation. After providing some preliminary information on Linked Data, we proceed to investigate current methods for open data consumption, including Linked Data exploration systems, SPARQL query builders, and data transformation and exploration systems. Through this research we identify a niche with regard to approaches that abstract the complexity beneath exploiting linked datasets; existing tools all require a certain degree of background knowledge that non-experts usually lack.

In Chapter 7 we propose the *ExConQuer Framework*. In order to provide more simple and workable views of the data, in this framework we transform RDF data into a number of different formats, whilst still preserving the semantic richness of the RDF data model. While during this process we do lose some of the richness of RDF representation, we compromise by preserving the link with the original RDF data through the ConQuer ontology, and still retain the full semantic richness through provenance information. As is evident through the evaluation we performed, the ExConQuer Framework is particularly useful to encourage the re-use of Linked Data by stakeholders who are not familiar with RDF, and are more acquainted with formats such as JSON or CSV. Our framework is also useful for more expert users who are however not able to exploit Linked Data to its full potential due to not being familiar with RDF, SPARQL or the data's underlying schema.

Through the contributions in Part III we hence encourage and enhance the data consumption process. Consumers are able to more easily access, re-use, and innovate upon previously-unaccessible data to participate in ensuing value creation processes.

Part IV

Value Creation as an Exploitation Strategy

With the aim of guiding and enabling stakeholders to exploit data to its fullest potential, in this part we focus on the value creating processes that can be used to enhance data products. We start with Chapter 8 by providing an overview of related literature and discussions about our motivation. We proceed by describing one of the main contributions in this part; the Data Value Network. In Chapter 9 we hence cover the various value creating processes within the network, as well as the different actors and the roles they can participate through. We also identify the challenges and resulting impacts of creating value. As a possible solution for the identified challenges we propose the use of Linked Data as a basis for value creation. We finally explore a number of use case scenarios and detail the value creating processes within. In Chapter 10 we propose the Value Creation Assessment Framework; a framework that stakeholders can use to determine the value creating *potential* of open data initiatives. With the aim of portraying how stakeholders can participate in the global data market through creating value, in Chapter 11 we propose the Demand and Supply Distribution Model. This model provides insight on how data producers/publishers and consumers can collaborate and exploit data products. In this chapter we also provide a concrete implementation of this model through a service.

The chapters in this part are based on the following publications:

- **Judie Attard**, Fabrizio Orlandi, Sören Auer. *Value Creation on Open Government Data*. In Proceedings of the 49th Hawaii International Conference on System Sciences, HICSS 2016, Koloa, Hawaii, USA, January 5-8, 2016.
- **Judie Attard**, Fabrizio Orlandi, Sören Auer. *Data Driven Governments: Creating Value through Open Government Data*. In Proceedings of the Transactions on Large-Scale Data- and Knowledge-Centered Systems Journal, 2016.
- **Judie Attard**, Fabrizio Orlandi, Sören Auer. *Data Value Networks: Enabling a New Data Ecosystem*. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Omaha, Nebraska, USA, October 13-16, 2016.
- **Judie Attard**, Fabrizio Orlandi, Sören Auer. *Exploiting the Value of Data through Data Value Networks*. In Proceedings of the 10th International Conference on Theory and Practice of Electronic Governance, ICEGOV, 2017.

CHAPTER 8

Value Creation and Data Value Chains

In our information society, data becomes increasingly a commodity and the basis for many products and services. Examples are Open Data, Linked Data or Big Data applications and services, such as government data portals[1], reviews, feedback, and product suggestion on e-commerce websites, weather emergencies forecast[2], patient monitoring[3], citizen participation and decision-making[4], etc. All data, whether addresses of schools, geospatial data, environmental data, weather data, transport and planning data, or budget data, has social and commercial value, and can be used for a number of purposes that could be different than the ones originally envisaged. Governments are one of the largest producers and collectors of data in many different domains [2, 58]. By publishing such data the government encourages stakeholders to innovate upon it, and create new services. The main challenge in releasing social and commercial value is that open data does not have intrinsic value, yet it becomes valuable when it is used [59], and there are many factors within an open government initiative that influence its success.

The value chain model describes value-adding activities that connect an industry's supply side to its demand side. The value chain model has been used to analyse and assess the linked activities carried out within traditional industries in order to identify where, within these activities, value is created. This was done with the aim to identify what activities are the source of competitive advantage within these industries. As successful as the value chain concept was to achieve this aim, during these last years products and services are becoming increasingly digital, and exist in a more non-tangible dimension [107]. In addition, the traditional value chain model does not consider when information is used as a source of value in itself [115]. Thus, the original concept of value chain is becoming an inappropriate method with which to identify value sources in today's industries that produce non-tangible products [107].

In recent years, in order to reflect this dataification [29], the concept of *data value chains* was introduced, building upon the concept of traditional value chains for tangible products [111]. The rationale of a data value chain is to extract the highest possible value from data by modifying, processing and re-using it. Value can be added to the generated raw data to make it re-usable, and thus a product within itself. The exploitation of this data with added value has the potential to feed a chain of innovative information products and services, making the data value chain the centre of the knowledge economy. Any traditional sector, such as health, transport, or retail, can thus benefit from new-found opportunities based on digital developments.

[1] https://open-data.europa.eu/en/data/ (Date accessed: 2 August 2016)
[2] http://centrodeoperacoes.rio/ (Date accessed: 2 August 2016)
[3] http://www.immunizeindia.org/ (Date accessed: 2 August 2016)
[4] https://www.fixmystreet.com/ (Date accessed: 2 August 2016)

Within an urban environment such as a city, the data value chain can have major impacts on the citizens, especially where a data product is used in a decision-making process and is a source of value in itself. The *decision-making process* is a very broad term used to encompass the practice of familiarising oneself with the relevant information before taking a particular decision. This concept was discussed as early as the 1970s, where Montgomery [95] describes the use of information systems to aid in the planning and decision-making processes within a marketing environment. Often used within a *smart city* environment, this decision-making process is becoming more popular and coincides with an increasing effort worldwide to transform cities into smart cities, particularly through the release of government data to the public, as well as through the exploitation of this data. Examples include Rio de Janeiro in Brazil[5], Dublin in Ireland[6], and London in the United Kingdom[7]. Examples of the impact of value creation on decision-making processes within urban environments include the following:

- Transportation - The analysis of traffic data can aid citizens to check the best time to use certain roads, public transport can be better managed through better prediction of arrival times, whilst the government can attempt to lessen traffic by providing alternative transportation options. For example, a live view of the car boarding areas for the ferry between the islands of Gozo and Malta is streamed[8] in order to enable citizens to check if there is currently a long queue and plan their travels accordingly. Moreover, traffic supervisors can be dispatched to control and manage the boarding process.

- Energy Consumption - The use of smart meters and other sensors can help in reducing energy consumption through monitoring use in real-time. For example, an initiative throughout the European Union is currently ongoing with the aim of controlling energy consumption and providing for a more sustainable environment[9].

- Weather Emergencies - Weather information can be used to predict if a weather-related emergency is incumbent, such as flooding, landslides, earthquakes, etc. This prediction can be used to issue warnings or evacuation orders in time. The city of Rio de Janeiro is a good example of this use case, as an operations centre[10] was established with the aim to prevent weather-related disasters (amongst other aims).

- Health - Patient data can be used to generally monitor a patient during an ongoing treatment or to issue reminders when check ups or vaccinations are due. The Immunize India initiative[11] is an example of the latter.

In this part of the thesis our aim is to identify the best value chain specifically suitable for a data product. This will enable us to provide guidelines for value creation which will enable the full exploitation of open data. We are hence answering the following research question, as defined in Section 1.2:

Research Question 3:

What aspects and processes play a role in value creation on a data product?

[5] http://www.centrodeoperacoes.rio.gov.br/ (Date accessed: 2 August 2016)
[6] http://www.dublinked.ie/ (Date accessed: 2 August 2016)
[7] http://citydashboard.org/london/ (Date accessed: 2 August 2016)
[8] http://www.visitgozo.com/en/content/live-ferry-queue-streaming-beta-1538/ (Date accessed: 2 August 2016)
[9] http://my-smart-energy.eu/ (Date accessed: 2 August 2016)
[10] http://centrodeoperacoes.rio/ (Date accessed: 2 August 2016)
[11] http://www.immunizeindia.org/ (Date accessed: 2 August 2016)

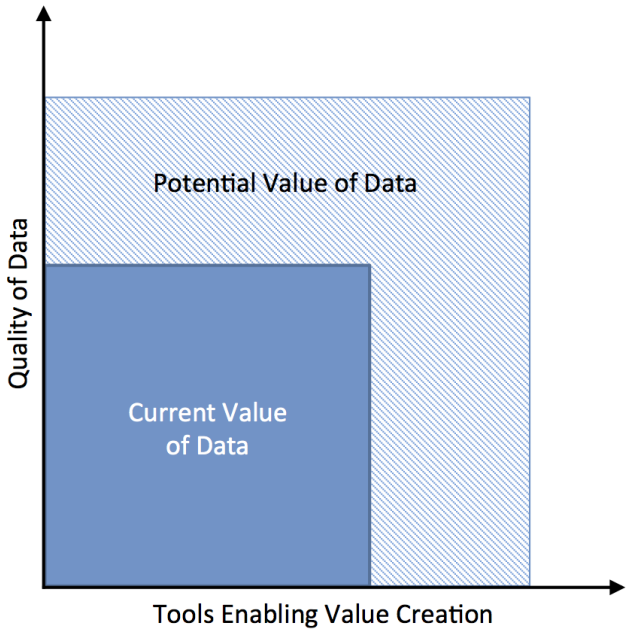

Figure 8.1: The potential increase in value of data through value creation.

We start by analysing various implementations of data value chains (Section 8.1) and build upon previous definitions with the aim of providing the best characterisation. The specification of the tasks and roles within this value chain will then aid us to define the best guidelines for participating stakeholders, who can align their contribution within the value chain accordingly. By delineating the resulting impacts of the value creation process we can then have a better perspective on why the value creation process is vital in our data economy. Finally, by measuring the value potential of a data product, we can therefore maximise the benefits of exploiting such data. Figure 8.1 shows (in an abstract manner) how the creation of value, through increasing the quality of data and the creation of tools, can increase the potential value of data.

8.1 Background and Related Work

The process of *Value Creation* is a somewhat subjective concept that depends on a consumer's perception on the usefulness of the product, the amount they are willing to pay for it, as well as the actual amount spent in such a transaction [96]. Many works in literature focus on various aspects in the context of value creation, such as how to best exploit value creation to achieve economic benefit [3], discussions on value creation within specific domains, such as mobile commerce [12], open data [128], and e-government [155], or how to achieve competitive advantage [68, 111, 112].

8.1.1 Traditional Value Chains

The term *Value Chain* was first introduced by Porter [111] in 1985 to identify how value is created in order to achieve a product. Porter defines a value chain to be the strategically relevant interdependent activities undertaken by a firm in order to achieve its goal. The value chain model hence describes value-adding activities that connect an industry's supply side, such as raw materials and production processes, to its

demand side, such as sales and marketing. The activities are physically and technologically distinct activities that are the building blocks by which a firm creates a product valuable to its buyers. The value chain can be considered as a tool that enables the analysis of the interactions between the different activities in order to identify the sources for competitive advantage, or, in other words, how and where the value is created. The activities within a value chain can be classified into five categories [111]:

1. *Inbound Logistics* - Activities related to receiving, storing, and disseminating inputs to the product; for example, the handling of materials and warehousing;
2. *Operations* - Activities associated with using the input and transforming it; for example, assembling materials or packaging;
3. *Outbound Logistics* - Activities associated with collecting, storing, and distributing the product to consumers; for example, the warehousing of the finished products and delivery operations;
4. *Marketing and Sales* - Activities associated with encouraging consumers to buy the product and providing the means to do so; for example, advertising the product, and creating a sales force; and
5. *Service* - Activities associated with providing services to enhance or maintain the value of the product; for example, installation and repair of the product.

Porter [111] states that there might be different value chains for each role in a more generic value chain. For example, different airline providers have different procedures for embarking operations or crew regulations, even though such entities all participate in the generic value chain through the airline industry. Furthermore, the value chain can also differ within the same entity, where, for example, the entity has different products that cater for different target markets.

8.1.2 Data Value Chains

As successful as the value chain concept was to achieve its aims, during these last years products and services are becoming increasingly digital, and exist in a more non-tangible dimension [107]. In addition, the traditional value chain model does not consider when information is used as a source of value in itself [115]. Thus, the original concept of value chain is becoming an inappropriate method with which to identify value sources in today's industries that produce non-tangible products [107]. Nowadays, in a digital data-centric world, the cost of processing data has drastically decreased, and the access to data from multiple sources such as networks, sensors, and the Internet, has skyrocketed the availability of data. Coupled with the dramatic decrease in the cost of data storage, this is enabling huge datasets to be generated or captured, stored, and processed. Newer definitions of the value chain concept, such as in [28, 72, 74, 93, 107], cater for these digital dimensions, taking into account factors and activities which set this dimension apart from the more physical one.

Building upon Porter's definition, Lee and Yang [74] define the *Knowledge Value Chain*. Their definition differs to Porter's in that the end product is not tangible, and they define a value chain for knowledge, including the knowledge infrastructure, the process of knowledge management, and the interaction between the required components that result in knowledge performance. Knowledge, a step further than information, is data organised in meaningful patterns. The process of reading, understanding, interpreting, and applying information to a specific purpose, transforms information into knowledge. This means that for an entity that is unable to understand knowledge, the knowledge is in fact still only information. This is the *data literacy* problem, where any effort invested in knowledge generation is lost if the target consumer is unable to actually understand the provided knowledge [133]. Similar

to Porter, Lee and Yang classify the activities within the knowledge value chain in five categories, namely knowledge acquisition, knowledge innovation, knowledge protection, knowledge integration, and knowledge dissemination. While essentially similar to Porter's definition, these categories are different in that they classify activities specifically involving a non-tangible information product. Yet, Lee and Yang's definition, whilst it is focused on a data product, only considers value creation to be based on the creation of knowledge.

In [28], Crié and Micheaux provide us with a more generic value chain than Lee and Yang, including raw data in their definition. Within their paper, the authors aim to highlight any issues within the value chain, to provide an overview of the current progress, and also to encourage entities to view the benefits of participating within the data value chain. They focus on four aspects of the *Data Value Chain*, namely:

- *Obtaining the right data* - Capturing the right data is the first step to forming an information chain that aims to provide the best customer service and result in profits;
- *Data quality management* - Ensuring the data is of good quality increases the potential towards maximising returns from the data for both the entity and its customers;
- *Deriving information and knowledge from raw data* - The act of extracting information from data, and interpreting knowledge from information; and
- *Using information and knowledge to satisfy customers and generate profits* - The use of good data increases the chance of making better decisions.

Alas, Crié and Micheaux still focus on a 'chain' structure which does not reflect the flexible value creation that is really possible on data products.

Peppard and Rylander [107] also discuss a value chain that is more suited where the product in question is digitised, and thus non-tangible. The authors introduce the concept of *Network Value*, where value is created by a combination of actors within the network. In contrast to the earlier definition of a value chain, network value does not necessarily follow a linear model, and accounts for the various interconnected actors that work together to *co-produce* value. While these actors or entities should be able to function independently, they operate together in a framework of common principles. This means that an action by a single entity can influence other entities within the network, or otherwise require further actions from them in order to achieve the final product. Morgan et al. [96] provide a similar discussion on the co-production of value through open-source software. Whilst these definitions improve on the former by defining the co-participation of actors in Network Value, they do not focus on value creation on a data product.

In line with more recent popular themes, Miller and Mork [93] and Latif et al. [72] focus on big data and Linked Data respectively. Miller and Mork discuss the data value chain concerning all required actions in aggregating heterogeneous data in an organised manner and creating value (information/knowledge) that can influence decision-making. The authors divide their data value chain in three main categories, namely data discovery, integration, and exploitation. In contrast, Latif et al. propose the *Linked Data Value Chain*. Motivated by the still limited commercial adoption of the Semantic Web, the authors aim to drive the Semantic Web and the use of Linked Data closer to commercial entities. The authors discuss the entities participating in the Linked Data value chain, their assigned Linked Data roles, as well as the types of data processed within the chain. An interesting aspect that distinguishes the proposed Linked Data value chain from the ones previously mentioned is that actors within the chain are not necessarily bound to one specific role. Rather the assignment of roles to entities is more flexible where, in extreme cases, an entity can even occupy all roles at once. Similar to previously-discussed literature Latif et al.

stick to a strict chain structure and only consider value creation to be the evolution of raw data to Linked Data and human-readable data. Miller and Mork, on the other hand, provide the most in-depth definition of value creation upon a data product, however they still stick to a rigid chain structure.

CHAPTER 9

Redefining Value Chains

In this chapter (and the following chapters in this part) we target the specific domain of *data* value chains. In Chapter 8 we identified the lack of literature that discusses the actual processes used to create this value. Moreover, rather than limiting our discussion to the economic impact, we identify a number of impact dimensions that influence, or are influenced by, value creation. Using a methodology similar to defining a life cycle, we improve upon existing representations and hence define the *Data Value Network* (DVN), where we also identify the different roles and activities within the network. Our definition caters specifically for non-tangible data products, thus we focus on the aspects specific to data that differ from the definitions of value chains in literature.

9.1 The Data Value Network

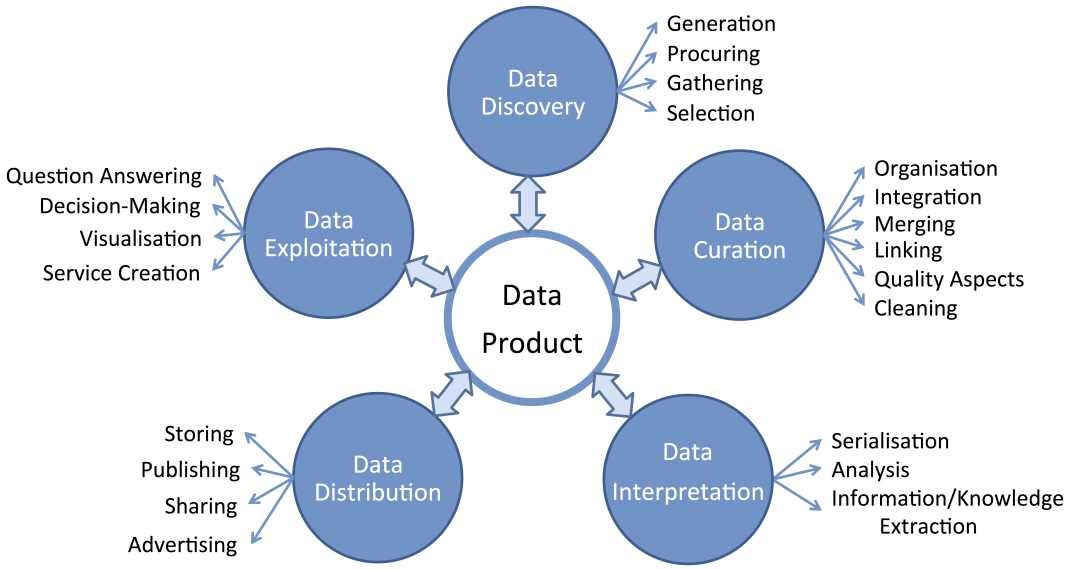

Figure 9.1: The Data Value Network (Activities and Value Creation Techniques).

After considering existing value chain definitions, and comparing them to real-life data value chains and the contained activities and roles, we define a **Data Value Network** (DVN) as shown in Figure 9.1.

Similarly to a life cycle, the DVN maps the ongoing processes through which value is created upon a data product. We included common activities executed on data products, where the final aim is usually the consumption of the data product. Due to the differing order of executing the relevant activities, a star network with the data product as its central node was deemed to be the best way to represent the interactive nature of adding value to data products. We define a DVN to be:

Definition 2: *A set of independent activities having the aim of creating value upon data in order to exploit it as a product*

where different **actors** (e.g. data producers, data consumers) can participate by executing one or more **activities** (e.g. Data Discovery, Data Exploitation), and each activity can consist of a number of **value creation techniques**, (e.g. Gathering, Visualisation, Service Creation). In turn, each value creation technique can consist of one or more **data value chains**, since they might need a series of processes to be executed in order (e.g. visualisation requires identifying the data to visualise, then deciding on a visualisation method, then rendering the visualisation). The value creation techniques will be described in more detail in Section 9.3.

The DVN has the following features which characterise it and distinguish it from other existing value chain definitions:

- **Non-Tangible Data Product** - While it can become outdated, the exploitation or use of data will *not deplete it*. Data can be re-used over and over, even for different purposes than the one it was originally planned for.
- **Non-Sequential** - The DVN does *not necessarily follow a sequential structure*, rather, any activity can follow, or precede, any other activity. Activities can be executed in tandem, and other activities can be skipped or repeated.
- **Multiple Actors** - One or more actors can participate in a DVN to produce value within an activity. Actors can also collaborate in order to *co-produce* value.
- **Nested Value Chains** - Each activity can be broken down into further, more specialised value creation techniques, each of which can be a data value chain within itself. For example, the *Data Discovery* activity can include both the *procurement* of data from a different entity, and the *generation* of new data specifically for the required purpose. In turn, the generation of new data is made up of a specific number of processes, hence making up a *data value chain*.
- **Recurring Value Network** - As opposed to the value chain which ends with the consumption of a product, the DVN can *recur as long as the data in question is still relevant*.
- **Independent Activities** - The activities and value creating processes are not interconnected, and can exist *independently*. Nevertheless, the output from one action or value creating process can act as an input to another.

Figure 9.2 shows how a data product can evolve over time under the execution of different activities led out by different actors. It is important to note that each data product in the diagram forms the core node in a new instance of a DVN, since an activity on a data product could result in a new version of the data product (e.g. by organising a dataset), or even a new data product (e.g. through knowledge extraction or merging). Then, the evolution of a data product will eventually result in a 'branching' out of various value-added versions, similar to the branching out of D_2 to two different versions; D_3 and D_4.

