Charles University in Prague
Faculty of Arts
Institute of Information Studies and Librarianship
New Media Studies

Diploma thesis

Jindřich Mynarz

Linked Open Data for Public Sector Information

Linked open data pro informace veřejného sektoru

Prague, 2012 Thesis supervisor: Doc. Ing. Vojtěch Svátek, Dr.

Affidavit

I hereby declare that I have written this diploma thesis on my own, that I have cited all of the resources used, and that the thesis was not used to obtain another university degree.

May 4, 2012 in Prague

. . . . . .
student's signature

Suggested citation

MYNARZ, Jindřich. *Linked open data for public sector information.* Prague, 2012. 92 p. Diploma thesis. Charles University in Prague, Faculty of Arts, Institute of Information Studies and Librarianship. Thesis supervisor Vojtěch SVÁTEK.

Abstract

The diploma thesis introduces the domain of proactive disclosure of public sector information via linked open data. At the start, the legal framework encompassing public sector information is expounded along with the basic approaches for its disclosure. The practices of publishing data as open data are defined as an approach for proactive disclosure that is based on the application of the principle of openness to data with the goal to enable equal access and equal use of the data. The reviewed practices range from necessary legal actions, choices of appropriate technologies, and ways in which the technologies should be used to achieve the best data quality. Linked data is presented as a knowledge technology that, for the most part, fulfils the requirements on open technology suitable for open data. The thesis extrapolates further from the adoption of linked open data in the public sector to recognize the impact and challenges proceeding from this change. The distinctive focus on the side supplying data and the trust in the transformative effects of technological changes are identified among the key sources of these challenges. The emphasis on technologies for data disclosure at the expense of a more careful attention to the use of data is presented as a possible source of risks that may undermine the overall merits of linked open data.

Key words

linked data, open data, public sector information

Abstrakt

Diplomová práce představuje problematiku zveřejňování informací veřejného sektoru pomocí principů otevřených a propojených dat. Nejprve přibližuje právní rámec, v němž se informace veřejného sektoru nacházejí, a základní přístupy ke zveřejňování těchto informací. Jsou popsány praktiky, kterými jsou definována otevřená data. Tyto praktiky vycházejí z aplikace principu otevřenosti na data a mají pro data zaručit rovný přístup a užití. Zahrnují nezbytné právní úkony, volbu vhodných technologií a jejich správné užití pro dosažení vysoké kvality dat. Propojená data jsou představena jako znalostní technologie, která po většině stránek naplňuje požadavky na otevřenou technologii pro otevřená data. Na základě užití principů otevřených a propojených dat ve veřejné správě jsou domýšleny jejich dopady a výzvy, které z této aplikace vyplývají. Mezi ústředními příčinami výzev, které s sebou převzetí těchto praktik ve veřejné správě nese, je identifikováno zejména vyhraněné zaměření na stranu nabídky dat a důvěra v transformativní účinky technologických změn. Zdůraznění technologií pro zpřístupnění otevřených dat na úkor ohledů ke způsobům užití dat je představeno jako možný zdroj rizik, která mohou výrazně oslabit pozitivní přínosy otevřených a propojených dat.

Klíčová slova

propojená data, otevřená data, informace veřejného sektoru

Contents

Preface

The choice of the topic for the thesis was informed by a number of related activities I have been involved in during the past several years. In the summer of 2010, I worked as a research intern at the Linked Data Research Centre[1] at the Digital Enterprise Research Institute in Ireland. The internship was focused on the conversion of legacy data sources from the Central Statistics Office of Ireland to linked data. I was one of the founding members of OpenData.cz[2], a Czech initiative promoting transparent data infrastructure for the public sector, which was established in late 2010. In December 2011, I co-founded the Czech chapter of the Open Knowledge Foundation[3] that is dedicated to building of a local community of supporters advancing the state of open knowledge in the Czech Republic. I have helped to organize a few events, the themes of which were closely related to the topic of the thesis, such as the Big Clean 2011[4] or Open Data and Public Sector,[5] which was held the Czech Chamber of Deputies in February 2012. Currently, I work for the University of Economics on the LOD2 research project,[6] that is dedicated to pushing the web of linked open data to the next level. In the light of my activity in the recent years, the topic of thesis has been the key research focus of mine, which resulted in it being a prime choice for my thesis assignment.

As compared with the structure in the thesis assignment several changes have been made to restructure the thesis. An introductory chapter about the domain of public sector information was added to the beginning of the thesis. A major addition consisted in supplementing the thesis with a chapter reviewing the impact and challenges for the application of linked open data for public sector information. On the contrary, the proposed use case demonstrating an example application was left out. The motive for this change was that the data exports obtained from the Czech Statistical Office, that were intended to serve as the source on which

[1] http://linkeddata.deri.ie/

[2] http://opendata.cz/

[3] http://cz.okfn.org/

[4] http://bigclean.org/praha/

[5] http://www.opendata.cz/en/event-austria12

[6] http://lod2.eu/

the application would be built, were impoverished and thus insufficient for the application to be developed.

I would like to thank to Vojtěch Svátek for supervising the evolution of this thesis and guiding it in the right direction.

1 Introduction

The public sector records data about what it does and about the environment in which it operates. Nowadays, improved and automated ways of data collection lead to a growth of the volume of data that is available in the public sector. Digitization allows to store the recorded data in a way that scales. Presently, researchers estimate that more than 5 exabytes is stored online every day [116]. Fortunately, there are scalable technologies for data storage and retrieval at our disposal.