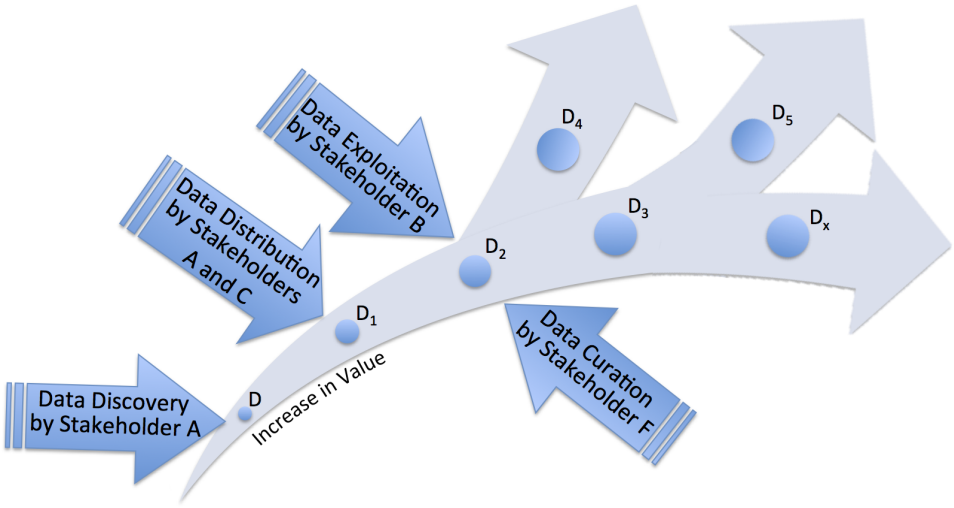

Figure 9.2: Tree structure of an evolving data product D, with interaction from different actors.

9.2 The Data in a Data Value Network

In order to be used within a DVN, data does not require any specific characteristics. It can be raw, previously-processed, machine or human readable, and it can also be data regarding any domain, such as statistical, financial, geographical, demographic, societal, etc. There can also be a variety of sources for data, such as sensors, mobile devices, applications, services, or social network profiles. Unfortunately, while data is abundantly available, there are also a number of obstacles which hinder its exploitation. Data is of a *heterogeneous nature*, with a plethora of formats, models, and schemas. This makes it very challenging to integrate, even if it belongs to the same domain. Furthermore, data is usually available from a *variety of sources* which can also be in isolated data repositories, or otherwise privately-owned spaces.

Data quality is an aspect of data that has a major effect on the value creation potential of the data itself. Due to the subjectiveness of data quality [101], it is very challenging to assess the quality of data, which can depend on a myriad of characteristics. Furthermore, there is no single agreed-upon definition of quality [71] due to the cross-disciplinary nature of the concept. However, data quality is commonly perceived to capture *fitness for use* [65], which in turn is a multi-dimensional concept that has both subjective perceptions and objective measurements based on the data in question [109]. Subjective data quality assessments reflect the requirements and experiences of the consumers of the data, whilst objective assessments can be both task-independent and task-dependent. The former metrics reflect the data itself, without any contextual knowledge of the application at hand, whilst the latter includes aspects that are specific to the application context.

Ochoa and Duval [100] propose a set of metrics to identify metadata quality, based on parameters used for human reviewing. Reiche and Höfig [116] build upon these metrics, adapting them for assessing the quality of the actual data, rather than the metadata. Similarly, in [71] and [79], the authors discuss a number of quality dimensions, as found in the majority of related literature. We here establish the following criteria which are considered by most efforts in the literature for calculating data quality aspects that aid, or hinder, an actor during a data-based, value creation process. Table 9.1 portrays the activities within the DVN that are impacted by each quality aspect.

	Data Discovery	Data Curation	Data Interpretation	Data Distribution	Data Exploitation
Usability		✓	✓	✓	✓
Data Format	✓	✓	✓		✓
Data Ambiguity	✓	✓	✓		✓
Data Discoverability	✓				
Data Representation	✓	✓	✓		✓
Provenance	✓	✓			
Accuracy		✓	✓		✓
Completeness		✓	✓		✓
Consistency		✓	✓		✓
Timeliness	✓			✓	✓
Accessibility	✓	✓	✓		✓
Openness	✓	✓	✓	✓	✓

Table 9.1: The impact of each data quality aspect on each Activity in the Data Value Network.

- **Usability** - This is the most generic quality criterion that generally regards how easily can the data be used.
- **Data Formats** - Data that is available in a machine processable data format which is non-proprietary usually provides the highest potential for value creation. However, in some cases such as for decision-making, data might be most useful if it is provided through visualisations or stories which portray aggregated data.
- **Data Ambiguity** - The use of more expressive data formats such as RDF are generally preferred, simply because they are more descriptive of the actual data they represent. This decreases the risks of ambiguity and misinterpretations [87].
- **Data Discoverability** - This aspect regards how easily existing data can be discovered by potential stakeholders. The discoverability of open data is bound to the quality of the metadata describing the data itself, which is not always complete or accurate [27, 71, 87, 116].
- **Data Representation** - Since in our information society data is very diverse and heterogeneous, its proper representation with interoperable models allows easier aggregation and integration.
- **Provenance** - Provenance refers to preserving details about the origins and processing of data, or, in other words, by whom and how the data was created, generated and processed.
- **Accuracy** - By accuracy we mean the extent to which a data/metadata record correctly describes the respective information [71, 87, 116].
- **Completeness** - This quality dimension deals with the number of completed fields in a data/-metadata record [100, 116, 136].
- **Consistency** - The consistency of record fields depends on whether they follow a consistent syntactical format, without contradiction or discrepancy within the entire dataset [71, 82].
- **Timeliness** - By this quality dimension we mean the extent to which the data is up to date.
- **Accessibility** - As identified by Ochoa and Duval in [100], the accessibility quality dimension has two measures, namely how easy it is for a data consumer to understand the published information (cognitive accessibility), and how easily a dataset can be discovered within the platform it is published (psychological/logical accessibility).

- **Openness** - As Kučera et al. [71] point out, open data can be technically defined to be open if it is available as a complete set in an open, machine readable format, at a reasonable price which is not more than the cost of reproduction.

9.3 Value Creation Techniques

Table 9.2 shows the various Value Creation Techniques within the DVN. While not comprehensive, we included the most popular and frequently-used techniques from various stakeholders participating in the DVN. The aim of all these techniques is to create or improve upon a data product, resulting in data that is (more) ideal to be used in the required application and increasing its value and re-use potential.

Data is produced in the day-to-day administration of a governing entity. The simple **generation** of this data is the first step towards its (re-)use as a data product. As opposed to data generation, data **procurement** involves obtaining data generated by a different entity through performing some sort of negotiation. Data **gathering**, on the other hand, refers to the aggregation of data from different entities or locations. Finally, data **selection** requires the stakeholder in question to choose a subset of available data and extract it, potentially with the aim of removing sensitive data. In order for the best value potential, all generated, procured, gathered, or selected data needs to be complete. This means a record has all the information required for an accurate representation of the described data.

The value creation techniques falling under the Data Curation activity have the purpose of making the data more usable. Data **organisation** requires the structuring of data in such a way that the data is more understandable, or that the data follows some pattern; for example government budget data can be organised by year. Data **integration** has the purpose of enriching an existing dataset with new data. For example, the integration of weather data to accident information can be done by an insurance company to

Government Data Life Cycle Processes (Activities)	Value Creation Techniques
Data Discovery	Generation Procuring Gathering Selection
Data Curation	Organisation Integration Merging Linking Quality Improvement Cleaning
Data Interpretation	Serialisation Analysis Information/Knowledge Extraction
Data Distribution	Storing Publishing Sharing Advertising
Data Exploitation	Question Answering Decision-Making Visualisation Service Creation

Table 9.2: Value Creation Techniques categorised according to the Data Value Network.

check the legitimacy of a claim. Another example is adding user feedback to product data in order to identify product faults. Data **merging** is somewhat similar, where different datasets are merged in order to obtain further information. For example, the merging of population data with geographical data can be used to obtain population density. On the other hand, the **linking** of different datasets is done in order to provide context, for example linking geographic data to textual descriptions about the locations in question. Finally, the **quality improvement** value creation technique represents the assessment and (if necessary) improvement and **cleaning** or repairing of data, such as removing duplicate data, ensuring the data is consistent, complete, timely, and trustworthy, and adding provenance data. This technique gives the data a higher level of quality and encourages its re-use. Similarly, metadata also enhances a dataset's re-use potential. By enriching a dataset's metadata, a dataset is made more easily discoverable by potential users [116].

The Data Interpretation activity involves some sort of reasoning where the data in question is made more understandable. In the simplest way, data **serialisation** involves the conversion of data into semantically richer or lower formats, such as PDF to RDB, or CSV to RDF. This conversion enables stakeholders with different backgrounds to still be able to exploit the data in question to its highest potential. Moreover, the use of non-proprietary, machine-readable formats will increase the value creation potential of the data in question. The implementation of **analysis** techniques, such as data mining, pattern identification, and trend analysis, enables stakeholders to identify any existing patterns, which can eventually aid actors in the DVN in value creation techniques such as decision-making. **Information/knowledge extraction** has a similar purpose, where raw data is interpreted manually (non-machine), and along with the available context information and the knowledge from the stakeholders in question it can be used to arrive to particular conclusions.

Techniques such as storing, publishing, advertising, and sharing, all have the purpose of adding the potential of the data to be distributed to different entities and re-used. The **storing** of data enables actors to re-use the data in question without requiring a local copy. By **publishing** the data in an open manner, **advertising** it, and making it **shareable**, it is also made available to many more external stakeholders. This publishing process creates value simply by making data available for re-use. The data distribution activity is a vital node within the DVN, as data that is not made available publicly is very limited in its re-use potential. Therefore, data that is provided in a timely manner (data is provided in a reasonable amount of time after creation/generation), without discrimination on its consumers (not requiring any registration), and made accessible for all, has the best value creation potential. Moreover, the addition of metadata enables the data to be more discoverable, thus enhancing this potential. Popular methods of publishing data include SPARQL[1] endpoints and Application Program Interfaces (APIs). Licensing is also vital here, as it has the purpose of declaring if and how data can be used. It is always preferable (at least from a consumer's perspective) that licences are of an open nature.

The Data Exploitation activity encompasses any value creation technique that involves consuming the data to solve a particular problem. **Visualisation** can be considered as an example of *passive* exploitation, where an actor consumes the data as information or knowledge. Visualisations involve a visual representation of data that, similar to data interlinking and data analysis, can provide us with a new insight. Visualisations can also be used to provide 'stories', since they are more easily interpreted than raw data. An example of a more *active* consumption of the data can be the use of data to influence **decision-making**, for example, a government might consider citizens' feedback before taking a decision. **Question answering** and **service creation** are other examples of active consumption of data. In the former data is collected and analysed in order to solve a specific question, whilst service creation is the provision of a service through the use of existing data, for example a public transport timetable mobile

[1] http://www.w3.org/TR/rdf-sparql-query/ (Date accessed: 2 August 2016)

application.

9.4 Actors' Roles in a Data Value Network

In this section we identify the various roles that an actor can adopt to participate within the DVN. The actors in question can be both beneficiaries and also contributors within the DVN, and are equivalent to the stakeholders who participate in open government initiatives, as defined in Section 3.3. An entity, be it a government entity, a public entity, a citizen, a company, or an enterprise, can participate as an actor within the DVN through one or more roles. This flexibility between the undertaken roles is due to the adaptive nature of the data, where value can be added in a number of different ways, as discussed previously. Figure 9.3 shows the activities that are involved in each role. Through the actors' interaction or collaboration, the outcome of a DVN is a data product of any kind or shape, that can be processed, integrated, maintained, shared and published in order to add value to it.

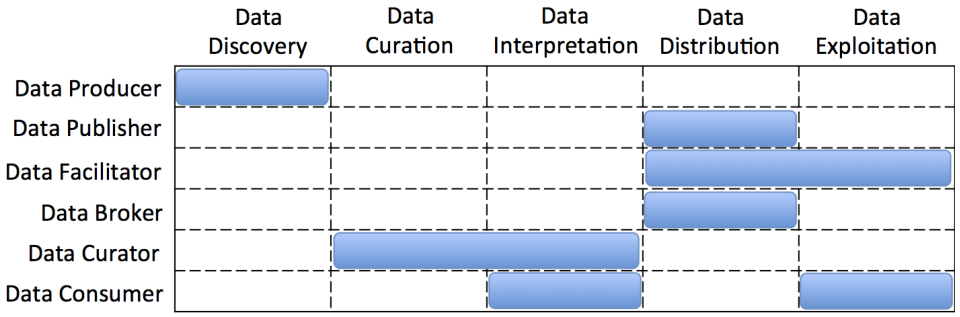

Figure 9.3: The Activities in which an Actor can participate in the Data Value Network through each Role.

- **Data Producer** - A data producer is the entity that creates, obtains, or generates the data. This can be achieved through a number of different value creation techniques, as defined in the *Data Discovery* activity. The role of a data producer can be considered as one of the most important roles within the DVN, as any activity or value creation technique in the network depends on the initial availability of data.

- **Data Publisher** - The data publisher role is also essential in the value network. This role involves the distribution of the data product, as defined in the *Data Distribution* activity. The distribution process enables other stakeholders to discover potentially useful data products.

- **Data Facilitator** - This role involves entities that, in some way or another, as defined through the *Data Distribution* and *Data Exploitation* activities, aid the other stakeholders in using, re-using, or exploiting, data products. This can be done through the provision of software, services, or other technologies. For example, the creator of a government data portal is facilitating the use and re-use of government data from other stakeholders by organising heterogeneous government data in a single location.

- **Data Broker** - A data broker has the role of acting as a 'matchmaker' and linking the roles of the data producers/publishers and the data consumers, enabling the balancing of the supply and demand. Through the *Data Distribution* activity, a data broker hence encourages data re-use throughout the DVN. Whilst similar to the role of a data publisher, a data broker does not only passively share the data, but actively provides the data to a relevant consumer.

- **Data Curator** - The role of a data curator is to modify or enhance the data in a manner that it is more usable for the target aim. This role involves executing any value creation techniques within the *Data Curation* and *Data Interpretation* activities. A data curator can influence the outcome of the DVN by adapting the data so that its highest value potential can be exploited.

- **Data Consumer** - The data consumer role can be considered as the final role in the DVN, however, this is not always the case. For example, when a consumer gives feedback, the feedback can in turn be used as a data product by the product manufacturer. In the case of crowdsourcing, the data consumer also has the role of a curator, blurring the lines between both roles. Actors in the role of a data consumer can exploit the data product in many ways, as defined in the *Data Interpretation* and *Data Exploitation* activities.

9.5 Barriers, Enablers, and Impacts of Value Creation

Within the DVN, value creation is both dependent on a number of dimensions, and also results in impact on other dimensions. Taking an open government initiative as a use case, based on efforts in the primary studies (See Part II), and other literature such as [27, 59, 144, 159], we identify the dimensions with the strongest impact. As open government initiatives are a subset of any open data initiative, the following dimensions apply to any initiative, albeit there can be some differences in the resulting impacts.

Figure 9.4 maps the relationship between the different dimensions, where a number of dimensions act as *enablers* or *barriers* towards value creation. In turn, the value creation process impacts a number of other dimensions. The stakeholders, while they give input for value creation, are also impacted through the results of their efforts.

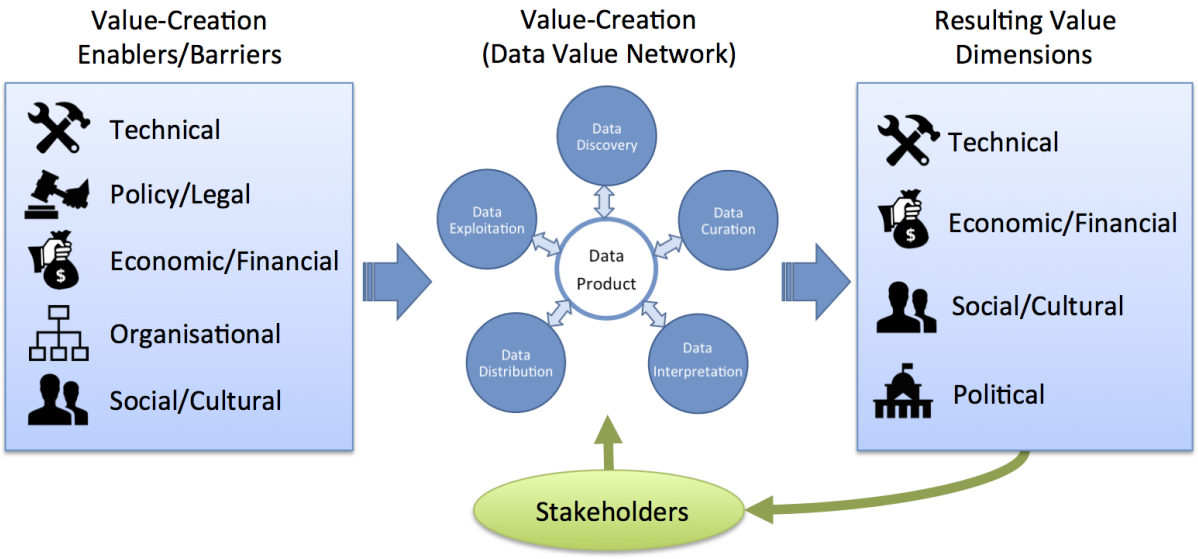

Figure 9.4: Dimensions impacting, and impacted by, value creation.

9.5.1 Value Creation Enablers/Barriers

The latter dimensions have a great impact on value creation in that they control to what extent value is created.

The **Technical Dimension** mostly regards aspects concerning the data itself. Data activities in a DVN all have the purpose of adding value to data. We can consider 'adding value' to be equivalent to 'making the data more usable, or more fit for use'. So, for example, while data in PDF format is easily human-readable, its conversion to RDF would make it more usable where the use case requires data to be machine readable. The opposite can also stand true. The format of the data is an essential aspect. Two of the eight Open Government Data Principles[2], in fact, regard the format in which data is made available to the public. They state that such data should be available in a *machine-processable* format which is *non-proprietary*. Such data would enable easier and un-restricted use of the data for value creation. Furthermore, if a format such as RDF is used, data ambiguity is reduced due to the format's expressivity, making the data more *understandable*. Additionally, the use of common schema aids to reduce interoperability issues caused by the large heterogeneity of the existing data. In order to encourage its use, data must also be easily *discoverable*. This is possible through the use of good quality metadata. The implementation of agreed-upon standards would aid reduce some, if not most, of the issues within this dimension.

The **Policy/Legal Dimension** regards issues with existing laws or policies that, through their ambiguity or due to being out-of-date, prevent data from being used to create value. On the other hand, well thought out policies encourage and enforce the creation of value, for example the publishing of data as Linked Data. Fortunately, there are growing efforts towards amending such laws and policies, but there is still a long way to go. Copyright and licensing of data can inhibit its unrestricted use. The incompatibility of licences, due to the data being created by various entities, further aggravates the issue. Privacy and data protection is another important aspect. Data providers need to strike a balance between making data freely available, whilst respecting the right to privacy [60].

The **Economic/Financial Dimension** is about aspects related to monetary issues and mainly concern the data publisher and the data producer roles. Being a relatively new concept, there might not be any budget allocation specifically for open government data efforts. In order to foster value creation, governmental entities cannot solely rely on existing data created in their day-to-day functionalities. Commitment is required, and hence also finances, for identifying and opening datasets with a high value creation potential. Using an Open Data Maturity Model, the authors of [23] have estimated the (total) market size of open data in the European Union to be between 193 and 209 billion Euro for 2016. A key capability to participate in the global data market is the capacity to innovate, and to ensure continuous improvement in product and process development [68], as well as identifying the correct competitive scope [111]. The DVN also increases competition. Having a global market where a number of actors provide a similar data product, consumers can identify the best product for their use case. Producers can compete with each other by attempting to provide the best data product for their target consumers and build up a reputation.

The **Organisational Dimension** is concerned with the strategic aspects of the involved stakeholders. This dimension is especially relevant for governmental institutions. Considering there probably isn't an institution specifically in charge of open government data initiatives, data can get lost in the various hierarchical levels of a government. Adequate workflows need to be put in place for all the processes within a government data life cycle.

Finally, the **Social/Cultural Dimension** regards the feeling of the public towards open government data. While efforts are well under way to increasing awareness about the potential of open government data, not all stakeholders are ready to jump on the bandwagon. Workers within governmental entities might not understand the value of the data they are gathering/creating. This results in lack of motivation towards providing this data to the public. Stakeholders can also have misconceptions about the opening

[2] http://opengovdata.org/ (Date accessed: 2 August 2016)

of public data. While open data can be considered as unfair competition for private entities (who invested to create their own data), public entities might consider the commercial appropriation of public open data unfair. The public also needs to be further informed on the advantages of public participation in creating value.

9.5.2 Impacts of Value Creation

Value creation has a number of different dimensions of impact, which in turn affect the stakeholders in creating *public value*. This term is used to define "what adds value to the public sphere" [15], where the public sphere is used to broadly indicate all of the following dimensions:

Technical Value is simply generated through the implementation of standards and the creation of services. As more value is created upon government data, the available data will be of better quality, and value creating services will increase.

Economic Value is defined as the worth of a good or service as determined by the market [62]. Value creation upon data enables the data itself to be considered as a product. Therefore, opening government data encourages its re-use in value creation, in turn stimulating competitiveness in the participating stakeholders and also encourages economic growth. For example, Mastodon C (a Big Data company) used open data to identify unnecessary spending in prescription medicine[3]. This will result in potentially huge savings from the National Health Service in the UK. Due to the potential of data to be used over and over (until it remains relevant), the economic impact of adding value to it and using it as a product is different when compared to the more traditional product manufacturing. First and foremost this is evident in the re-use of data in another context, or domain, that it was originally envisaged for. For example, e-commerce businesses use historic purchase data to identify patterns and suggest items to users. Moreover, the data can be processed repetitively in order to make it more usable for a specific use case, for example, by changing its format, removing irrelevant data, or linking it with other data. Data can also be interpreted and made human-readable by extracting knowledge from it. For example, in the case of government data, this data processing would enable all citizens to exploit the data, and potentially even give their feedback. In turn, this feedback could be added value that the governmental entity can exploit.

Social/Cultural Value is generated through creating innovative services that can aid relevant entities in consuming such data. For example, the use of public school locations can help a new family identify the best school for their child in their neighbourhood. Another example is the use of public transport timetables in mobile applications, which help commuters plan their trips on the go. The development of such services based on adding value to data also results in in the creation of jobs. DVNs can also help to preserve and better showcase the cultural identity and diversity of a region. A perfect example in this regard are digital museum, archive and library aggregators, who collect metadata about millions of cultural heritage artefacts. Here, museums, libraries and archives are data providers to a hierarchic network of geographic or thematic data aggregator nodes. The German Digital library[4], for example, aggregates semantic metadata from more than 2,000 memory institutions and feeds this data into the European cultural heritage portal Europeana[5], as similarly do many other national aggregators. As a result, this vast DVN allows citizens to explore the cultural heritage in completely novel ways, facilitating exploration across institutional, administrative and thematic boundaries.

Political Value is created through the stimulation of democratic dialogue. Through participatory

[3] http://theodi.org/news/prescription-savings-worth-millions-identified-odi-incubated-company (Date accessed: 2 August 2016)

[4] https://www.deutsche-digitale-bibliothek.de/?lang=en (Date accessed: 2 August 2016)

[5] http://www.europeana.eu/ (Date accessed: 2 August 2016)

governance, citizens can gain a better insight as to how the governing process works. Stakeholders can possibly also participate in improving the policy-making process, hence increasing *citizen social control* and also being more informed. As a result citizens are able to make better decisions [118]. Besides, the efforts of governmental entities to be more transparent and accountable increases citizens' trust in their government.

9.6 Linked Data

As discussed in Section 6.1, many publishers are following Linked Data practices. This trend is also reflected in recent open government data initiatives, where Linked Data practices are being followed by an increasing number of data publishers/producers such as `http://data.gov.uk` and `http://data.gov`. Yet, the use of Linked Data in open government initiatives is still quite low [130]. This might be due to a number of reasons, as the use of Linked Data is a process involving a high number of steps, design decisions and technologies [150].

While significant efforts in literature cover advantages of using Linked Data (for example [35, 51, 130, 132]), there is no evident effort targeted towards the benefits of using Linked Data specifically in open government data value creation. We here therefore proceed to focus on the value creation techniques described in Section 9.3 and the advantages and benefits provided through the use of Linked Data. While still having similar barriers, enablers, and impacts, as described in Section 9.5, the use of Linked Data can result in different levels of impact, since the use of Linked Data techniques directly reduces some barriers of the technical level.

9.6.1 Linked Data as a Basis for Value Creation

Linked Data and Semantic Web technologies have the potential of solving many challenges in open (government) data, as well as possibly lowering the cost and complexity of developing data-based applications such as government portals.

Starting from the most common starting point of creating value, in general, data **generation** is the least impacted from the use of Linked Data since essentially the data is still being created. Data **procurement** is similarly not impacted to a high level. Yet, the data **gathering** process can be enhanced through the use of Linked Data. Consider the example of providing feedback based on a linked open dataset consisting of budget data. The use of Linked Data enables feedback providers to have further context on the available data through the links. This would aid them in making a more informed decision. Furthermore, the high level of granularity of Linked Data has the potential of providing a deeper insight on the resource at hand. Also, since the data publisher is not necessarily the data producer, Linked Data will enable the access to primary data through the use of provenance information located within the metadata. In the case of data **selection**, the use of Linked Data is particularly useful in querying for subsets of an existing dataset. Query languages such as SPARQL enable actors to generate complex queries and get very specific subsets of data.

The value creation techniques within the Data Curation activity are some of the highest impacted techniques within the DVN through the use of Linked Data. Linked Data is based on models (schema) or ontologies that are best suited to represent the data at hand. In this way, the **organisation** of data is very easily achieved through the manipulation of the model at hand. If an entity is working with Linked Data, we can safely assume the data is represented in a semantically rich, machine-processable format. Hence, links with or between other datasets are more easily identified through the implemented models, and thus, the data **linking** process is simplified. Thereafter, data **integration** and **merging** follow easily through joining the existing models. Through the use of the standards required to obtain Linked Data, the

fitness for use of data, and hence its **quality**, is immediately increased. For example, data ambiguity is decreased through the use of a semantically rich format, and data consistency can be ensured through the implemented data model. Moreover, in some instances, the quality assessment of data (and the ensuing data repairing or **cleaning**) can be more easily executed. For example, having a model for a linked dataset enables a stakeholder to assess the schema completeness for the dataset. Linked Data also enables (semi) automated cleaning and repairing of datasets through the use of reasoners. In this way, the violation of logical constraints is easily identified through the dataset's underlying model. Through the use of metadata, a consumer can also check the provenance of the data, and ensure that it is a reliable source. Timeliness and versioning information can be obtained in the same manner.

Having Linked Data means that the available data already conforms to some standards with regard to formatting, however this does not necessary make it easier to **serialise** to other formats. Yet, the use of agreed-upon standards positively affects the accessibility, discoverability, and re-usability potential of the data in question. Since Linked Data standards demand the use of a semantic representation such as RDF, Linked Data is automatically more accessible than other standards such as CSV or PDF. Data **analysis** is also enhanced through the use of Linked Data. As explained above, Linked Data enables easier integration and merging of datasets, which in turn affect the implementation of analysis techniques. Moreover, through the existence of links it is easier to get further context and information on the data at hand, enhancing pattern identification. Similarly, the use of Linked Data in **information/knowledge extraction** also provides further insight and context to actors through links between the datasets, and within datasets themselves. This increased information directly affects the data interpretation process, as the data consumer can interpret the data in a more informed manner, and generate knowledge from the existing information.

The aim of the value creation techniques within the Data Distribution activity is to make the data more accessible as a data product. As mentioned above, the use of Linked Data standards automatically makes the data more accessible and discoverable. Hence, **stored** or **published** Linked Data has the potential to be easily accessed and manipulated through a variety of manners, such as RESTful APIs and public endpoints (queryable through SPARQL). This means that while Linked Data alternatives might require a consumer to download a data dump, the use of Linked Data enables the same consumer to access the specific subset of data he/she needs, and manipulate it easily. Additionally, each data resource is dereferenceable, i.e. the resource URI can be resolved into a web document on the Web of Data. The **sharing** of data is also impacted through the use of Linked Data technologies, as the links in between different datasets make them more easily discovered through the crawling of web resources, which potentially could lead to the addition of the dataset to the more known Linked Open Data Cloud, in turn making its **advertising** easier.

Data Exploitation is possibly the activity that has the highest impact from the use of Linked Data. Similarly to the knowledge/information extraction process, **question answering** and **decision-making** are enhanced through the existence of links and the provision of further context. Hence a more informed stakeholder is more capable of making the best decision, or obtaining the best answer for the problem at hand. The creation of **visualisations** is also affected through the existence of links between multiple datasets. Visualising a dataset against a related dataset has the potential of providing the consumer with a new and different understanding of the data. Finally, **service creation** on top of Linked Data has the advantage of easier data consumption (through the use of standards), and more interoperability.

The above benefits of using data for value creation are only a few, yet they collectively motivate and enhance the exploitation of open (government) data. Of course, this does not mean the implementation of a Linked Data approach does not have its challenges. Various efforts in literature, such as [132], provide discussions on the topic.

9.6.2 An Example of Linked Open Government Data

http://publicspending.net is a data portal created with the scope of demonstrating the power of economic Linked Open Data in analysing the situation with regard to market, competition conditions, and public policy, on a global scale. The creators of this portal consume and create value upon public spending data of seven governments around the world. Results of the analysis led on the data are then published on the portal as tables, graphs, and statistics. The stakeholders here participate through five of the value-creating roles described in Section 9.4 and execute value creation processes accordingly.

Firstly, the public spending data is produced by the various governments (Data Producers). The data is then subject to pre-processing and data-preparation. Through the role of a Data Enhancer, the stakeholders here homogenise and link the data through the Public Spending Ontology and other widely used vocabularies such as Dublin Core[6] and FOAF[7]. The resulting data in RDF is then published (Data Publisher) on the portal and is available both as bulk datasets and through a SPARQL endpoint. The Data Facilitator role is then fulfilled through the application built on top of the data. These stakeholders use the internal data, along with other cross-referenced and external data, to provide a portal acting as an information point. Finally, the Data Consumer can view and exploit the provided data in a myriad of ways, including exploring and scrutinising spending data that giving them a good insight as to what is being spent, where, and by whom. Such an open government data initiative enhances accountability and prevents corruption since it aids citizens to be more informed about how their country is being led, and if it is being led in a suitable manner. This can also help them decide who to vote for in an upcoming election.

9.7 Use Case Scenarios

In order to provide a better understanding of the DVN and its contribution to generating an economic data ecosystem, we here provide three scenarios. In the latter, data is generated and consumed as a product, and the dimensions defined in Section 9.5 are affected as a result.

9.7.1 Exploiting Weather Data

Weather data is one of the most popular domains of data that is very frequently used not only for the obvious weather forecasting, but also in many different domains such as insurance firms, aviation, military operations, agricultural decision-making, constructions, and forensic science. Weather data is particularly special in that it is mostly important in the relatively short timespan after it is collected, however weather data is also important in the long term, such as for research on climate change.

The DVN for weather data starts with *Data Discovery*. Larger weather companies around the world, such as AccuWeather[8] have a large number of sensors and instruments spread around various locations. The company in question (in this case AccuWeather) has the role of either *procuring* or *gathering* the required data, such as wind force, rainfall, temperature, humidity, etc. It is important to note that such data is handled in real-time, so in this scenario the Data Discovery activity is ongoing.

The *Data Curation* activity is the next in line to be executed within the DVN. In this case the collected data is *cleaned* and incorrect or incomplete records are removed. This will ensure that the remaining data is of the best quality possible for the planned usage. Data from different instruments or sensors can

[6] http://dublincore.org/ (Date accessed: 2 August 2016)
[7] http://xmlns.com/foaf/spec/ (Date accessed: 2 August 2016)
[8] http://www.accuweather.com/ (Date accessed: 2 August 2016)

also be *organised* by location and integrated or merged in order to have a complete representation of the weather data in question.

In the next phase, different actors pertaining to different domains participate in the DVN differently. For example, AccuWeather proceeds in the DVN by *analysing* the data in the *Data Interpretation* activity and creating weather forecasts. These forecasts are then *shared* to an audience during the *Data Distribution* activity. On the other hand, firms such as Weather Analytics[9], *curate* and *interpret* the data further with the intention to use it for *decision-making* in the *Data Exploitation* activity. For example, when an insurance company receives insurance claims due to weather related damage, interpreted data from Weather Analytics can be used to identify whether the claims are valid or otherwise. Forensic science is another field that benefits from the use of weather data. AccuWeather is one example of companies who *interpret* and *exploit* data by *analysing* and *extracting knowledge*, and then using the results to help in a *decision-making* process. For example, such a firm is able to provide testimony in court for cases ranging from collapsing roofs, late arrivals of ships, and even murder[10]. Another use of weather data is by the governments who use severe weather warnings to decide whether or nor to proceed with evacuation plans[11].

As is evident through this use case scenario, weather data has a very high impact especially on the economic and societal dimensions of the DVN. Apart from simply influencing our everyday lives through changing plans to reflect current weather, weather data can provide an even higher impact. Through the timely provision of the relevant weather data, citizens can rightly get compensation from an insurance company. On the other side of the coin, an unfounded claim is not compensated. Weather data can be used to determine the outcome of a court proceedings, and a timely weather warning can even save lives.

9.7.2 Real-Time Event Detection

Social media form a very important aspect in many people's lives. They are also an extremely important source of real-life information about a huge number of topics, ranging from mundane activities such as what someone ate for lunch, to more profound events such as the occurrence of an earthquake, or the death of a famous person. Clearly, the detection of events of interest and their interpretation would be considered relevant to develop as a data product. In fact, the authors of [103] follow a similar approach to the one we describe here.

The DVN here starts with data *gathering*. Events are identified on social media through the use of tags, where tags recurring frequently in a short amount of time (trending) would be indicative that an event of particular importance has occurred, and that it is relevant to a large number of people. It is important to note that such data should be handled in real-time, otherwise it might not be relevant any longer. Thus, in this scenario, the *Data Discovery* activity is ongoing.

The *Data Curation* activity is the next in line to be executed within the DVN. In this case the gathered data, due to being crowdsourced by social media users, can contain duplicated, incomplete, imprecise, and incorrect information. Here, a very important action is to assess the quality aspects relevant to the use case in question. Data streaming from the *Data Discovery* activity that cannot be interpreted is either corrected or discarded.

Unfortunately, the *Data Discovery* and *Curation* activities are not sufficient to give the data producer insight on the relevance of the gathered data. Hence, the *Data Interpretation* activity needs to implement

[9] http://www.weatheranalytics.com/wa/ (Date accessed: 2 August 2016)

[10] http://enterprisesolutions.accuweather.com/legal-forensics/cases (Date accessed: 2 August 2016)

[11] http://enterprisesolutions.accuweather.com/assets/documents/AccuWeather_Success_April_2011_Tornadoes_Lives_Saved.pdf (Date accessed: 2 August 2016)

suitable actions to discover the relevancy of the detected events. Let us consider the event of a fire. After analysing the data gathered from social media, the *analysis* action identifies a trending tag: 'fire'. The analysis also points out that the persons posting this event all posted from a particular location. Through leading out a *knowledge extraction* action, it is evident that there is currently a fire event occurring in a specific location.

Considering the real-time approach in this scenario, the identification of an event that could have serious implications is vital. Continuing with the fire example, the information about the event should be immediately communicated to the fire station and authorities of the location where the event occurred. Through the *Data Distribution* activity, the data product can be quickly *shared*, and tragedies can be avoided if, in the *Data Exploitation* activity in the DVN, the data product is used wisely in the *decision-making* action.

9.7.3 Participatory Budgeting

Governments are one of the largest data-producing and collecting entities of multiple domains [2, 58]. The main challenge in releasing value is that such data does not have any intrinsic value, yet it becomes valuable when it is used [59]. Open government initiatives, such as the Public Sector Information (PSI) Directive[12] in Europe, U.S. President's Obama open data initiative[13], the Open Government Partnership[14], and the G8 Open Data Charter[15], are a very popular approach how to exploit government data and create value.

Participatory Budgeting is one approach within an open government initiative that aims to encourage the participation of citizens in budget decision-making. In other words, citizens get to decide how their taxes get spent. Participatory budgeting is a powerful tool that can increase civic participation and community engagement. Basically, the citizens demand transparency, and the government or governmental entity supplies the information required for decision-making as a data product. New York is an example of a city implementing this approach[16].

As opposed to the previous scenarios, this DVN starts with the *distribution* of relevant data; that is, the New York City Council provides the information that the citizens should provide their feedback on. After *sharing* this information, the council proceeds to *gather* the citizens' feedback. Depending on the feedback process, which can include the submission of forms, or simply voting online, the data will require to be *organised* through the *Data Curation* activity, and *cleaned* if necessary. The *Data Interpretation* activity and the *information extraction* and *analysis* value creating techniques would give an insight on the citizens' feedback. In the case of this scenario, the citizens feedback is only relevant for the decision-making of a specific budget allocation, therefore, the council *exploits* and consumes the data product to lead out the *decision-making* value creation technique.

This scenario is real-life reflection of the DVN in participatory budgeting. While, in this case, there is no competitive intentions directly correlated to the data product, there is still an economic impact through the decisions undertaken based on the data.

[12] http://ec.europa.eu/digital-agenda/en/european-legislation-reuse-public-sector-information (Date accessed: 2 August 2016)

[13] http://www.whitehouse.gov/open/documents/open-government-directive (Date accessed: 2 August 2016)

[14] http://www.opengovpartnership.org/ (Date accessed: 2 August 2016)

[15] https://www.gov.uk/government/publications/open-data-charter (Date accessed: 2 August 2016)

[16] http://council.nyc.gov/html/pb/home.shtml (Date accessed: 2 August 2016)

CHAPTER 10

Assessing the Value Potential of Data Products

In order to assess the success of open data initiatives, there exist a large number of assessment frameworks that aim to evaluate the effectiveness of an initiative in achieving its goals and objectives. Yet, rather than assessing the resulting impacts of such an initiative, real-life assessments, as documented in literature (see Section 3.1), mostly involve checking whether open data initiatives are obeying existing policies and regulations [124]. Since the latter are not necessarily up to date with current technologies and approaches, this assessment is not really representative of the success of an initiative.

Consider the example of an open government initiative where data is published in PDF. While the entity would be obeying existing laws requiring opening up such data, the use of PDF makes it pretty inconvenient for re-use and re-distribution. In this case, one could argue that the open government initiative is not really a success. For this reason, a number of assessment frameworks analyse open government data initiatives based on different criteria [18, 79]. The latter include nature of the data, citizen participation, and data openness.

While there is still the problem that there is no agreed-upon assessment framework to evaluate open data initiatives, there is also limited literature (such as [140]) that focuses on the *impact of value creation*. Considering many resulting benefits of open data depend on the creation of value (through the execution of one or more value creation techniques), we deem it essential to assess open data initiatives on their *potential for enabling value creation*.

10.1 Value Creation Assessment Framework

With the aim of defining the ideal assessment framework to analyse open data initiatives, in this section we revisit the literature covered in Section 3.1 and identify the aspects currently being assessed to analyse open government initiatives. Being an instance of the generic open data initiatives, the aspects used to assess open government initiatives also apply to any open data initiative in general. In fact, since governments are one of the largest producers of open data, open government initiatives usually cover data in a large variety of domains. In Figure 10.1 we provide an overview of commonly evaluated aspects (in blue). These mostly concern implementation aspects, such as the format of the data, and how the initiative respects the requirements set from existing laws and policies.

Through the publishing guidelines specified in Section 4.1.2 and the data quality aspects specified in Sections 4.3 and 9.2, we came up with important aspects in an open data initiative that are currently not

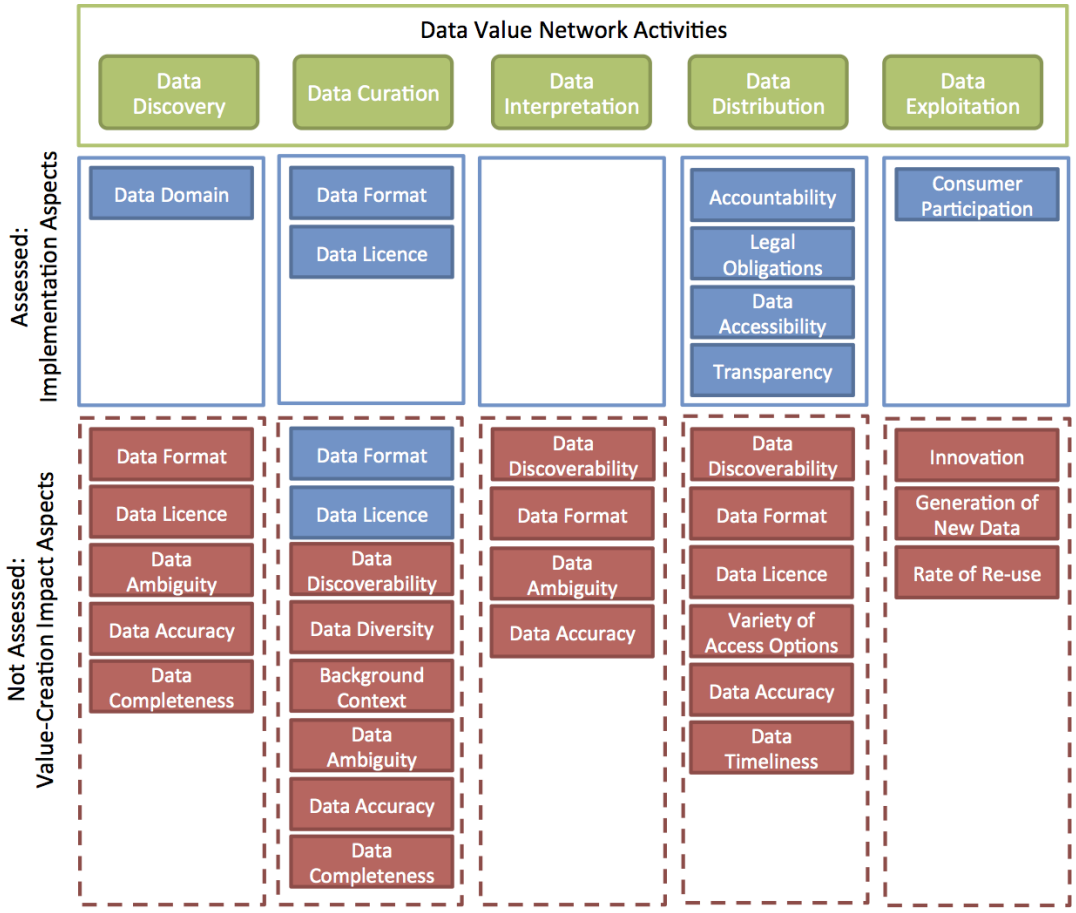

Figure 10.1: Aspects assessed in existing frameworks (blue), aspects proposed for Value Creation Assessment Framework (Red).

being assessed in existing assessment frameworks. These aspects impact the overall re-usability of the data, and therefore also any value creation upon it. The bottom part of Figure 10.1 portrays the missing aspects (in red), i.e. those that are currently not being considered when evaluating the success of an open data initiative.

We propose the aspects shown in the bottom part of Figure 10.1 as part of a *Value Creation Assessment Framework*. The aim of this framework is to provide a guideline as to what aspects of an open data initiative should be assessed to determine the potential of an open data initiative to enable value creation, and thus exploit open data to its highest potential. Since one of the major aims of open data initiatives, particularly open government initiatives, is the release of social and commercial value, we deem that the proposed aspects are vital to determine the success of an initiative. Here we briefly describe the aim of each aspect.

- **Data Format** - Formats such as CSV and RDF are much more usable then PDF. This is because they allow easier re-use of the represented data.
- **Data Licence** - Other than allowing for reasonable privacy, security, and privilege restrictions, data has the highest value creation potential if it is not subject to any limitations on its use due to copyright, patent, trademark or other regulations. Hence, data with an open licence has the best value creation potential.

- **Data Ambiguity** - Data ambiguity is reduced when a representationally rich format (e.g. RDF) is used.
- **Data Accuracy** - The extent to which data accurately represents the respective information.
- **Data Completeness** - Data is complete when all required information is available, for the representation of the data in question.
- **Data Discoverability** - This aspect depends on the metadata annotating the data in question, and enables stakeholders to more easily find data that is relevant to their needs. Data Discoverability is also affected by the search functions provided by a government portal or catalogue.
- **Data Diversity** - In the Linking value creation technique within the DVN, the use of diverse datasets has the potential of releasing new insights or unforeseen results.
- **Background Context** - The linking of datasets provides further context to the data in question, enabling stakeholders to have a deeper understanding.
- **Variety of Access Options** - Providing various access options to the available data, such as APIs and SPARQL endpoints, encourages stakeholders to create value upon the data as they are able to access the data in their preferred manner.
- **Data Timeliness** - Certain data might only be valuable if it is made openly available shortly after its creation.
- **Innovation** - Creating new products (data or otherwise) based on open government data is a direct impact of value creation. Innovations include services and applications.
- **Generation of New Data** - The value creation techniques in the Data Exploitation process can result in the generation of new data, such as visualisations, that provide new interpretations or insight on the existing government data.
- **Rate of Re-use** - The participation of stakeholders in consuming the data is essential for value creation. There is no use in having data made openly available if it is not exploited. The rate of re-use of open government data is directly indicative of the value creation potential in the assessed initiative.

10.1.1 Value Creation Assessment Framework in Action

In this section we implement the proposed assessment framework on two open government data initiatives, namely `http://www.govdata.de` and `http://www.gov.mt`, in order to portray its relevance and applicability in the context of value creation on open data. Keeping in mind that this implementation is acting as a proof of concept, we restrain our metrics to assess the portal on a high level, as we consider a thorough and more accurate implementation to require significant more research. We therefore base the provided metrics on ground research. In Table 10.1 we provide a description of the metrics used, and the results of the portals[1]. We assign marks according to the assessed aspect, and where relevant we average the marks based on the number of available datasets. For example, to assess the data format of eight datasets, if four datasets are in RDF and linked to other datasets (4 x 5 marks) and four datasets are in CSV (4 x 2 marks), then the result for the data format aspect is 3.5 marks.

[1] As per 29th of December 2015.

Value Creation Impact Aspects	Assessment Metrics	Results govdata.de	Results gov.mt
Data Format	5 star scheme for LOD: 1-5 marks according to format	2.71 out of 5	2.39 out of 5
Data Licence	0 marks if no licence specified, 1 mark if licence has some restrictions, 2 marks if open and enabling re-use	1.85 out of 2	0 out of 2
Data Ambiguity	1 mark if using semantically rich formats (e.g. RDF)	0 out of 1	0 out of 1
Data Accuracy	Requires use of a gold standard[a]	-	-
Data Completeness	Requires use of a gold standard[a]	-	-
Data Discoverability	1 mark if metadata is available, 1 mark if portal offers search functions on the data (2 marks max)	2 out of 2	0 out of 2
Data Diversity	1 mark if there is more than one dataset on a specific domain	1 out of 1	1 out of 1
Background Context	1 mark if datasets are linked to other external datsets	0 out of 1	0 out of 1
Variety of Access Options	1 mark if more than one access option is available	1 out of 1	0 out of 1
Data Timliness	1 mark if data has a timestamp, 1 mark if recently updated data is available (2 marks max)	2 out of 2	0 out of 2
Innovation	1 mark if portal provides innovations based on published data, 2 marks if different innovations are provided (e.g. services, applications) (3 marks max)	3 out of 3	0 out of 3
Generation of New Data	1 mark if portal enables users to generate new data (e.g. visualisations)	0 out of 1	0 out of 1
Rate of Re-Use	1 mark if portal provides links and information on re-use of the published data	0 out of 1	0 out of 1
	Total	13.56 out of 20	3.39 out of 20

Table 10.1: Value Creation Assessment Framework metrics and results for two open government data initiatives.

[a]This aspect cannot be assessed on a high level as it requires the use of an algorithm that analyses each dataset in a portal and compares it to a gold standard.

Having a value creation potential of 13.56 marks out of 20, http://www.govdata.de can do with some improvements, especially with regard to the use of RDF and the linking to other documents. The portal could also benefit from enabling users to both create new innovations or data through the portal itself, and also from providing some sort of documentation to portray any innovations based on the data in question. In summary, http://www.govdata.de is on the right track towards the opening of governmental data, however it definitely requires more effort towards encouraging stakeholders to create value upon the published data.

On the other hand, http://www.gov.mt does not really excel in publishing government data. Apart from providing very few datasets, some require logging in with a government-issued e-id to download, and others are not even available (404 error given). Moreover, no search functions are provided to aid a user search within the provided datasets, such as a faceted browser. Whilst there is a statement encouraging stakeholders to innovate upon the data, no actual data licence is provided, leading room towards uncertainty.

CHAPTER 11

Mapping the Demand and Supply of Data Products

In this chapter we propose a Demand and Supply Distribution Model (Figure 11.1) that portrays how data producers and consumers (in the generic sense of the word, i.e. entities who use data in any way) can balance out their efforts whilst participating in the global data market. These two stakeholders can participate through all activities in a DVN, however usually a data producer participates through the Data Discovery and Data Distribution activities, whilst a consumer usually participates through all the other activities. From the perspective of a data producer, the Demand and Supply Distribution Model provides an entry point for stakeholders to create value and participate in a DVN by enabling and enhancing the Data Discovery and the Data Distribution activities. On the other hand, from the perspective of a consumer, the model mitigates a number of barriers that hinder stakeholders from finding data, such as data fragmentation, language barriers, lack of information on the datasets, lack of search functions, and issues within coordination processes [159]. We also provide a concrete implementation of the model through an online service and lead out a preliminary evaluation to assess its usefulness.

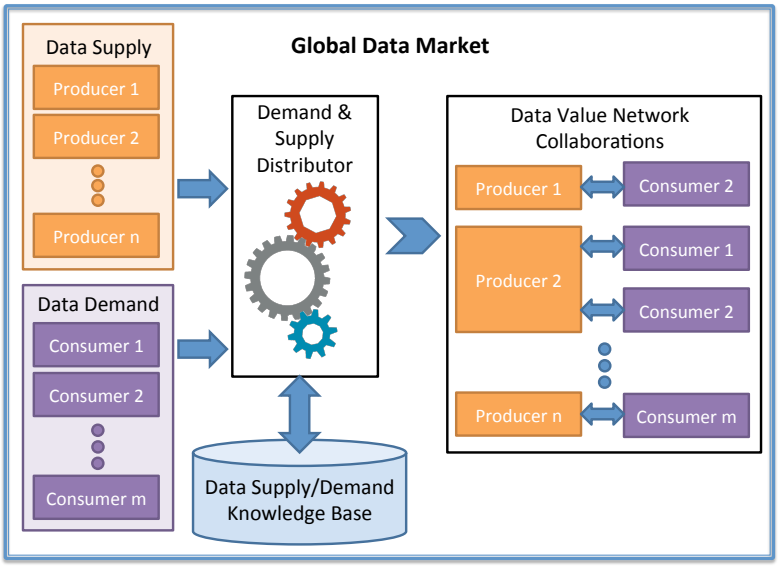

Figure 11.1: Demand and Supply Distribution Model.

11.1 Demand and Supply Distribution Model

The Demand and Supply Distribution Model is based upon the knowledge base shown in Table 11.1, and acts as a *dynamic leveler* between data supply and demand. In this knowledge base we indicate various datasets and related relevant information, such as the domain of the data product, and the way it was consumed. The purpose of this knowledge base is to portray the development and exploitation of a number of data products with an economic motivation. As can be seen, there is a large variety in the domains of the data products. Any type of data can be used within a DVN, besides also being re-used in use cases other then the one originally envisaged, as is particularly evident in the fifth entry in the table.

	Publisher	Domain	Access Method/ Data Format	Consumer	Aim	Short Description
1	MusicBrainz[1]	Music Data	API	BBC[2]	News Enhancement	Music data from MusicBrainz is linked to news data on the BBC music site[3] in order to provide discographies and track listings across about 700 artist pages.
2	Europeana[4]	Cultural Data	API, SPARQL Endpoint, Datadump	Historiana.eu	Educational Portal	Data from the Europeana database is aggregated in the Historiana.eu portal in order to act as an educational website.
3	Her Majesty's Treasury[5]	Government Expenditure Data	Sparql Endpoint, CSV	wheredoesmymoneygo.org	Informative Portal	Where Does My Money Go? aims to promote transparency and citizen engagement through the analysis and visualisation of information about UK public spending, extracted from the COINS[6] dataset.
4	Safecast[7]	Environmental Data	API, CSV	Fukushima Government	Radiation Awareness	The Fukushima Government used radiation measurements data from Safecast in order to populate maps, showing the radiation levels in different locations[8].
5	Vehicle	Vehicular Data	n/a	Progressive[9]	Insurance Services	Progressive aggregates sensor data from a vehicle in order to identify a person's driving style, and then adapts insurance policies according to how safe the person drives.
6	Office for National Statistics[10]	Crime	CSV, Spreadsheets	Walkonomics[11]	Environmental Safety	Walkonomics use crime statistics provided by the Office for National Statistics in order to develop an app. Using this app, a person can check the, "walkability" of a given street based upon various categories, such as fear of crime, road safety, and pavement quality.

Table 11.1: Demand and Supply Knowledge Base excerpt.

Entities participating as data producers or publishers in the DVN can be overwhelmed by the amount of competition in the global market. Likewise, data consumers can find it difficult to identify whether the data product they need is already on the market. Moreover, if the data is created with a specific use case in mind, it might be difficult to envision or implement its use in a different domain. This model we propose can be a solution to these problems, where information about data products resulting from entities' DVNs are indexed in a knowledge base, making them available for easier search and discovery. Using this knowledge base, data consumers can easily identify publishers or producers that are providing the data product that they require. Similarly, data producers can be aware of the data products already on

[1] https://musicbrainz.org/ (Date accessed: 2 August 2016)
[2] http://www.bbc.com/ (Date accessed: 2 August 2016)
[3] http://www.bbc.co.uk/music (Date accessed: 2 August 2016)
[4] http://www.europeana.eu/ (Date accessed: 2 August 2016)
[5] https://www.gov.uk/government/organisations/hm-treasury (Date accessed: 2 August 2016)
[6] http://data.gov.uk/dataset/coins (Date accessed: 2 August 2016)
[7] http://blog.safecast.org/ (Date accessed: 2 August 2016)
[8] http://fukushima-radioactivity.jp/pc/ (Date accessed: 2 August 2016)
[9] https://www.progressive.com/auto/snapshot/ (Date accessed: 2 August 2016)
[10] http://www.ons.gov.uk/ons/index.html (Date accessed: 2 August 2016)
[11] http://www.walkonomics.com/ (Date accessed: 2 August 2016)

the market, thus having the opportunity to target a niche, if it exists, rather than attempting to compete with established data producers. Basically, this model maps the real-life data demand and supply picture and undertakes the role of a Data Broker, as defined in Section 9.4. By following this model stakeholders can hence optimise their process of participating in the value creating process upon data products by having a clear picture of the supply and demand, and acting accordingly.

11.2 Demand and Supply as a Service

In order to act as proof of concept to the model, we created a cloud service in the form of a portal. We provide the Demand and Supply as a Service (DSAAS - available online at `http://butterbur22.iai.uni-bonn.de/dsaas/`); an entry point to the DVN and hence also to the *Economic Data Ecosystem*. The portal caters for two discrete roles, reflecting the Demand and Supply Distribution Model, namely data publishers (Supply) and data consumers (Demand), and aids the value creation process through enabling and enhancing data discovery and re-use, collaborations, and providing contributions to the data market.

The DSAAS provides two different ways for consuming data; a faceted browser and a RESTful API. The faceted browser enables data consumers (humans) to browse the Data Supply and Demand Knowledge Base (as shown in Figure 11.1) of existing data that they can consume or even contribute to. The RESTful API, on the other hand, provides automated access to the Knowledge Base. This enables third parties to provide their own applications based on the available data. Following, we explore the various functionalities of the DSAAS:

- **Browsing existing datasets** - Data Consumers, or other entities who want to discover the current state of the data market, can browse existing datasets through a faceted browser, as shown in Figure 11.2. Results can be filtered according to various aspects of the datasets, including the licence, publisher, data format, data content, geographical coverage, etc. This faceted browser hence enables stakeholders to quickly search for datasets that are relevant to their current need. Once a particularly interesting dataset is identified, a user can also view all the details about the dataset, as well as any use cases of other stakeholders who already used the dataset. These use cases are provided with the intention of giving potential consumers further insight on the (re-)usability of the dataset in question, especially in contexts not in the domain of the original dataset use.

- **Adding new data** - Apart from browsing existing data, stakeholders can also use the DSAAS to add new datasets. This is very simply done by filling the relevant information in a template, as shown in Figure 11.3. The latter includes information such as the producer of the dataset, the formats available for consumption, the licence under which the dataset can be used, the language of the dataset, the temporal and geographic coverage, etc. Moreover, if the stakeholder knows another stakeholder who used their dataset in a use case, this information can be added as well. The more information that a stakeholder can provide about the dataset in question, the more easily the dataset can be discovered (and re-used) by data consumers.

- **Browsing the data requests** - Similar to the functionality of browsing existing datasets, stakeholders can browse through data requests posted on the DSAAS. This functionality is shown in Figure 11.4. These requests indicate a need for specific data. This will allow data producers to identify a niche in the data market that was not previously catered for by any other data producer, and provide them with the opportunity to target it in order to be on the forefront of the competition. Data consumers, on the other hand, can second any data request in order to raise awareness about its relevance to various stakeholders.

- **Adding a new data request** - Adding a data request, or a demand, is simply done through filling a template, as shown in Figure 11.5, where stakeholders can enter information about the data that they require to varying degrees of details.

Through the above functionalities, we hence provide a 'matchmaking' service that acts as a broker between data producers and consumers. Acting as a sort of manual recommendation system, the DSAAS allows consumers to easily identify datasets of interest by filtering the results through the faceted browser. Similarly, data producers can identify the current needs of consumers by looking in the posted data requests. While this matchmaking service is currently manual, it can be at least partially automated by implementing comparison metrics that match the domain, keywords, descriptions, etc. between a data demand and existing datasets. Preferences can also be saved, and an intelligent system can eventually 'learn' to provide more accurate results.

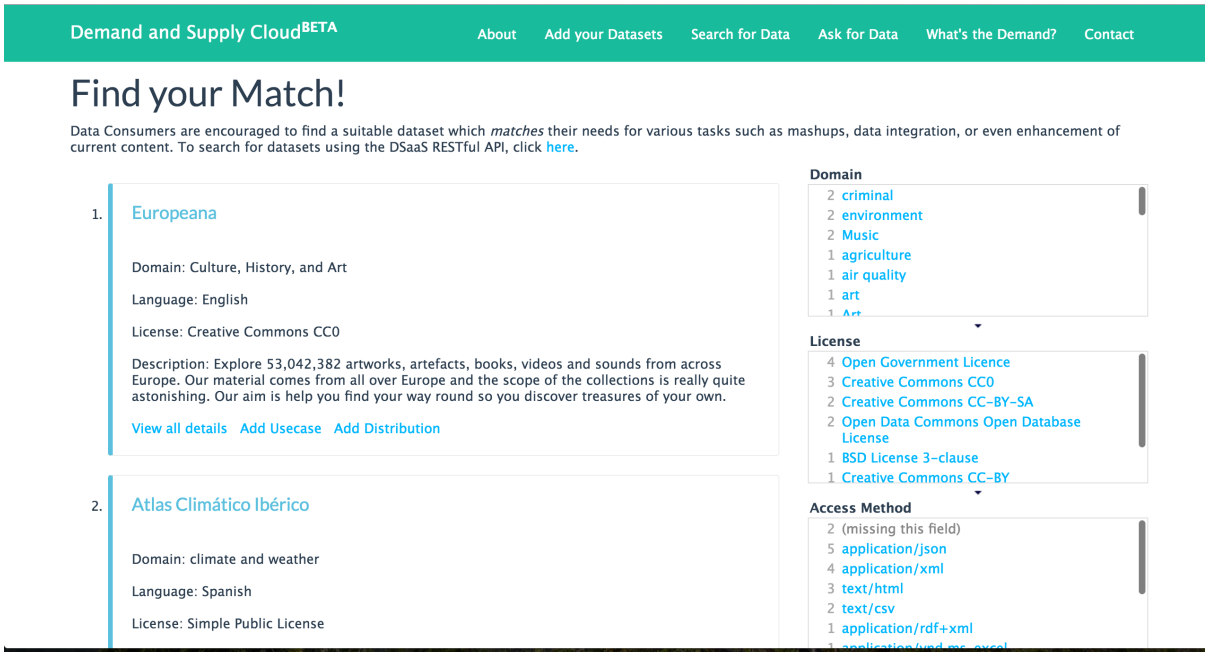

Figure 11.2: DSAAS: Browsing existing datasets.

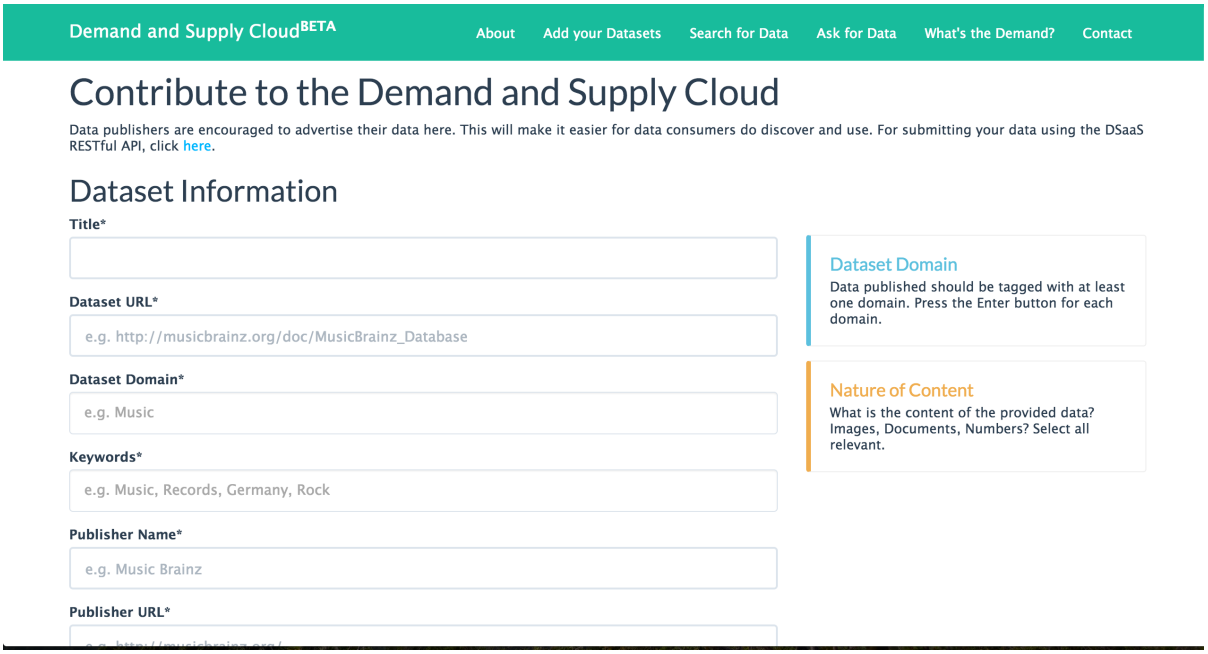

Figure 11.3: DSAAS: Adding new datasets.

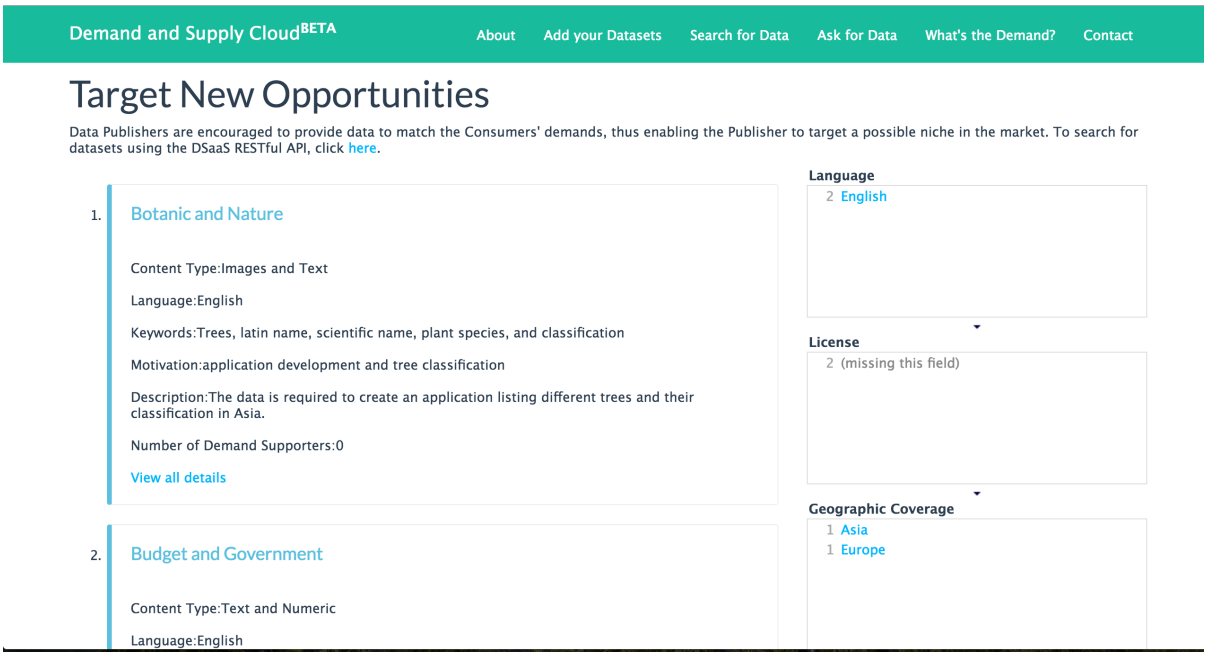

Figure 11.4: DSAAS: Browsing requests for new datasets.

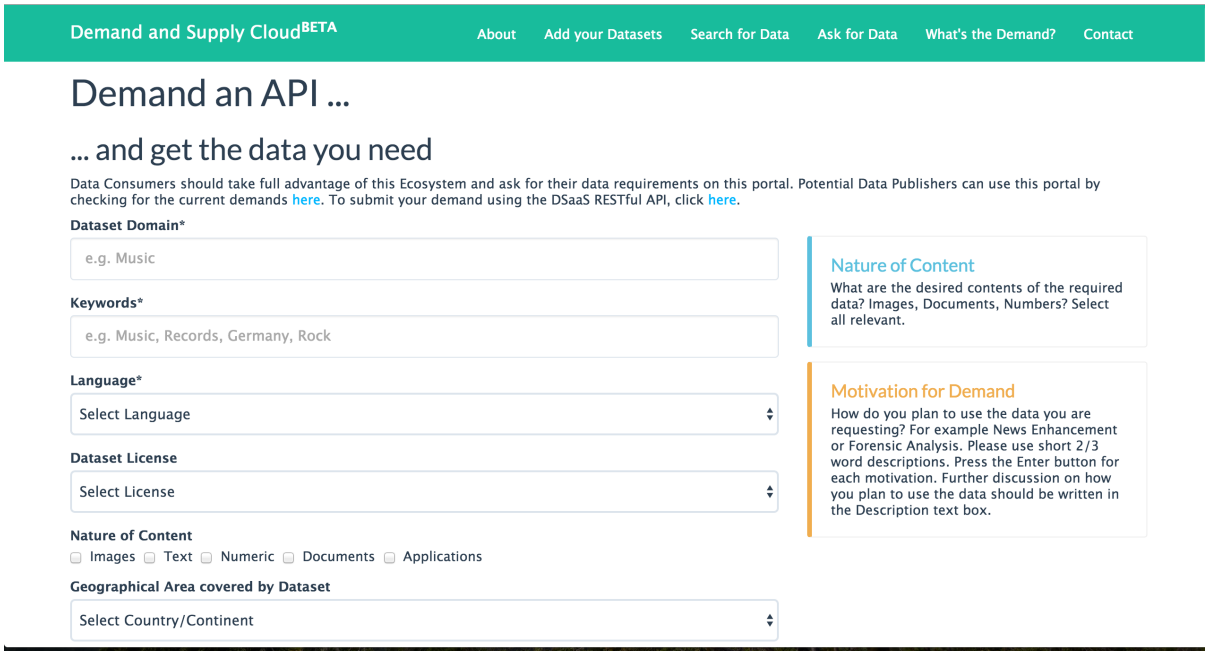

Figure 11.5: DSAAS: Adding a request for a new dataset.

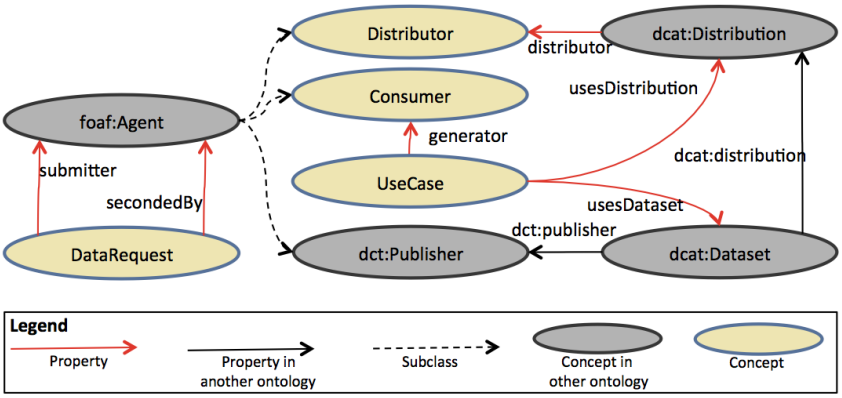

Figure 11.6: The main concepts in the Demand and Supply Ontology (DSO).

In order to best represent the supply and demand of data, we defined the **Demand and Supply Ontology** (DSO - available online at https://w3id.org/dso) to act as the underlying schema to the DSAAS. We re-use existing concepts from DCAT[12], Dublin Core, and FOAF to ensure interoperability and easier interlinking. The DSO improves upon existing schemas and initiatives such as DCAT, CKAN, and Datahub[13] in that it enables us to represent not only the dataset in question (resulting in a catalogue of datasets), but also all the involved actors, as well as their relationships with the data at hand, hence providing some context on the provenance of the data. Moreover, the DSO also enables us to represent the context of the re-use of a dataset. This representation allows us to holistically portray the picture of the supply and demand within a data market. It also enables us to store and publish data using Linked Data principles. The core concepts of the DSO (shown in Figure 11.6) are:

[12] https://www.dcat.org/ (Date accessed: 2 August 2016)
[13] https://datahub.io/ (Date accessed: 2 August 2016)

- **Dataset** - This concept represents information about existing datasets that can be consumed. This concept has properties that describe the dataset at hand, including the licence, theme, language, temporal coverage, description, date issued, data modified, etc.
- **Distribution** - This concept reflects information about the access method of a Dataset. Hence, a dataset can have one or more Distributions. A Distribution has properties that describe the data format, the access and download URLs, the date modified, etc.
- **UseCase** - A UseCase is an example of a 'success story' of the use of a Dataset. Properties of a UseCase include a description of the use case, the distribution and the dataset it uses, the stakeholders' motivation, etc.
- **DataRequest** - This concept encompasses similar details to a Dataset concept, except that this concept reflects the *need* for a dataset.
- **Publisher** - A Publisher is the entity or stakeholder responsible for publishing (and possibly also creating) a Dataset.
- **Consumer** - A Consumer is responsible for generating a UseCase through consuming an existing dataset.
- **Distributor** - A Distributor is the entity responsible for creating a distribution of a Dataset. For example, whilst the original dataset was created by a governmental entity on paper, a Distributor would be the one responsible for creating an RDF version of the same data.
- **Agent** - Apart from being a super-class to the Publisher, Consumer, and Distributor concepts, an Agent is an entity who is responsible for submitting a DataRequest, or otherwise seconding an existing one.

11.2.1 Demand and Supply as a Service in Use

With the aim of determining the potential benefits of using the DSAAS, we lead out a preliminary evaluation where a number of data producers and consumers were requested to fill out a survey, available in Appendix C. The results are available online at `http://eis.iai.uni-bonn.de/Projects/Demand_and_Supply_as_a_Service.html`. At the moment of writing, 15 persons responded the survey, of which 10 are both data consumers and publishers, whilst 4 are only data consumers and 1 is only a data publisher. When asked about the most common challenges in consuming open datasets, the respondents of the survey indicated that the low discoverability of the dataset in question is the challenge they faced most (14), with the lack of provenance and licence information being a tied second (11 each), uncertainty whether a dataset even existed followed (10), lack of use cases of previous use (9), and finally one evaluator identified as a challenge the varying quality of datasets (1). These challenges are in line with aims of the DSAAS, in that we provide the essential data to enhance dataset discoverability, as well as provenance, licence, and use case information. For the rest of the questions in the survey, which directly concern the foreseen benefits of using the DSAAS, we use the Likert scale to evaluate the degree to which the evaluators agree with the specified benefits. The results, also shown in Figure 11.7, are as follows (SA-Strongly Agree, A-Agree, N-Neither agree not disagree, D-Disagree, SD-Strongly Disagree):

- The DSAAS can help stakeholders to easily identify the demand in the data market (by enabling stakeholders to submit data requests) - Figure 11.7 (a) - SA:1, A:9, N:2, D:2, SD:1

- The DSAAS can help stakeholders to easily identify the supply in the data market (by listing existing datasets) - Figure 11.7 (b) - SA:2, A:10, N:1, D:0, SD:2
- The DSAAS can help stakeholders to identify a niche in the data market, and hence target it specifically (through catering for a data request) - Figure 11.7 (c) - SA:2, A:7, N:4, D:2, SD:0
- Stakeholders are encouraged to re-use datasets if success stories (use cases) are provided - Figure 11.7 (d) - SA:4, A:8, N:2, D:0, SD:1
- The DSAAS encourages stakeholders to collaborate with each other by showing their interests in specific dataset domains - Figure 11.7 (e) - SA:4, A:5, N:5, D:1, SD:0
- The DSAAS would be a good tool to showcase datasets and encourage their consumption - Figure 11.7 (f) - SA:5, A:8, N:1, D:0, SD:1
- By allowing consumers to put a request for a dataset, the DSAAS could possibly make the acquirement process faster - Figure 11.7 (g) - SA:3, A:11, N:1, D:0, SD:0

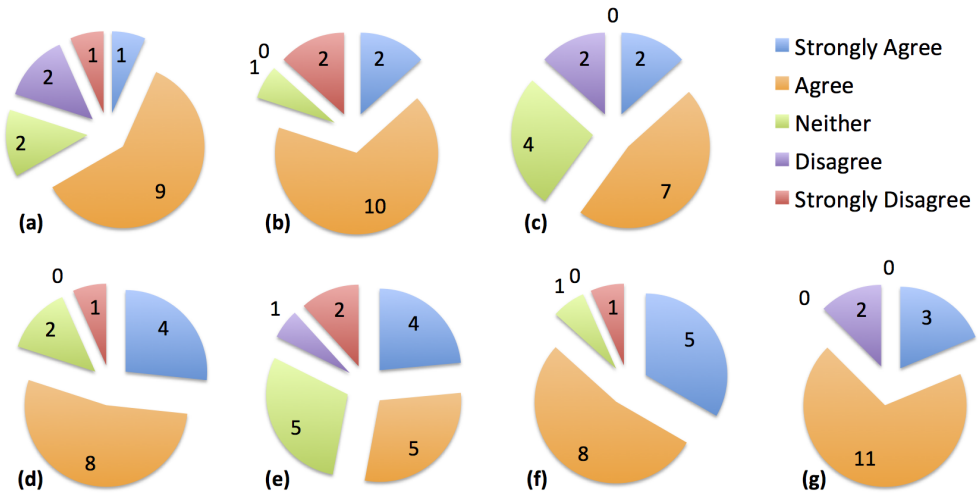

Figure 11.7: Pie charts of the results for the preliminary survey.

Through this evaluation we can conclude that overall the survey responders agree with the benefits of using the DSAAS that we portray. Ten out of twelve respondents agree that the DSAAS encourages dataset sharing and consumption, and eleven respondents agree that it can help the data acquirement process. Whilst there are varying opinions on the benefits, the majority of the respondents always agree that the tool will improve their participation in the data market. These results, while not conclusive, certainly indicate the potential of our approach.

In order to further establish its validity, the DSAAS is already being used in the ODINE Project[14,15]. The latter is an open data incubator that provides access to hundreds of companies and SMEs working on open data businesses. Starting in July 2015 (and ending in August 2016), the calls for such entities

[14] https://opendataincubator.eu/ (Date accessed: 2 August 2016)

[15] https://www.theguardian.com/technology/2014/nov/04/eu-commits-144m-to-support-open-data-across-europe (Date accessed: 2 August 2016)

have attracted the participation of over 300 companies or SMEs. Use cases and datasets used within the latter SMEs are being fed into the DSAAS knowledge base, hence creating a network of connections and collaborations between the datasets and their producers/consumers. We envisage that once the knowledge base is more substantial, we can also provide additional functions, such as a crowdsourced effort for knowledge base curation, importing of existing data catalogs, and the provision of a recommender system built on top of the knowledge base.

Concluding Remarks for Part IV: Value Creation as an Exploitation Strategy

The increasing dataification within our information society has required the need for the specification and implementation of new value chains. The main challenge in creating value is that open data has no value in itself, yet it becomes valuable when it is used. In Chapter 8 we describe how our information society value creation processes have the potential of extracting the maximum value from data by building on its intelligent use. Yet, existing definitions of value chains hardly manage to adequately represent the fluid, interconnected but independent processes of creating value on a data product.

In response to the following research question:

Research Question 3:

What aspects and processes play a role in value creation on a data product?

we proceed to identify the various processes that create value on a data product. With the aim of projecting our vision of generating a new *Economic Data Ecosystem* based on data value chains, in Chapter 9 we propose the *Data Value Network* (DVN). The DVN models the co-production of value through the interaction of a number of actors who participate through a number of roles. All stakeholders of value creation can participate through different roles, yet they have one common goal; that of creating a data product. Different dimensions impact the creation of such a product, namely technical, policy/legal, economic/financial, organisational, and social/cultural. Some of these dimensions are in turn also impacted by value creation. The use of Linked Data in creating value enhances the process, and also aids us to gradually proceed through various types of data products: starting with data, to information, and ultimately to knowledge.

In order to assess the value creation process of an open data initiative, in Chapter 10 we propose an assessment framework that focuses on the potential impact achievable from a data product generated through a value creating process, and implement it on a high level on two government data portals. Whilst this assessment framework can currently provide a good indication, it is as yet not very accurate, however the implementation of appropriate metrics can considerably enhance the accuracy of the results. Moreover, the framework also needs to be used to assess different types of open data initiatives, to ensure its compatibility with different types of initiatives and validity in assessing their value creating potential. Step by step the vision of having open data exploited to its full potential can be acquired.

In Chapter 11 we also define the *Demand and Supply Distribution Model*, which provides an insight on how an entity can successfully enter the global data market, whilst maintaining a competitive edge. The Demand and Supply as a Service (DSAAS) application then acts as concrete implementation of the proposed model. Acting as a dynamic leveller, this service enables stakeholders to more easily advertise existing data products, or otherwise create a request for specific data. This match-making service has the potential of creating a sustainable environment of data re-use, enhancing the value creation cycle within the DVN.

Whilst still not fully implemented, the Demand and Supply as a Service has the potential to be so much more than a catalogue providing details about existing datasets. Here we provide an outline of planned improvements and additional features to be implemented:

- Forum - The provision of a forum will enable stakeholders to discuss and collaborate in the publishing and consumption of data. Ideas on how to best exploit data can be shared and the data can be further exploited to its full potential.
- Data Mashups - Datasets from different domains can be mashed up in order to achieve further context and insight. The provision of such enhanced datasets through APIs would make them more accessible and discoverable, ultimately affecting their re-use.
- Recommendations - The stored information can be exploited to provide recommendations to consumers based on their interests. For example, if a consumer is interested in weather data, then datasets tagged as being weather-related can be suggested to the consumer.
- Bridge Service - The Demand and Supply as a Service can be used to act as a bridge to other data portals or catalogues. If a consumer enters his interest in a specific dataset, we can query datasets listed within other datastores and suggest the results to the user directly through our service.

Part V

Epilogue

CHAPTER 12

Conclusion

Information and data products are prevalent aspects of our society that affect a large number of dimensions. Valuable data is being generated at an astonishing pace, however in many cases the data is not published in a way that it is accessible for re-use by possible consumers. This acts as a disincentive for many stakeholders from participating in open data initiatives, and therefore many achievable benefits of data re-use are lost. It is therefore extremely important to tackle such issues at the root. Once existing data is made available for re-use, stakeholders are then able to exploit the data and create value upon it, therefore even increasing the potential resulting benefits and impacts. Value creation on a data product can be defined to be the manipulation of data with the aim to make it more fit for the intended use. The resulting consumption of this data with added value has then the potential to form the basis for many innovative information products and services, impacting the knowledge economy in the process. All data, whether addresses of schools, geospatial data, environmental data, weather data, transport and planning data, or budget data, has social and commercial value, and its use will result in a number of different impacts.

Following the ever-rising importance of this data-pervasion, in this thesis our research was targeted towards answering the following research question, as proposed in Section 1.2:

> *What strategies, methods and technologies can be used to maximise the exploitation of open data?*

Starting with a systematic survey on open government initiatives, in Part II we research existing initiatives and identified key aspects that determine their success in publishing data and enabling its consumption. This provided a good foundation for the rest of the thesis, which focuses more on the actual exploitation of data. In Part III we therefore identify and cater for the lack of approaches that can be used by non-expert consumers to re-use data. In this part we provided the relevant tools which enable such consumers to more easily exploit open data without requiring significant background knowledge on the underlying processes. Finally, in Part IV we focus specifically on the value creating processes that stakeholders can undertake to enhance data products and exploit them to their full potential, whilst also providing insight on how stakeholders can participate in the global data economy.

Through the research and contributions provided in this thesis we are bringing the benefits of exploiting open data closer to stakeholders. The latter can be any entity, and vary from governments and public entities, to private individuals, SMEs, businesses, and non-profit organisations. We provide detailed

information on how the re-use of data can impact a number of different dimensions within society. This provides stakeholders with reasonable understanding on the importance of data and motivation to participate in open data initiatives. In this thesis we also provide guidelines and tools that enhance the publication and consumption processes. The latter processes are crucial in any open data initiative, as such an initiative cannot even exist without the initial existence of data and its subsequent use. Moreover, by identifying challenges that (i) hinder an open data initiative from reaching its full potential, (ii) hinder data from being truly open, and (iii) discourage stakeholders from joining an open data initiative, we make stakeholders more aware of the challenges they might face and also how they can mitigate them.

Once data is published in its ideal state to be exploited, and stakeholders, particularly non-experts, have the tools to do so, value can be created on top of data to obtain a data product. As opposed to traditional value chains on more tangible products, the value creation process on data has a number of advantages, including the fact that data can be re-used over and over until it remains relevant, the flexibility in the order that value creation process are executed, and most importantly, the fact that data can be consumed at any state during the value creation process, as value creation is actually ongoing throughout the relevancy of data. These advantages enable the value of data to be exploited any number of times by different stakeholders. By identifying the various value creating processes, we provide stakeholders with the possibility to more easily identify and target the value creation process that is ideal in their context and potentially provide them with a competitive edge upon other competing stakeholders, whether they are using data to improve their product or service, or whether they are using data as a product within itself. Our Demand and Supply as a Service also helps balance out the data demand and supply such that data producers can specifically target to produce data that is pointedly required by consumers, making it easier to participate in a Data Value Network and create value on data.

12.1 Answering the Research Questions

In this section we go through the research sub-questions defined in Section 1.2 and summarise the contributions of this thesis in response.

Research Question 1:

What are existing approaches and techniques that enable the publishing and consumption of open data?

There are numerous approaches for publishing and consuming open data. In the context of open government data, such approaches are commonly data portals, catalogues, and other services. The latter approaches have the common functionality of making the published data available, and enable stakeholders to consume it. However, even though there exist numerous guidelines, there are no agreed-upon standards for the publication and consumption of open data. Moreover, a number of challenges hinder both stakeholders from participating in an open data initiative, and also the initiative itself from achieving its full potential. If this potential is appropriately harnessed and exploited, open data can provide a large number of direct and indirect impacts and benefits to the relevant stakeholders. Our contributions towards this research question include an in-depth discussion of open government initiatives, including impacts, challenges, assessment frameworks, and participating stakeholders, the proposal of guidelines, the exploration of existing approaches, tools, and standards for publishing and consuming open data, and the exploration of an open budget data initiative as a use case.

Research Question 2:

How can we enhance the consumption process of a data product in order to enable further value creation?

Most approaches in the context of (linked) open data are focused on enabling the publishing of data and making the required processes easier to execute. Less attention was afforded towards consuming approaches, and most of the latter require a certain amount of expertise, including knowledge about RDF or the Linked Data paradigm and the SPARQL query language. Hence, after identifying strengths and weaknesses of existing consuming approaches, we devised a framework that enables non-expert stakeholders to easily consume open data without requiring such knowledge. In this contribution we obscured the underlying processes in such a way that consumers can more simply identify and re-use the required data to create value, without being overwhelmed. Using the ExConQuer framework, consumers can very easily query RDF data and transform the results into a number of different formats. We retain the semantic richness of the underlying RDF data model through the ConQuer ontology, which also enables previously-executed queries and transformations to be re-used or modified to get different results. As evidenced by the evaluations we lead on the framework, the provided tools are deemed to be particularly useful to encourage the re-use of Linked Data by non-experts. Our framework hence paves the way towards making open data more accessible for exploitation and value creation.

Research Question 3:

What aspects and processes play a role in value creation on a data product?

In answer to this question we identify the different processes that can be undertaken on data products to make them more fit for use (whatever the context). We thus define a Data Value Network (DVN) that acts as a model of a number of independent processes that have the aim of creating value upon data in order to exploit it as a product. As opposed to more traditional value chains, the one we propose is data-specific, in that is specifically tailored for digital products. We also delineate the various roles through which stakeholders can participate in the DVN to create value. Hence any interested stakeholders can identify the role and processes that best fit their context for creating value on a data product. To further guide stakeholders to create value, we define the Demand and Supply Distribution Model. This model portrays how stakeholders, more specifically data producers/publishers and consumers, can participate in a DVN to create value and consequently exploit data. Through the concrete implementation of this model as an online service, we also encourage data producers/publishers to advertise their datasets, whilst consumers can easily find the data they need, or otherwise ask for it. Finally, in order to encourage the publishing of data with a high value creation potential, we provide an assessment framework. Here we specify a number of aspects that should be assessed to evaluate value creation *potential* of the data in question.

12.2 Future Directions

There are various opportunities for future directions as a continuation to the contributions presented in this thesis, as well as improvements upon current limitations. In this section we therefore explore a number of possible directions, whilst keeping in line with the aim and motivation of this thesis.

12.2.1 Short-Term Directions

In this thesis we focused on strategies and approaches for exploiting the value of open data. Through our contributions we provided guidelines and solutions on how data can be re-used and exploited as a

data product. This aids and motivates stakeholders to actually participate in an open data initiative and innovate upon the data. We here discuss further improvements to our contributions in this thesis, with the aim of furthering our aim of exploiting open data to its highest potential.

- **Investigating all the processes of the government data life cycle**
  In this thesis we limit our research to the publishing and consumption of data, as they are the most important processes within a life cycle. This is because without the availability of data, and its consumption, then the life cycle would not exist. Yet, this does not mean that the other processes are not important. For instance, the production of data is very important to the eventual creation of value on data as some data is more useful and can potentially yield more value than others [161]. The investigation of all the processes within the data life cycle will consequently provide a more wholesome picture. The identification of challenges and possible solutions, along with other guidelines, will ultimately provide stakeholders with the best opportunity to make an open government data initiative succeed and obtain the highest amount of benefits possible.

- **Adapting the ExConQuer framework to cater for different stakeholders and increase its functionality**
  Currently, the ExConQuer framework caters for non-experts and is somewhat limited with regard to the results that can be achieved. Further research can enable us to optimise this tool to enable both experts and non-experts to have an entry point to consuming open data which is suitable for more requirements. For instance, the tool can be improved by enabling the generation and execution of SPARQL queries other than SELECT. A current limitation of the tool is that it requires the explicit definition of classes and properties through the *owl:Class*, *rdfs:Class*, *owl:DataTypeProperty*, and *owl:ObjectProperty*. Moreover, if the properties do not have a domain and range, then they will not be previewed by the tool since we cannot identify if there is a relationship between such a property and any classes. These definitions should also be accessible through the datasource's SPARQL endpoint in order for the tool to correctly preview the contents of the dataset. These limitations can all be solved if the tool identifies classes and properties based on the actual instances in the datasource, rather than through its schema, although this might result in a less-efficient tool.

- **Validation of the Value Creation Assessment Framework**
  The current version of the assessment framework we propose is as yet not fully validated. Its evaluation on different open data initiatives will ensure it accurately represents the relevant aspects that affect the final impact of a data product after value is created on it. The aspects within the framework will then provide a more accurate guideline for evaluating existing open data initiatives with regard to how well they enable stakeholders to re-use and create value on the data. Moreover, more accurate metrics to assess the specified aspects will provide more precise results and give a better indication of the potential impacts of creating value on the data in question.

12.2.2 Long-Term Directions

The publishing of data in order to make it publicly available is already quite popular, particularly in the context of open government, where the main motivation for publishing data is transparency and accountability. Yet, the value of data lies in its re-use. Various studies, such as [23], [34], and [83], have suggested an estimate on the value of open data. More specifically, Manyika et al. [83] estimate that open data can help unlock between 3 to 5 trillion U.S. Dollars in economic value annually, the authors of [34] estimate that the value of public sector information (government data) to consumers, businesses and the public sector in 2011/12 was approximately 1.8 billion Pounds, whilst Carrara et al. [23] estimated

the (total) market size of open data in the European Union to be between 193 and 209 billion Euro for 2016. The authors of [34] also indicate that there is a link between the provision and (re-)use of public sector information and economic growth, where the benefits include increased efficiency, development of new products and services, cost savings, and better quality products. We here explore various future directions that could act as a continuance to the contributions in this thesis.

- **Measuring the impact of open (government) data**
  A plethora of publications focus on the challenges and benefits of open (government) data. They also provide guidelines on how to best lead out an open data initiative. Yet, few publications measure the actual results and impacts of open data initiatives, such as the actual number of new jobs, any economic or societal benefits, political impact, etc. For example, Viscusi et al. [151] provide a classificatory framework having the aim of assessing the social value of open government initiatives, however they do not consider other types of impact or value. Therefore, additional research on the topic could be useful to determine which impact dimensions to assess, and the appropriate metrics to do so. Admittedly, some impacts are quite tough to measure accurately since they might depend on a large number of different factors or are otherwise non quantifyable. For example, if there is an economic boost in a country, it would be quite difficult to determine to what extent did open data contribute to this benefit. However, even an estimated quantitative indication of the benefits resulting from open data will provide stakeholders with a more tangible notation and therefore act as a motivation for participation in open data initiatives.

- **Measuring the impact of value creation**
  As an extension to measuring the impact of open data (i.e. the simple provision of data for public use), it is also vital to measure the impacts of creating value on data (e.g. improving quality, using open data to aid decision-making, etc). As opposed to the Value Creation Assessment Framework we proposed in this thesis, we here have a different motivation. Rather than measuring the *potential* for the resulting impacts (i.e. how should an open data initiative provide data in order to best enable its re-use for value creation), we here would measure the actual impacts of creating value. The eventual quantification of the resulting impacts will provide stakeholders with a concrete appraisal on why value creation is important and portray the real resulting benefits. Such a study would probably best be led over a longer period of time, and be based upon the monitoring of specific initiatives and the involved stakeholders. This is because certain impacts of value creation, such as the aftermath of data-based decision-making and knowledge extraction, can only be calculated after an adequate amount of time has passed.

- **Data literacy**
  This term can be can be defined as the skills required to use and analyse data. Whilst an astonishing amount of data is available for re-use, not all stakeholders own the necessary skills to properly be able to understand and analyse this data. Following the Data Literacy workshop[1] we co-organised with the WebScience 2015 conference, we would like to further explore this topic in context of how data literacy would affect the value creating processes on a data product. The workshop was the ideal environment to discuss the various challenges and implications of data literacy, as well as to set an agenda for future directions on the topic. Through the workshop it was also quite evident that the topic has as yet received very little academic attention. This further motivates us to research this recently-established topic.

[1] http://www.dataliteracy.eita.org.br/1st-dl-workshop/ (Date accessed: 8 August 2016)

Bibliography

[1] C. Alexopoulos, L. Spiliotopoulou and Y. Charalabidis,
Open Data Movement in Greece: A Case Study on Open Government Data Sources,
Proceedings of the 17th Panhellenic Conference on Informatics, PCI '13,
New York, NY, USA: ACM, 2013, p. 279,
URL: http://doi.acm.org/10.1145/2491845.2491876 (cit. on pp. 20, 29, 31).

[2] C. Alexopoulos et al., *Designing a Second Generation of Open Data Platforms: Integrating Open Data and Social Media*, Electronic Government - 13th {IFIP} {WG} 8.5 International Conference, {EGOV} 2014, Dublin, Ireland, September 1-3, 2014. Proceedings, 2014, p. 230,
URL: http://doi.org/10.1007/978-3-662-44426-9_19
(cit. on pp. 14, 21, 34, 38, 45, 46, 93, 113).

[3] R. Amit and C. Zott, *Creating Value Through Business Model Innovation*,
MIT Sloan Management Review vol. **53** (2012),
URL: http://sloanreview.mit.edu/article/creating-value-through-business-model-innovation/ (cit. on p. 95).

[4] J. Arcelus, *Framework for Useful Transparency Websites for Citizens*, Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance, ICEGOV '12,
New York, NY, USA: ACM, 2012, p. 83,
URL: http://doi.acm.org/10.1145/2463728.2463749 (cit. on pp. 27, 28, 33).

[5] J. Attard, F. Orlandi and S. Auer,
ExConQuer Framework - Softening RDF Data to Enhance Linked Data Reuse,
Proceedings of the ISWC 2015 Posters & Demonstrations Track co-located with the 14th International Semantic Web Conference (ISWC-2015), Bethlehem, PA, USA, October 11, 2015.,
2015, URL: http://ceur-ws.org/Vol-1486/paper_39.pdf (cit. on p. 8).

[6] J. Attard, F. Orlandi and S. Auer,
Data Driven Governments: Creating Value through Open Government Data,
Transactions on Large-Scale Data- and Knowledge-Centered Systems, ed. by D. Kalisch et al.,
Berlin, Heidelberg: Springer Berlin Heidelberg, 2016 (cit. on p. 8).

[7] J. Attard, F. Orlandi and S. Auer, *Data Value Networks: Enabling a New Data Ecosystem*,
IEEE/WIC/ACM International Conference on Web Intelligence, 2016, Omaha, Nebraska, USA, October 13-16, 2016, 2016 (cit. on pp. 8, 9).

[8] J. Attard, F. Orlandi and S. Auer, *Value Creation on Open Government Data*,
In Proceedings of the 49th Hawaii International Conference on System Sciences, HICSS 2016,
Kauai, HI, USA, January 5-8, 2016, 2016, p. 2605,
URL: http://dx.doi.org/10.1109/HICSS.2016.326 (cit. on p. 8).

[9] J. Attard et al., *A systematic review of open government data initiatives*, Government Information Quarterly vol. **32**.4 (2015), p. 399, URL: http://doi.org/10.1016/j.giq.2015.07.006 (cit. on p. 8).

[10] J. Attard et al., *ExConQuer: Lowering barriers to RDF and Linked Data re-use*, Semantic Web Journal (To appear, accepted 12 October 2016), URL: http://www.semantic-web-journal.net/content/exconquer-lowering-barriers-rdf-and-linked-data-re-use-0 (cit. on p. 8).

[11] T. Bakıcı, E. Almirall and J. Wareham, *A Smart City Initiative: the Case of Barcelona*, Journal of the Knowledge Economy vol. **4**.2 (2013), p. 135, URL: http://dx.doi.org/10.1007/s13132-012-0084-9 (cit. on p. 25).

[12] S. J. Barnes, *The mobile commerce value chain: analysis and future developments*, International Journal of Information Management vol. **22**.2 (Apr. 2002), p. 91, URL: http://doi.org/10.1016/S0268-4012(01)00047-0 (cit. on p. 95).

[13] N. Beghin and C. Zigoni, *Measuring Open Data's Impact of Brazilian National and Sub-National Budget Transparency Websites and its Impacts on Peoples's rights*, tech. rep., Institute for Socioeconomic Studies (INESC), 2014, URL: http://www.opendataresearch.org/sites/default/files/publications/Inesc_ODDC_English.pdf (cit. on p. 51).

[14] K. Belhajjame et al., *PROV-O: The PROV Ontology*, tech. rep., W3C, 2012, URL: http://www.w3.org/TR/prov-o/ (cit. on p. 82).

[15] J. Benington, *From private choice to public value*, Public value: Theory and practice, ed. by J. Benington, ed. by M. Moore, Palgrave Macmillan, 2011, p. 31 (cit. on p. 108).

[16] J. C. Bertot, B. S. Butler and D. Travis, *Local big data: the role of libraries in building community data infrastructures*, 15th Annual International Conference on Digital Government Research, dg.o '14, Aguascalientes, Mexico, June 18-21, 2014, 2014, p. 17, URL: http://doi.acm.org/10.1145/2612733.2612762 (cit. on pp. 25, 35, 39).

[17] C. Bizer, T. Heath and T. Berners-Lee, *Linked Data - The Story So Far*, International Journal on Semantic Web and Information Systems vol. **5**.3 (2009), p. 1, URL: http://doi.org/10.4018/jswis.2009081901 (cit. on pp. 20, 21, 68).

[18] S. Bogdanović-Dinić, N. Veljković and L. Stoimenov, *How Open Are Public Government Data? An Assessment of Seven Open Data Portals*, Measuring E-government Efficiency, vol. 5, Public Administration and Information Technology, Springer New York, 2014, p. 25, URL: http://doi.org/10.1007/978-1-4614-9982-4_3 (cit. on pp. 28, 29, 115).

[19] C. Böhm et al., *GovWILD: Integrating Open Government Data for Transparency*, Proceedings of the 21st International Conference Companion on World Wide Web, WWW '12 Companion, New York, NY, USA: ACM, 2012, p. 321, URL: http://doi.acm.org/10.1145/2187980.2188039 (cit. on pp. 38, 67).

[20] M. Bovens, *Analysing and Assessing Accountability: A Conceptual Framework*, European Law Journal vol. **13**.4 (2007), p. 447, URL: http://dx.doi.org/10.1111/j.1468-0386.2007.00378.x (cit. on p. 34).

[21] S. Campinas, *Live SPARQL Auto-Completion.*,
International Semantic Web Conference (Posters and Demos), 2014, p. 477,
URL: http://dl.acm.org/citation.cfm?id=2878573 (cit. on p. 71).

[22] R. Caplan et al.,
Towards common methods for assessing open data: workshop report & draft framework,
tech. rep., Workshop on common methods for assessing open data, 2014,
URL: http://opendataresearch.org/sites/default/files/posts/Common%20Assessment%20Workshop%20Report.pdf (cit. on p. 51).

[23] W. Carrara et al., *Creating Value through Open Data*, 2015,
URL: http://www.europeandataportal.eu/sites/default/files/edp_creating_value_through_open_data_0.pdf (cit. on pp. 3, 107, 138).

[24] J. J. Carroll et al., *Named Graphs, Provenance and Trust*,
Proceedings of the 14th International Conference on World Wide Web, WWW '05,
New York, NY, USA: ACM, 2005, p. 613,
URL: http://doi.acm.org/10.1145/1060745.1060835 (cit. on p. 38).

[25] C. M. L. Chan, *From Open Data to Open Innovation Strategies: Creating E-Services Using Open Government Data*, 2014 47th Hawaii International Conference on System Sciences,
Los Alamitos, CA, USA: IEEE Computer Society, 2013, p. 1890,
URL: http://doi.org/10.1109/HICSS.2013.236 (cit. on pp. 35, 36).

[26] L. Clark, *SPARQL Views: A Visual SPARQL Query Builder for Drupal*,
9th International Semantic Web Conference (ISWC2010), Nov. 2010,
URL: http://dl.acm.org/citation.cfm?id=2878440 (cit. on p. 71).

[27] P. Conradie and S. Choenni,
Exploring Process Barriers to Release Public Sector Information in Local Government,
Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance, ICEGOV '12, New York, NY, USA: ACM, 2012, p. 5,
URL: http://doi.acm.org/10.1145/2463728.2463731
(cit. on pp. 4, 14, 36–38, 40, 102, 106).

[28] D. Crié and A. Micheaux,
From customer data to value: What is lacking in the information chain?, en,
Journal of Database Marketing & Customer Strategy Management vol. **13**.4 (July 2006), p. 282,
URL: http://doi.org/10.1057/palgrave.dbm.3240306 (cit. on pp. 96, 97).

[29] N. Cukier and V. Mayer-Schoenberger, *The Rise of Big Data*, Foreign Affairs (2013),
URL: https://www.foreignaffairs.com/articles/2013-04-03/rise-big-data (cit. on pp. 4, 93).

[30] R. Cyganiak, F. Maali and V. Peristeras,
Self-service linked government data with dcat and gridworks,
Proceedings of the 6th International Conference on Semantic Systems - I-SEMANTICS '10,
New York, New York, USA: ACM Press, Sept. 2010, p. 1,
URL: http://doi.org/10.1145/1839707.1839754 (cit. on p. 68).

[31] T. Davies, *Supporting open data use through active engagement*,
Proceedings of the W3C Using Open Data Workshop, Brussels, 2012, p. 1,
URL: https://www.w3.org/2012/06/pmod/pmod2012_submission_5.pdf
(cit. on p. 58).

[32] T. Davies and M. Frank, *'There's No Such Thing As Raw Data': Exploring the Socio-technical Life of a Government Dataset*, Proceedings of the 5th Annual ACM Web Science Conference, WebSci '13, New York, NY, USA: ACM, 2013, p. 75, URL: http://doi.acm.org/10.1145/2464464.2464472 (cit. on pp. 37, 38).

[33] J. Debattista, C. Lange and S. Auer, *Representing Dataset Quality Metadata Using Multi-dimensional Views*, Proceedings of the 10th International Conference on Semantic Systems, SEM '14, New York, NY, USA: ACM, 2014, p. 92, URL: http://doi.acm.org/10.1145/2660517.2660525 (cit. on p. 48).

[34] Deloitte, *Market Assessment of Public Sector Information*, tech. rep., 2013, URL: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/198905/bis-13-743-market-assessment-of-public-sector-information.pdf (cit. on pp. 138, 139).

[35] D. DiFranzo et al., *The Web is My Back-end: Creating Mashups with Linked Open Government Data*, Linking Government Data, ed. by D. Wood, Springer New York, 2011, p. 205, URL: http://dx.doi.org/10.1007/978-1-4614-1767-5_10 (cit. on pp. 46, 109).

[36] K. dos Santos Brito et al., *Brazilian Government Open Data: Implementation, Challenges, and Potential Opportunities*, Proceedings of the 15th Annual International Conference on Digital Government Research, dg.o '14, New York, NY, USA: ACM, 2014, p. 11, URL: http://doi.acm.org/10.1145/2612733.2612770 (cit. on pp. 29, 31).

[37] K. dos Santos Brito et al., *Using Parliamentary Brazilian Open Data to Improve Transparency and Public Participation in Brazil*, Proceedings of the 15th Annual International Conference on Digital Government Research, dg.o '14, New York, NY, USA: ACM, 2014, p. 171, URL: http://doi.acm.org/10.1145/2612733.2612769 (cit. on pp. 29, 31, 33).

[38] K. Dos Santos Brito et al., *Experiences Integrating Heterogeneous Government Open Data Sources to Deliver Services and Promote Transparency in Brazil*, Computer Software and Applications Conference (COMPSAC), 2014 IEEE 38th Annual, July 2014, p. 606, URL: http://doi.org/10.1109/COMPSAC.2014.87 (cit. on pp. 38, 41, 42).

[39] T. Dyba, T. Dingsoyr and G. K. Hanssen, *Applying Systematic Reviews to Diverse Study Types: An Experience Report*, Proceedings of the First International Symposium on Empirical Software Engineering and Measurement, ESEM '07, Washington, DC, USA: IEEE Computer Society, 2007, p. 225, URL: http://dx.doi.org/10.1109/ESEM.2007.21 (cit. on p. 15).

[40] S. Eckartz, W. Hofman and A. Van Veenstra, *A Decision Model for Data Sharing*, Electronic Government, ed. by M. Janssen et al., vol. 8653, Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2014, p. 253, URL: http://dx.doi.org/10.1007/978-3-662-44426-9_21 (cit. on pp. 37, 40).

[41] N. Edelmann, J. Höchtl and M. Sachs,
Collaboration for Open Innovation Processes in Public Administrations.,
Empowering Open and Collaborative Governance, ed. by Y. Charalabidis and S. Koussouris,
Springer, 2012, p. 21, URL: http://doi.org/10.1007/978-3-642-27219-6_2
(cit. on pp. 33–35, 39, 45).

[42] I. Egger-Peitler and T. Polzer,
Open Data: European Ambitions and Local Efforts. Experiences from Austria,
Open Government, ed. by M. Gascó-Hernóndez, vol. 4,
Public Administration and Information Technology, Springer New York, 2014, p. 137,
URL: http://doi.org/10.1007/978-1-4614-9563-5_9 (cit. on pp. 29, 31).

[43] European Comission, *EU Anti-Corruption Report*, tech. rep., European Comission, 2014,
URL: http://ec.europa.eu/dgs/home-affairs/e-library/documents/policies/organized-crime-and-human-trafficking/corruption/docs/acr_2014_en.pdf (cit. on p. 49).

[44] M. Foulonneau, S. Martin and S. Turki, *How Open Data Are Turned into Services?*,
Exploring Services Science, ed. by M. Snene and M. Leonard, vol. 169,
Lecture Notes in Business Information Processing, Springer International Publishing, 2014,
p. 31, URL: http://doi.org/10.1007/978-3-319-04810-9_3
(cit. on pp. 25, 35, 39).

[45] R. Fuentes-Enriquez and Y. Rojas-Romero, *Developing Accountability, Transparency and Government Efficiency Through Mobile Apps: The Case of Mexico*, Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance, ICEGOV '13,
New York, NY, USA: ACM, 2013, p. 313,
URL: http://doi.acm.org/10.1145/2591888.2591944
(cit. on pp. 25, 29, 30, 39).

[46] J. C. González et al., *Government 2.0: A Conceptual Framework and a Case Study Using Mexican Data for Assessing the Evolution Towards Open Governments*,
Proceedings of the 15th Annual International Conference on Digital Government Research,
dg.o '14, New York, NY, USA: ACM, 2014, p. 124,
URL: http://doi.acm.org/10.1145/2612733.2612742 (cit. on pp. 29, 30).

[47] K. Granickas, *Understanding the impact of releasing and re-using open*, tech. rep.,
European Public Sector Information Platform, 2013, p. 1, URL: https://www.europeandataportal.eu/sites/default/files/2013_understanding_the_impact_of_releasing_and_re_using_open_data.pdf (cit. on p. 51).

[48] M. B. Gurstein, *Open data: Empowering the empowered or effective data use for everyone?*,
First Monday vol. **16**.2 (2011), p. 1,
URL: http://dx.doi.org/10.5210/fm.v16i2.3316 (cit. on pp. 50, 59).

[49] F. Haag et al., *Visual SPARQL querying based on extended filter/flow graphs*,
International Working Conference on Advanced Visual Interfaces, 2014,
URL: http://doi.acm.org/10.1145/2598153.2598185 (cit. on p. 71).

[50] T. Heath, *How Will We Interact with the Web of Data?*, English,
IEEE Internet Computing vol. **12**.5 (Sept. 2008), p. 88,
URL: http://doi.org/10.1109/MIC.2008.101 (cit. on p. 68).

[51] T. Heath and C. Bizer, *Linked Data: Evolving the Web into a Global Data Space*, Synthesis Lectures on the Semantic Web, Morgan & Claypool Publishers, 2011, URL: http://dx.doi.org/10.2200/S00334ED1V01Y201102WBE001 (cit. on pp. 56, 109).

[52] P. Heim, T. Ertl and J. Ziegler, *Facet Graphs: Complex Semantic Querying Made Easy*, The Semantic Web: Research and Applications, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, URL: http://link.springer.com/10.1007/978-3-642-13486-9 (cit. on p. 70).

[53] J. Hendler et al., *US Government Linked Open Data: Semantic.data.gov*, IEEE Intelligent Systems vol. **27**.3 (2012), p. 25, URL: http://doi.ieeecomputersociety.org/10.1109/MIS.2012.27 (cit. on pp. 26, 37, 38).

[54] J. Höchtl and P. Reichstädter, *Linked Open Data - A Means for Public Sector Information Management.*, Electronic Government and the Information Systems Perspective - Second International Conference, EGOVIS 2011, Toulouse, France, August 29 - September 2, 2011. Proceedings, ed. by A. Kim Normann et al., vol. 6866, Lecture Notes in Computer Science, Springer, 2011, p. 330, URL: http://doi.org/10.1007/978-3-642-22961-9_26 (cit. on pp. 28, 42).

[55] W. Hofman and M. Rajagopal, *A Technical Framework for Data Sharing*, Journal of theoretical and applied electronic commerce research vol. **9** (2014), p. 45, URL: http://dx.doi.org/10.4067/S0718-18762014000300005 (cit. on p. 44).

[56] International Budget Partnership, *Open Budget Survey 2012*, tech. rep., International Budget Partnership, 2012, URL: http://www.internationalbudget.org/wp-content/uploads/OBI2012-Report-English.pdf (cit. on p. 51).

[57] V. Janev et al., *Modeling, Fusion and Exploration of Regional Statistics and Indicators with Linked Data Tools*, Electronic Government and the Information Systems Perspective, ed. by A. Kö and E. Francesconi, vol. 8650, Lecture Notes in Computer Science, Springer International Publishing, 2014, p. 208, URL: http://doi.org/10.1007/978-3-319-10178-1_17 (cit. on p. 46).

[58] K. Janssen, *The influence of the PSI directive on open government data: An overview of recent developments*, Government Information Quarterly vol. **28**.4 (Oct. 2011), p. 446, URL: http://doi.org/10.1016/j.giq.2011.01.004 (cit. on pp. 14, 93, 113).

[59] M. Janssen, Y. Charalabidis and A. Zuiderwijk, *Benefits, Adoption Barriers and Myths of Open Data and Open Government*, Information Systems Management vol. **29**.4 (2012), p. 258, URL: http://dx.doi.org/10.1080/10580530.2012.716740 (cit. on pp. 4, 6, 93, 106, 113).

[60] M. Janssen and J. van den Hoven,
Big and Open Linked Data (BOLD) in government: A challenge to transparency and privacy?,
Government Information Quarterly vol. **32**.4 (2015), p. 363,
URL: http://dx.doi.org/10.1016/j.giq.2015.11.007 (cit. on pp. 40, 107).

[61] M. Janssen and H. van der Voort,
Adaptive governance: Towards a stable, accountable and responsive government,
Government Information Quarterly vol. **33**.1 (2016), p. 1,
URL: http://dx.doi.org/10.1016/j.giq.2016.02.003 (cit. on p. 37).

[62] T. Jetzek, M. Avital and N. Bjørn-Andersen, *Generating Value from Open Government Data*,
Proceedings of the International Conference on Information Systems, {ICIS} 2013, Milano, Italy,
December 15-18, 2013, ed. by R. Baskerville and M. Chau,
Association for Information Systems, 2013, URL:
http://aisel.aisnet.org/icis2013/proceedings/GeneralISTopics/5
(cit. on p. 108).

[63] T. Jetzek, M. Avital and N. Bjørn-Andersen,
Generating Sustainable Value from Open Data in a Sharing Society,
Creating Value for All Through IT, ed. by B. Bergvall-Kåreborn and P. Nielsen, vol. 429,
IFIP Advances in Information and Communication Technology,
Springer Berlin Heidelberg, 2014, p. 62,
URL: http://doi.org/10.1007/978-3-662-43459-8_5
(cit. on pp. 20, 29, 31, 46).

[64] Z. Jiříček and F. Di Massimo,
Microsoft Open Government Data Initiative (OGDI), Eye on Earth Case Study,
Environmental Software Systems. Frameworks of eEnvironment,
ed. by J. Hřebíček, G. Schimak and R. Denzer, vol. 359,
IFIP Advances in Information and Communication Technology,
Springer Berlin Heidelberg, 2011, p. 26,
URL: http://dx.doi.org/10.1007/978-3-642-22285-6_3 (cit. on pp. 37, 45).

[65] J. M. Juran, *Juran's Quality Handbook*, 4th, Mcgraw-Hill (Tx), 1974 (cit. on pp. 21, 101).

[66] E. Kalampokis, M. Hausenblas and K. A. Tarabanis,
Combining Social and Government Open Data for Participatory Decision-Making., ePart,
ed. by E. Tambouris, A. Macintosh and H. de Bruijn, vol. 6847,
Lecture Notes in Computer Science, Springer, 2011, p. 36,
URL: http://doi.org/10.1007/978-3-642-23333-3_4 (cit. on p. 36).

[67] E. Kalampokis, E. Tambouris and K. Tarabanis,
A Classification Scheme for Open Government Data: Towards Linking Decentralised Data,
Int. J. Web Eng. Technol. vol. **6**.3 (June 2011), p. 266,
URL: http://dx.doi.org/10.1504/IJWET.2011.040725
(cit. on pp. 25, 41, 42, 46, 68).

[68] R. Kaplinsky and M. Morris, *A Handbook for Value Chain Research*, 2002,
URL: http://oro.open.ac.uk/5861/ (cit. on pp. 95, 107).

[69] B. Kitchenham, *Procedures for Performing Systematic Reviews*, tech. rep., Departament of Computer Science, Keele University, 2004, URL: http://people.ucalgary.ca/~medlibr/kitchenham_2004.pdf (cit. on p. 15).

[70] H. Knublauch, J. A. Hendler and K. Idehen, *SPIN - Overview and Motivation*, tech. rep., 2011, URL: http://www.w3.org/Submission/spin-overview/ (cit. on p. 82).

[71] J. Kučera, D. Chlapek and M. Nečaský, *Open Government Data Catalogs: Current Approaches and Quality Perspective*, Technology-Enabled Innovation for Democracy, Government and Governance, vol. 8061, Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2013, p. 152, URL: http://doi.org/10.1007/978-3-642-40160-2_13 (cit. on pp. 20, 21, 25, 26, 38, 46–48, 101–103).

[72] A. Latif et al., *The Linked Data Value Chain: A Lightweight Model for Business Engineers*, Proceedings of International Conference on Semantic Systems, 2009, p. 568, URL: http://www.know-center.tugraz.at/download_extern/papers/ldvc.pdf (cit. on pp. 96, 97).

[73] K. Layne and J. Lee, *Developing fully functional E-government: A four stage model*, Government Information Quarterly vol. **18**.2 (2001), p. 122, URL: http://doi.org/http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.181.950 (cit. on p. 20).

[74] C. C. Lee and J. Yang, *Knowledge value chain*, Journal of Management Development vol. **19**.9 (2000), p. 783, URL: http://dx.doi.org/10.1108/02621710010378228 (cit. on p. 96).

[75] R. Likert, *A technique for the measurement of attitudes.*, Archives of Psychology vol. **22** (1932), p. 1, URL: http://www.voteview.com/pdf/Likert_1932.pdf (cit. on p. 83).

[76] C. S. Lin and H.-C. Yang, *Data Quality Assessment on Taiwan's Open Data Sites*, Multidisciplinary Social Networks Research, ed. by L. S.-L. Wang et al., vol. 473, Communications in Computer and Information Science, Springer Berlin Heidelberg, 2014, p. 325, URL: http://dx.doi.org/10.1007/978-3-662-45071-0_26 (cit. on pp. 25, 29, 31).

[77] Q. Liu et al., *Linking Australian Government Data for Sustainability Science - A Case Study*, Linking Government Data, ed. by D. Wood, Springer New York, 2011, p. 181, URL: http://dx.doi.org/10.1007/978-1-4614-1767-5_9 (cit. on pp. 29, 30, 37, 38, 42, 44, 46).

[78] S. López-Ayllón and D. Arellano Gault, *Estudio en materia de transparencia de otros sujetos obligados por la Ley Federal de Transparencia y Acceso a la Información Pública Gubernamental*, Centro de Investigación y Docencia Económicas: Instituto Federal de Acceso a la Información: UNAM. Instituto de Investigaciones Jurídicas, 2008 (cit. on pp. 4, 33, 34).

[79] R. P. Lourenço,
Open Government Portals Assessment: A Transparency for Accountability Perspective., EGOV, ed. by M. Wimmer, M. Janssen and H. J. Scholl, vol. 8074, Lecture Notes in Computer Science, Springer, 2013, p. 62, URL: http://doi.org/10.1007/978-3-642-40358-3_6 (cit. on pp. 28, 34, 46, 101, 115).

[80] R. Lourenço and L. Serra, *An Online Transparency for Accountability Maturity Model*, Electronic Government, ed. by M. Janssen et al., vol. 8653, Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2014, p. 35,
URL: http://dx.doi.org/10.1007/978-3-662-44426-9_3 (cit. on p. 44).

[81] J. von Lucke and K. Große, *Open Government Collaboration*, Open Government, ed. by M. Gascó-Hernández, vol. 4, Public Administration and Information Technology, Springer New York, 2014, p. 189,
URL: http://dx.doi.org/10.1007/978-1-4614-9563-5_12 (cit. on p. 35).

[82] F. Maali, R. Cyganiak and V. Peristeras,
Enabling Interoperability of Government Data Catalogues, EGOV, ed. by M. Wimmer et al., Lecture Notes in Computer Science, Springer, 2010, p. 339,
URL: http://doi.org/10.1007/978-3-642-14799-9_29
(cit. on pp. 25, 26, 37–39, 45, 47, 102).

[83] J. Manyika et al., *Open Data: Unlocking Innovation and Performance with Liquid Information*, tech. rep. October, McKinsey, 2013, p. 24,
URL: http://www.mckinsey.com/business-functions/business-technology/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information (cit. on pp. 3, 50, 138).

[84] G. Marchionini, *Exploratory search*, Communications of the ACM vol. **49**.4 (Apr. 2006), p. 41,
URL: http://doi.org/10.1145/1121949.1121979 (cit. on p. 70).

[85] N. Marie and F. Gandon, *Survey of linked data based exploration systems*,
IESD 2014 - Intelligent Exploitation of Semantic Data, Oct. 2014,
URL: https://hal.inria.fr/hal-01057035/en (cit. on pp. 25, 33–35, 39, 46, 72).

[86] F. Marienfeld et al., *Metadata Aggregation at GovData.De: An Experience Report*,
Proceedings of the 9th International Symposium on Open Collaboration, WikiSym '13,
New York, NY, USA: ACM, 2013, 21:1,
URL: http://doi.acm.org/10.1145/2491055.2491077
(cit. on pp. 26, 29, 38, 39).

[87] S. Martin, M. Foulonneau and S. Turki,
1-5 Stars: Metadata on the Openness Level of Open Data Sets in Europe., MTSR, ed. by E. Garoufallou and J. Greenberg, vol. 390,
Communications in Computer and Information Science, Springer, 2013, p. 234,
URL: http://doi.org/10.1007/978-3-319-03437-9_24
(cit. on pp. 26, 29, 31, 37, 38, 46, 102).

[88] S. Martin et al., *Open Data: Barriers, Risks, and Opportunities*,
European Conference on eGovernment, Como, Italy, June 13-14 (2013), 2013,
URL: https://www.academia.edu/8334526/Open_Data_Barriers_Risks_and_Opportunities (cit. on p. 46).

[89] R. Matheus, M. M. Ribeiro and J. C. Vaz, *New Perspectives for Electronic Government in Brazil: The Adoption of Open Government Data in National and Subnational Governments of Brazil*, Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance, ICEGOV '12, New York, NY, USA: ACM, 2012, p. 22, URL: http://doi.acm.org/10.1145/2463728.2463734 (cit. on pp. 29, 31).

[90] R. Matheus et al., *Anti-corruption Online Monitoring Systems in Brazil*, Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance, ICEGOV '12, New York, NY, USA: ACM, 2012, p. 419, URL: http://doi.acm.org/10.1145/2463728.2463809 (cit. on pp. 25, 29, 30, 33, 39).

[91] R. Meijer, P. Conradie and S. Choenni, *Reconciling Contradictions of Open Data Regarding Transparency, Privacy, Security and Trust*, Journal of theoretical and applied electronic commerce research vol. **9** (2014), p. 32, URL: http://dx.doi.org/10.4067/S0718-18762014000300004 (cit. on pp. 40, 45).

[92] E. Mercado-Lara and J. R. Gil-Garcia, *Open Government and Data Intermediaries: The Case of AidData*, Proceedings of the 15th Annual International Conference on Digital Government Research, dg.o '14, New York, NY, USA: ACM, 2014, p. 335, URL: http://doi.acm.org/10.1145/2612733.2612789 (cit. on pp. 25, 34).

[93] H. G. Miller and P. Mork, *From Data to Decisions: A Value Chain for Big Data*, English, IT Professional vol. **15**.1 (Jan. 2013), p. 57, URL: http://doi.org/10.1109/MITP.2013.11 (cit. on pp. 96, 97).

[94] C. G. Mkude, C. Pérez-Espés and M. a. Wimmer, *Participatory budgeting: A framework to analyze the value-add of citizen participation*, Proceedings of the Annual Hawaii International Conference on System Sciences (2014), p. 2054, URL: http://doi.org/10.1109/HICSS.2014.260 (cit. on p. 52).

[95] D. B. Montgomery and G. L. Urban, *Marketing Decision-Information Systems: An Emerging View*, Journal of Marketing Research vol. **7**.2 (1970), arXiv: 9809069v1 [arXiv:gr-qc], URL: http://doi.org/10.2307/3150113 (cit. on p. 94).

[96] L. Morgan, J. Feller and P. Finnegan, *Exploring value networks: Theorising the creation and capture of value with open source software*, European Journal of Information Systems (2013), p. 569, URL: http://doi.org/10.1057/ejis.2012.44 (cit. on pp. 95, 97).

[97] S. Mouzakitis et al., *Challenges and opportunities in renovating public sector information by enabling linked data and analytics*, Information Systems Frontiers (2016), p. 1, ISSN: 1572-9419, URL: http://dx.doi.org/10.1007/s10796-016-9687-1 (cit. on p. 8).

[98] L. N. Mutuku and J. Colaco, *Increasing Kenyan Open Data Consumption: A Design Thinking Approach*, Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance, ICEGOV '12, New York, NY, USA: ACM, 2012, p. 18, URL: http://doi.acm.org/10.1145/2463728.2463733 (cit. on pp. 25, 33, 34, 36, 39).

[99] J. Nielsen and T. Landauer, *A mathematical model of the finding of usability problems*, Proceedings of the INTERACT'93 and CHI'93 ... (1993), p. 206, URL: http://doi.org/10.1145/169059.169166 (cit. on p. 83).

[100] X. Ochoa and E. Duval, *Quality Metrics for Learning Object Metadata*, Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2006, ed. by E. Pearson and P. Bohman, Chesapeake, VA: AACE, June 2006, p. 1004, URL: http://www.editlib.org/p/23127 (cit. on pp. 46, 47, 101, 102).

[101] K. O'Hara, *Enhancing the Quality of Open Data*, The Philosophy of Information Quality, ed. by L. Floridi and P. Illari, vol. 358, Synthese Library, Springer International Publishing, 2014, p. 201, URL: http://dx.doi.org/10.1007/978-3-319-07121-3_11 (cit. on pp. 34, 38, 46, 101).

[102] OpenSpending, *Budget Data Package*, tech. rep., Open Knowledge Foundation, 2014, URL: https://github.com/openspending/budget-data-package (cit. on pp. 50, 59).

[103] M. Osborne et al., *Real-time detection, tracking, and monitoring of automatically discovered events in social media*, June 2014, URL: http://eprints.gla.ac.uk/95243/ (cit. on p. 112).

[104] M. Palmirani, M. Martoni and D. Girardi, *Open Government Data Beyond Transparency*, Electronic Government and the Information Systems Perspective, ed. by A. Kö and E. Francesconi, vol. 8650, Lecture Notes in Computer Science, Springer International Publishing, 2014, p. 275, URL: http://dx.doi.org/10.1007/978-3-319-10178-1_22 (cit. on pp. 29, 31).

[105] P. Parycek, J. Hochtl and M. Ginner, *Open Government Data Implementation Evaluation*, Journal of theoretical and applied electronic commerce research vol. **9** (2014), p. 80, URL: http://dx.doi.org/10.4067/S0718-18762014000200007 (cit. on pp. 25, 29, 31, 36).

[106] T. Peixoto, *Beyond Theory: e-Participatory Budgeting and its Promises for eParticipation*, European Journal of ePractice vol. **1**.March (2009), p. 1, URL: http://www.uquebec.ca/observgo/fichiers/91130_eparticipation.pdf (cit. on p. 52).

[107] J. Peppard and A. Rylander, *From Value Chain to Value Network:* European Management Journal vol. **24**.2-3 (Apr. 2006), p. 128, URL: http://doi.org/10.1016/j.emj.2006.03.003 (cit. on pp. 93, 96, 97).

[108] M. Petychakis et al., *A State-of-the-Art Analysis of the Current Public Data Landscape from a Functional, Semantic and Technical Perspective*, Journal of theoretical and applied electronic commerce research vol. **9** (2014), p. 34, URL: http://dx.doi.org/10.4067/S0718-18762014000200004 (cit. on pp. 29, 31).

[109] L. L. Pipino, Y. W. Lee and R. Y. Wang, *Data Quality Assessment*, Commun. ACM vol. **45**.4 (Apr. 2002), p. 211, URL: http://doi.acm.org/10.1145/505248.506010 (cit. on pp. 21, 101).

[110] I. O. Popov et al., *Connecting the dots: a multi-pivot approach to data exploration*, Proceedings of the 10th International Conference on The Semantic Web - Volume Part I, Springer-Verlag, Oct. 2011, URL: http://dl.acm.org/citation.cfm?id=2063016.2063052 (cit. on p. 71).

[111] M. E. Porter, *Competitive Advantage:Creating and sustaining superior performance*, vol. 15, The Free Press, New York, 1985, URL: http://94.236.206.206/dohodi.net/books/en/Business%20Books/Michael%20Porter/Michael.Porter.-.Competitive.Advantage.pdf (cit. on pp. 4, 93, 95, 96, 107).

[112] M. E. Porter and V. E. Millar, *How information gives you competitive advantage*, Harvad Business Review vol. **63**.4 (1985), p. 149, URL: https://hbr.org/1985/07/how-information-gives-you-competitive-advantage/ar/1 (cit. on p. 95).

[113] C. Pradel, O. Haemmerlé and N. Hernandez, *Natural Language Query Translation into {SPARQL} using Patterns*, Proceedings of the Fourth International Workshop on Consuming Linked Data, {COLD} 2013, Sydney, Australia, 2013, URL: http://ceur-ws.org/Vol-1034/PradelEtAl_COLD2013.pdf (cit. on p. 71).

[114] L. M. Prieto, A. C. Rodríguez and J. Pimiento, *Implementation Framework for Open Data in Colombia*, Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance, ICEGOV '12, New York, NY, USA: ACM, 2012, p. 14, URL: http://doi.acm.org/10.1145/2463728.2463732 (cit. on pp. 29, 31).

[115] J. F. Rayport and J. J. Sviokla, *Exploiting the Virtual Value Chain*, Harvard Business Review vol. **73** (1995), p. 75, URL: https://hbr.org/1995/11/exploiting-the-virtual-value-chain (cit. on pp. 93, 96).

[116] K. J. Reiche and E. Höfig, *Implementation of Metadata Quality Metrics and Application on Public Government Data*, COMPSAC Workshops, 2013, p. 236, URL: http://doi.org/10.1109/COMPSACW.2013.32 (cit. on pp. 20, 21, 26, 38, 39, 46–48, 101, 102, 104).

[117] P. D. Renzio and J. Wehner, *The Impacts of Fiscal Openness: A Review of the Evidence*, tech. rep. March, GIFT - Global Initiative for Fiscal Transparency, 2015, p. 35, URL: https://ebape.fgv.br/sites/ebape.fgv.br/files/SSRN-id2602439.pdf (cit. on p. 51).

[118] L. A. Rodriguez Rojas, G. M. Tarazona Bermudez and J. M. Cueva Lovelle, *Open Data and Big Data: A Perspective from Colombia*, Knowledge Management in Organizations, vol. 185, Lecture Notes in Business Information Processing, Springer International Publishing, 2014, p. 35, URL: http://dx.doi.org/10.1007/978-3-319-08618-7_4 (cit. on pp. 14, 29, 31, 109).

[119] M. de Rosnay and K. Janssen, *Legal and Institutional Challenges for Opening Data across Public Sectors: Towards Common Policy Solutions*,
Journal of theoretical and applied electronic commerce research vol. **9** (2014), p. 1,
URL: http://dx.doi.org/10.4067/S0718-18762014000300002
(cit. on pp. 25, 36, 37, 40, 44).

[120] Royal Society, *G8 Science Ministers Statement*, tech. rep.,
Foreign & Commonwealth Office, 2013, URL: https://www.gov.uk/government/news/g8-science-ministers-statement
(cit. on p. 56).

[121] A. Russell and P. R. Smart, *NITELIGHT: A Graphical Editor for SPARQL Queries*,
Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC2008), Karlsruhe, Germany, {CEUR} Workshop Proceedings, 2008,
URL: http://ceur-ws.org/Vol-401/iswc2008pd_submission_11.pdf
(cit. on p. 71).

[122] S. D. J. B. Samur F. C. Araújo, Daniel Schwabe, *Experimenting with Explorator: a Direct Manipulation Generic RDF Browser and Querying Tool*, 2008, URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.164.6772
(cit. on p. 70).

[123] P. Sanabria, C. Pliscoff and R. Gomes,
E-Government Practices in South American Countries: Echoing a Global Trend or Really Improving Governance? The Experiences of Colombia, Chile, and Brazil, Open Government, ed. by M. Gascó-Hernóndez, vol. 4, Public Administration and Information Technology, Springer New York, 2014, p. 17,
URL: http://dx.doi.org/10.1007/978-1-4614-9563-5_2 (cit. on pp. 29, 31).

[124] R. Sandoval-Almazan and J. Gil-Garcia,
Towards an Evaluation Model for Open Government: A Preliminary Proposal,
Electronic Government, ed. by M. Janssen et al., vol. 8653, Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2014, p. 47,
URL: http://dx.doi.org/10.1007/978-3-662-44426-9_4 (cit. on pp. 28, 115).

[125] R. Sandoval-Almazan et al.,
Open Government 2.0: Citizen Empowerment Through Open Data, Web and Mobile Apps,
Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance, ICEGOV '12, New York, NY, USA: ACM, 2012, p. 30,
URL: http://doi.acm.org/10.1145/2463728.2463735 (cit. on pp. 29, 30).

[126] D. S. Sayogo, T. A. Pardo and M. Cook,
A Framework for Benchmarking Open Government Data Efforts,
System Sciences (HICSS), 2014 47th Hawaii International Conference on, Jan. 2014, p. 1896,
URL: http://doi.org/10.1109/HICSS.2014.240 (cit. on pp. 29, 31, 39).

[127] F. Scharffe et al., *Enabling linked data publication with the Datalift platform*,
Proc. AAAI workshop on semantic cities, Toronto, Canada, July 2012,
URL: https://hal.inria.fr/hal-00768424 (cit. on p. 72).

[128] D. Schlagwein and D. Schoder, *The Management of Open Value Creation*, 44th Hawaii International International Conference on Systems Science {(HICSS-44} 2011), Proceedings, 4-7 January 2011, Koloa, Kauai, HI, {USA}, {IEEE} Computer Society, 2011, p. 1, URL: http://dx.doi.org/10.1109/HICSS.2011.424 (cit. on p. 95).

[129] O. Seneviratne, *QueryMed: An Intuitive SPARQL Query Builder for Biomedical RDF Data ABSTRACT*, 2010, URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.389.4282 (cit. on p. 71).

[130] N. Shadbolt and K. O'Hara, *Linked Data in Government*, {IEEE} Internet Computing vol. **17**.4 (2013), p. 72, URL: http://doi.ieeecomputersociety.org/10.1109/MIC.2013.72 (cit. on p. 109).

[131] N. Shadbolt et al., *eGovernment*, John Domingue, Dieter Fensel & James Hendler (eds.), Handbook of Semantic Web Technologies, Springer-Verlag, 2011, p. 840, URL: http://eprints.soton.ac.uk/271711/ (cit. on pp. 37, 39).

[132] N. Shadbolt et al., *Linked open government data: lessons from Data.gov.uk*, May 2012, URL: http://eprints.soton.ac.uk/340564/4/Linked%20OGD.pdf (cit. on pp. 109, 110).

[133] S. Shah, A. Horne and J. Capella, *Good Data Won't Guarantee Good Decisions - Harvard Business Review*, Harvard Business Review vol. **90**.April (2012), URL: https://hbr.org/2012/04/good-data-wont-guarantee-good-decisions (cit. on p. 96).

[134] A. Sheffer Correa et al., *Really Opened Government Data: A Collaborative Transparency at Sight*, Big Data (BigData Congress), 2014 IEEE International Congress on, June 2014, p. 806, URL: http://doi.org/10.1109/BigData.Congress.2014.131 (cit. on p. 39).

[135] Y. Sintomer and C. Herzberg, *Participatory Budgeting in Europe: Potentials and Challenges*, International Journal of Urban and Regional Research vol. **32**.1 (2008), p. 164, URL: http://doi.org/10.1111/j.1468-2427.2008.00777.x (cit. on p. 52).

[136] M. Solar, G. Concha and L. Meijueiro, *A Model to Assess Open Government Data in Public Agencies.*, EGOV, ed. by H. J. Scholl et al., vol. 7443, Lecture Notes in Computer Science, Springer, 2012, p. 210, URL: http://doi.org/10.1007/978-3-642-33489-4_18 (cit. on pp. 25, 39, 42, 44, 47, 102).

[137] M. Solar, L. Meijueiro and F. Daniels, *A Guide to Implement Open Data in Public Agencies.*, EGOV, ed. by M. Wimmer, M. Janssen and H. J. Scholl, vol. 8074, Lecture Notes in Computer Science, Springer, 2013, p. 75, URL: http://doi.org/10.1007/978-3-642-40358-3_7 (cit. on pp. 34, 37).

[138] A. Stolz, B. Rodriguez-Castro and M. Hepp, *RDF Translator: A RESTful Multi-Format Data Converter for the Semantic Web*, tech. rep., E-Business and Web Science Research Group, Universität der Bundeswehr München, 2013, URL: http://arxiv.org/abs/1312.4704 (cit. on p. 72).

[139] E. Styrin, N. Dmitrieva and A. Zhulin, *Openness Evaluation Framework for Public Agencies*, Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance, ICEGOV '13, New York, NY, USA: ACM, 2013, p. 370, URL: http://doi.acm.org/10.1145/2591888.2591964 (cit. on p. 25).

[140] I. Susha et al., *Benchmarks for Evaluating the Progress of Open Data Adoption: Usage, Limitations, and Lessons Learned*, Social Science Computer Review (2014), p. 1, URL: http://doi.org/10.1177/0894439314560852 (cit. on pp. 28, 115).

[141] M. Svihla and I. Jelinek, *Benchmarking RDF Production Tools*, Database and Expert Systems Applications: 18th International Conference, DEXA 2007, Regensburg, Germany, September 3-7, 2007. Proceedings, ed. by R. Wagner, N. Revell and G. Pernul, Springer Berlin Heidelberg, 2007, p. 700, URL: http://dx.doi.org/10.1007/978-3-540-74469-6_68 (cit. on p. 72).

[142] M. Tvarozek and M. Bieliková, *Generating Exploratory Search Interfaces for the Semantic Web*, Human-Computer Interaction - Second {IFIP} {TC} 13 Symposium, {HCIS} 2010, Springer, 2010, URL: http://dx.doi.org/10.1007/978-3-642-15231-3_18 (cit. on p. 70).

[143] A. F. Tygel et al., *"How Much?" is not Enough: an Analysis of Open Budget Initiatives*, In Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance, ICEGOV 2016, Montevideo, Uruguay, March 1-3, 2016, 2016, p. 276, URL: http://doi.acm.org/10.1145/2910019.2910054 (cit. on p. 8).

[144] B. Ubaldi, *Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives*, OECD Working Papers on Public Governance No. 22 (OECD 2013) (May 2013), URL: http://dx.doi.org/10.1787/5k46bj4f03s7-en (cit. on pp. 51, 106).

[145] M. Vafopoulos et al., *Insights in global public spending*, Proceedings of the 9th International Conference on Semantic Systems - I-SEMANTICS '13, 2013, p. 135, URL: http://doi.org/10.1145/2506182.2506201 (cit. on p. 50).

[146] M. Vasa and S. Tamilselvam, *Building Apps with Open Data in India: An Experience*, Proceedings of the 1st International Workshop on Inclusive Web Programming - Programming on the Web with Open Data for Societal Applications, IWP 2014, New York, NY, USA: ACM, 2014, p. 1, URL: http://doi.acm.org/10.1145/2593761.2593763 (cit. on pp. 29, 31).

[147] N. Veljković, S. Bogdanović-Dinić and L. Stoimenov, *Web 2.0 as a Technological Driver of Democratic, Transparent, and Participatory Government*, Web 2.0 Technologies and Democratic Governance, ed. by C. G. Reddick and S. K. Aikins, vol. 1, Public Administration and Information Technology, Springer New York, 2012, p. 137, URL: http://dx.doi.org/10.1007/978-1-4614-1448-3_9 (cit. on p. 39).

[148] N. Veljković, S. Bogdanović-Dinić and L. Stoimenov, *Benchmarking open government: An open data perspective*, Government Information Quarterly vol. **31**.2 (2014), p. 278, URL: http://doi.org/10.1016/j.giq.2013.10.011 (cit. on pp. 28, 29, 51).

[149] N. Verma and M. P. Gupta,
Open Government Data: Beyond Policy & Portal, a Study in Indian Context, Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance, ICEGOV '13, New York, NY, USA: ACM, 2013, p. 338,
URL: http://doi.acm.org/10.1145/2591888.2591949 (cit. on pp. 36, 37).

[150] B. Villazón-Terrazas et al., *Methodological Guidelines for Publishing Government Linked Data*, Linking Government Data, ed. by D. Wood, Springer, 2011, chap. 2,
URL: http://doi.org/10.1007/978-1-4614-1767-5_2 (cit. on p. 109).

[151] G. Viscusi, M. Castelli and C. Batini,
Assessing Social Value in Open Data Initiatives: A Framework,
Future Internet vol. **6**.3 (2014), p. 498, URL: http://doi.org/10.3390/fi6030498 (cit. on p. 139).

[152] V. Vlasov and O. Parkhimovich, *Development of the Open Budget Format*,
Proceedings of the 16th conference of fruct association association, Oulu, 2014, p. 129,
URL: http://doi.org/10.1109/FRUCT.2014.7000922 (cit. on p. 50).

[153] S. van der Waal et al., *Lifting Open Data Portals to the Data Web*,
Linked Open Data – Creating Knowledge Out of Interlinked Data,
ed. by S. Auer, V. Bryl and S. Tramp, Lecture Notes in Computer Science,
Springer International Publishing, 2014, p. 175,
URL: http://dx.doi.org/10.1007/978-3-319-09846-3_9 (cit. on pp. 29, 30).

[154] S. T. Walker,
Budget mapping: Increasing citizen understanding of government via interactive design,
Proceedings of the Annual Hawaii International Conference on System Sciences, 2010, p. 1,
URL: http://doi.org/10.1109/HICSS.2010.87 (cit. on p. 57).

[155] A. Wassenaar, *E-Governmental Value Chain Models*, 11th International Workshop on Database and Expert Systems Applications (DEXA'00), 6-8 September 2000, Greenwich, London, {UK}, {IEEE} Computer Society, 2000, p. 289,
URL: http://dx.doi.org/10.1109/DEXA.2000.875041 (cit. on p. 95).

[156] T.-M. Yang et al., *Open Data Development and Value-added Government Information: Case Studies of Taiwan e-Government*, Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance, ICEGOV '13, New York, NY, USA: ACM, 2013, p. 238,
URL: http://doi.acm.org/10.1145/2591888.2591932 (cit. on pp. 29, 31, 33).

[157] Z. Yang and A. Kankanhalli, *Innovation in Government Services: The Case of Open Data*, Grand Successes and Failures in {IT.} Public and Private Sectors - {IFIP} {WG} 8.6 International Working Conference on Transfer and Diffusion of IT, {TDIT} 2013, Bangalore, India, June 27-29, 2013. Proceedings, 2013, p. 644,
URL: http://dx.doi.org/10.1007/978-3-642-38862-0_47
(cit. on pp. 33–36, 39).

[158] A. Zuiderwijk and M. Janssen,
A Coordination Theory Perspective to Improve the Use of Open Data in Policy-Making., EGOV, ed. by M. Wimmer, M. Janssen and H. J. Scholl, vol. 8074, Lecture Notes in Computer Science, Springer, 2013, p. 38, URL: http://doi.org/10.1007/978-3-642-40358-3_4
(cit. on p. 37).

[159] A. Zuiderwijk and M. Janssen, *Barriers and Development Directions for the Publication and Usage of Open Data: A Socio-Technical View*, Open Government, vol. 4, Public Administration and Information Technology, Springer New York, 2014, p. 115, URL: http://dx.doi.org/10.1007/978-1-4614-9563-5_8 (cit. on pp. 4, 14, 36, 38, 40, 106, 121).

[160] A. Zuiderwijk and M. Janssen, *Open data policies, their implementation and impact: A framework for comparison*, Government Information Quarterly vol. **31**.1 (2014), p. 17, URL: http://dx.doi.org/10.1016/j.giq.2013.04.003 (cit. on pp. 51, 58).

[161] A. Zuiderwijk and M. Janssen, *The Negative Effects of Open Government Data - Investigating the Dark Side of Open Data*, Proceedings of the 15th Annual International Conference on Digital Government Research, dg.o '14, New York, NY, USA: ACM, 2014, p. 147, URL: http://doi.acm.org/10.1145/2612733.2612761 (cit. on pp. 36–38, 40, 50, 138).

[162] A. Zuiderwijk, M. Janssen and A. Parnia, *The Complementarity of Open Data Infrastructures: An Analysis of Functionalities*, Proceedings of the 14th Annual International Conference on Digital Government Research, dg.o '13, Quebec, Canada: ACM, 2013, p. 166, URL: http://doi.acm.org/10.1145/2479724.2479749 (cit. on pp. 29, 30).

[163] A. Zuiderwijk et al., *Socio-technical impediments of open data*, Electronic Journal of eGovernment vol. **10**.2 (2012), p. 156, URL: http://www.ejeg.com/issue/download.html?idArticle=255 (cit. on p. 21).

[164] A. Zuiderwijk et al., *Special Issue on Innovation through Open Data: Guest Editors' Introduction*, Journal of theoretical and applied electronic commerce research vol. **9** (2014), p. i, URL: http://dx.doi.org/10.4067/S0718-18762014000200001 (cit. on p. 39).

[165] A. Zuiderwijk et al., *Special Issue on Transparency and Open Data Policies: Guest Editors' Introduction*, J. Theor. Appl. Electron. Commer. Res. vol. **9**.3 (Sept. 2014), p. i, URL: http://doi.org/10.4067/S0718-18762014000300001 (cit. on pp. 13, 22, 40).

APPENDIX A

Usability Evaluation Survey

*Required

Generic Information on you and your work affiliation

1. Please enter the time when you started the survey*

   _____

2. Are you Male or Female *
   - ☐ Male
   - ☐ Female
3. What is your age? *
   - ☐ 18-24
   - ☐ 25-34
   - ☐ 35-44
   - ☐ 45-54
   - ☐ 55-64
   - ☐ 65 or older
4. What is the purpose of your SME/Company or group in an Academic Entity? *

   _____

   _____

5. What is your position within the SME/Company/Academic Entity?*
   (Please mark only one)
   - ☐ Student
   - ☐ Researcher
   - ☐ Software Developer

- □ Engineer
- □ Owner/CEO
- □ (Project/Product) Manager
- □ Academic
- □ Other: _____

6. Do you or your SME/Company/Academic Entity use Linked Data?*
   (Please mark only one)
   - □ I use Linked Data
   - □ SME/Company/Academic Entity uses Linked Data
   - □ Both of the above
   - □ None of the above

If you or your SME/Company/Academic Entity use Linked Data please answer the following questions, if not please skip to the following section.

1. What specific tasks does your position require with regard to using Linked Data?
   (Please write tasks as short points in bullet form, for e.g. analysing data or accessing data)

   _____
   _____
   _____
   _____

2. What is the current process within your SME/Company/Academic Entity for accessing such data?
   (Please write tasks as short points in bullet form, for e.g. identifying published dataset, exploring data, cleaning data)

   _____
   _____
   _____
   _____

3. After accessing Linked Data, do you re-use it in some way?
   For example, to lead out analytics tasks
   (Please mark only one)
   - □ Yes
   - □ No

4. If yes, what tasks do you execute on this data?
   (Please mark all that apply)
   - □ Analytics

- ☐ Data Curating (e.g. cleaning, removing duplicates)
- ☐ Mashups
- ☐ Integration with enterprise data
- ☐ Visualisation
- ☐ Other _____

5. What are challenges faced in the process of re-using Linked Data?
   (Please mark all that apply)
   - ☐ Data not accessible
   - ☐ Incomplete data
   - ☐ Copyright/Licence issues
   - ☐ Data format problems
   - ☐ Invalid data
   - ☐ Data duplicates
   - ☐ Data not relevant
   - ☐ Other _____
6. How does the use of such data impact the SME/Company/Academic Entity?
   (Please mark all that apply)
   - ☐ More competitive
   - ☐ Better tackling of target market
   - ☐ Better service provision
   - ☐ New service provision
   - ☐ Other _____

Query Builder

This page regards questions about the Query Builder and its use.

1. Do you know how to query Linked Data using the SPARQL query language?*
   (Please mark only one)
   - ☐ Yes
   - ☐ No
2. If you use Linked Data, what is the current process to query Linked Data?
   (Please write tasks as short points in bullet form, for e.g. exploring dataset, testing with sample queries)

   _____
   _____
   _____
   _____

Please execute the following tasks:

(i) Open Query Builder: http://butterbur22.iai.uni-bonn.de:3000/query/builder
(ii) Select a dataset to Query
(iii) Select Class for required data (e.g. Actor, Animal, Politician, Film)
(iv) Select 3 Properties (e.g. Birth date, country, colour, age)
(v) Add a filter on one of the properties by clicking on the property name (e.g. country = Germany, colour = black, age > 25)
(vi) Preview the results
(vii) Convert the results and download in any format except PDF (this is because PDF is only added for the sake of completeness, as PDF is not re-usable as a format)

NOTE: If DBPedia is not responding please try using DBPedia Live, or vice versa. This is because the DBPedia service might not stable at the moment.

1. Do you agree that it was easy to execute this task?*
(Please mark only one)
   - □ Strongly disagree
   - □ Disagree
   - □ Neither agree nor disagree
   - □ Agree
   - □ Strongly Agree

2. Which parts did you find most challenging to do?*
(Please mark all that apply)
   - □ Select dataset to Query
   - □ Select Class for required data
   - □ Select 3 Properties
   - □ Add a filter on one of the properties by clicking on the property name
   - □ Preview the results
   - □ Convert the results and download in any format
   - □ None
   - □ Other _____

3. Do you feel using this tool is better than your normal way of executing such a data accessing task?
(Please mark only one)
   - □ Yes
   - □ No
   - □ Not sure

4. If no, how is your method better?

_____

_____

5. Do you agree that this tool would be useful in your SME/Company/Academic Entity to access, explore and query Linked Data? *
(Please mark only one)
   - ☐ Strongly disagree
   - ☐ Disagree
   - ☐ Neither agree nor disagree
   - ☐ Agree
   - ☐ Strongly Agree
6. Can you think of any other features or functions that would be useful for you in this tool?

Transformation Explorer

This part regards questions about the Transformation Explorer and its use.

Please execute the following task:
From the transformation explorer (http://butterbur22.iai.uni-bonn.de/pam/), find the query you just executed in the query builder. You can use the facets on the side to help you by ordering the results accordingly. Load the query into the Query Builder and edit the query so that you add another property.

1. Do you agree that it was easy to execute this task?*
(Please mark only one)
   - ☐ Strongly disagree
   - ☐ Disagree
   - ☐ Neither agree nor disagree
   - ☐ Agree
   - ☐ Strongly Agree
2. Do you agree that this tool would be useful in your SME/Company/Academic Entity to re-use saved data access queries?*
(Please mark only one)
   - ☐ Strongly disagree
   - ☐ Disagree
   - ☐ Neither agree nor disagree
   - ☐ Agree
   - ☐ Strongly Agree

3. Can you think of any other features or functions that would be useful for you in this tool?*

   _____

   _____

4. Please enter the time you finished the survey*

   _____

APPENDIX B

Effort Evaluation Survey

*Required

The aim of this survey is to assess how the Query Builder tool fares in comparison to getting data from DBpedia in other ways. We are mostly interested in the time and effort required to obtain the needed data.

In order to better understand how to operate the tools, please watch the short demo video:
https://vimeo.com/164145033

Link to the Query Builder Tool:
http://butterbur22.iai.uni-bonn.de:3000/query/builder

Link to DBpedia:
http://wiki.dbpedia.org/

Task using any method

1. Using your own method, how much time do you spend to obtain the data you need for the below task from DBpedia, and convert it to RDF?*
   Task: Get all Actors whose nationality is a country where the national language is English.
   (e.g. 4.03.32 - 4 hours, 3 minutes, 32 seconds)

   _____

2. What method did you use to get this data?*
   (e.g. using the DBpedia SPARQL endpoint)

   _____

3. If you don't know how to get the data in the given Task from DBpedia, do you have an alternative solution?

   _____

4. How easy is it to get the data you need?*
   (Please mark only one)

5. Do you use SPARQL queries frequently?*
   (Please mark only one)
   - ☐ Yes
   - ☐ No
   - ☐ I do not know SPARQL

Task using Query Builder

1. Using the Query Builder, how much time do you spend to obtain the data you need for the below task from DBpedia, and convert it to RDF?*
   Task: Get all Actors whose nationality is a country where the national language is English.
   (e.g. 4.03.32 - 4 hours, 3 minutes, 32 seconds)

   _____

2. How easy is it to get the data you need using the Query Builder?*
   (Please mark only one)

3. Do you think the Query Builder is useful to learn SPARQL?*
   (Please mark only one)

APPENDIX C

DSAAS Preliminary Usability Evaluation Survey

*Required

A survey on the benefits of using a tool which aims to balance the demand and supply of data

The Demand and Supply as a Service (DSAAS) is a knowledge base of existing datasets and their use cases. Data publishers can fill in the relevant (extensive) information on the datasets they want to publish, as well as any success stories as use cases. Thereafter consumers can browse different datasets through a faceted browser, and easily discover any dataset of interest. If consumers do not find what they need, then they can fill in a request for a datasets. Data publishers can then easily find niches in the data market, and provide for them respectively.

The aim of the DSAAS is to provide a match-making service that balances the demand and supply in the data market. This will aid publishers, consumers, as well as other stakeholders, though enabling and enhancing data discovery and re-use, collaborations, and providing contributions to the data market.

This survey hence requires stakeholders to provide their input on how this service would affect their participation in the data market, keeping in mind that the current knowledge base as yet only contains a sample of datasets. Eventually, the knowledge base will be expanded as more and more publishers and consumers collaborate in using this service.

Please visit the DSAAS (http://butterbur22.iai.uni-bonn.de/dsaas/) and try it out for yourself!

1. What is the role of your company in the data market?*
   (Please mark all that apply)
   - ☐ Data publishers (you create datasets and make them available publicly)
   - ☐ Data consumers (you re-use existing datasets)
   - ☐ Other: _____

2. What are the challenges you face when searching for and consuming datasets?*
   (Please mark all that apply)
   - ☐ Dataset not easily discoverable
   - ☐ Dataset re-use not specified (no knowledge of existing use cases)

- □ Lack of provenance information
- □ No idea if a dataset already exists
- □ Conditions of use unspecified/not clear (licence of use)
- □ Other: _____

Please read the following statements and answer whether you agree with the statement or otherwise.

3. The DSAAS can help stakeholders to easily identify the DEMAND in the data market (by enabling stakeholders to submit data requests).*
(Please mark only one)
   - □ Strongly disagree
   - □ Disagree
   - □ Neither agree nor disagree
   - □ Agree
   - □ Strongly Agree

4. The DSAAS can help stakeholders to easily identify the SUPPLY in the data market (by listing existing datasets).*
(Please mark only one)
   - □ Strongly disagree
   - □ Disagree
   - □ Neither agree nor disagree
   - □ Agree
   - □ Strongly Agree

5. The DSAAS can help stakeholders to identify a niche in the data market, and hence target it specifically (through catering for a data request).*
(Please mark only one)
   - □ Strongly disagree
   - □ Disagree
   - □ Neither agree nor disagree
   - □ Agree
   - □ Strongly Agree

6. Stakeholders are encouraged to re-use datasets if success stories (use cases) are provided.*
(Please mark only one)
   - □ Strongly disagree
   - □ Disagree
   - □ Neither agree nor disagree
   - □ Agree

- ☐ Strongly Agree

7. The DSAAS encourages stakeholders to collaborate with each other by showing their interests in specific dataset domains.*
(Please mark only one)
   - ☐ Strongly disagree
   - ☐ Disagree
   - ☐ Neither agree nor disagree
   - ☐ Agree
   - ☐ Strongly Agree
8. The DSAAS would be a good tool to showcase datasets and encourage their consumption.*
(Please mark only one)
   - ☐ Strongly disagree
   - ☐ Disagree
   - ☐ Neither agree nor disagree
   - ☐ Agree
   - ☐ Strongly Agree
9. By allowing consumers to put a request for a dataset, the DSAAS could possibly make the acquirement process faster.*
(Please mark only one)
   - ☐ Strongly disagree
   - ☐ Disagree
   - ☐ Neither agree nor disagree
   - ☐ Agree
   - ☐ Strongly Agree
10. Do you have any suggestions to improve the DSAAS?