The Web enables zero cost reproduction of digital information that makes it possible to share the information in a frictionless manner. Building on the premise that data deemed useful for the public sector is useful for the private sector as well, online exchange of public sector data allows to maximize its value by reaching members of the public that may recycle it and reuse it for their own purposes. In fact, the increased access and reuse of the disclosed public data is driven by technologies making it feasible [14].

Digital data may be represented in structured ways that make it machine-readable. Raw, machine-readable representations of data are amenable to automated processing and enable to retain the generative value of data, so that people and computers might use the data in a non-predefined way. Machine readability makes possible a wide array of interactions with data that go far beyond displaying it. In this way, disclosure of public sector data in a machine-readable format allows members of the public to find new uses for the data.

Adoption of the available technologies for data representation and storage may prove to have a disruptive effect on the public sector. Graham Vickery emphasizes two technological developments that, in his opinion, completely redefined the possibilities for public sector information [110, p. 6]. First, he points out to the technologies that enable digitization of public resources. Second, he highlights the role of broadband telecommunications that enable better access to public sector information.

The technologies for representing and exchanging data constitute the basic components for open disclosure of data. Open access to public sector data is considered as a key ingredient for a government that is open. Open government is *"the notion that the people have the right to access the documents and proceedings of government"* [65, p. xix], which is necessary for an open society that *"reflects the*

universal values of intellectual autonomy, equality and trust" [49, p. 8]. Coupled with the demand for openness of the public sector, the technologies stimulated numerous initiatives promoting open data world-wide. Open data is a set of practices for data disclosure that strives to provide for an equal access and an equal use of the data.

The foundations of open data draw from related approaches. Driven by the recognition of freedom of information as a basic human right, open data transposes the principles of open access, close to those of open source, onto data. It complements the adoption of the approaches of e-government, which promotes use of information and communication technologies to improve government processes, and coincides with the call for government 2.0, which makes a better use of online collaborative technologies to create a more participatory government.

The application of open data, and more specifically linked open data, to the information held by public sector bodies constitutes the main theme of this thesis. Public sector information represents the content, to which the principles of open data are applied using the technologies recommended by the linked data publication model. The goal od this thesis is twofold. Its first part is to explore the competitive advantage of linked data for release of public sector information under the terms of open data principles. Its second part is to extrapolate the impact and challenges associated with the adoption of linked open data for public sector information.

2 Public Sector Information

Access to proceedings of the public sector is a fundamental underpinning of democracy. *"Quality of public discussion would be significantly impoverished without the nourishment of information from public authorities"* [75]. Moreover, economic and research activities in the private sector would be vastly impoverished if public sector information was kept concealed within the public sector. Reuse of public sector information in the private sector is a pivotal goal of its disclosure.

The disclosure of public sector information constitutes the subject matter of this thesis. The aim of this chapter is to delineate the scope of the described domain by providing its basic conceptualization, along with lexical and extensional definitions of the concepts involved.

To cater for this goal, the introductory section is concerned with definitions. Once an elemental characterization of the domain in question is presented, the legal environment for public sector information is described. The practical side of the subject matter is covered in sections dealing with the models for disclosure of public sector information and with the ways how to set the price charged for information provision. The chapter concludes with a remark discussing reuse, which is the goal of the disclosure of public sector information.

2.1 Definitions

This section defines the key concepts for the purposes of the following parts of the thesis. It explains the notions of the public sector, public body, and public sector information.

First, how can the borders of the public sector be circumscribed? Boundaries of the public sector are demarcated by private ownership. The institutions the public sector consists of are not private property [68, p. 5]. Instead, the public sector is publicly owned.

Other definitions of the public sector employ the viewpoints of policy control or financial control. A common way of how to give a definition to the public sector in law is to use an extensional definition enumerating the public bodies that fall within its scope.

However, the boundary between public and private sector is getting blurry, since a lot of the functions traditionally performed by public bodies have been

outsourced within public-private partnerships. The public sector may also start to take on some characteristics of the private sector, such as the models of finance management.

The public sector is constituted of public bodies. Public body is an institution with legal subjectivity that belongs to the public sector. It is set up under law by the state or other public sector body. Public bodies are established for a specific purpose of meeting the needs in the general interest. They do not have a commercial character and so the majority of their budgets is funded from tax revenue [34, p. 55]. Among the public bodies that are deemed to be most important from the perspective of the data they produce are offices of cadaster, mapping agencies, statistical offices, or company registrars [110, p. 10].

Public bodies produce public sector information, or public data, which is the subject matter of this chapter. UK Public data transparency principles offer a working definition of "public data". Public data is thought of as *"the objective, factual, non-personal data on which public services run and are assessed, and on which policy decisions are based, or which is collected or generated in the course of public service delivery"* [106]. It is usually a by-product of the delivery of functions of public sector bodies, which makes it serve as an official public record as well [4]. The term "public sector data" is in most contexts used in the same way as "government data", and can be thus treated as synonymous.

Given the generic definition of public sector information, enumerating all of the types of public data would be unnecesary. Instead, a few prototypical examples will be mentioned. In 2010, a survey by Socrata identified several high-value categories of data. Among the top-ranked categories were data about public safety, revenues and expenditures, and education [100]. The most commonly used data categories in *publicdata.eu*,[1] a catalogue of Europe's public data, are "Finance and budgeting", "Social questions", and "Education and communication". Among the other frequently mentioned types of public data are statistical or geospatial data, the types that are particularly important from the perspective of their reuse by businesses. Paul Clarke sorted out public data into four categories [19]:

- *Historical data*, such as statistics
- *Planning data*, including legal regulations in progress
- *Infractructural data*, for example, reference concepts such as postcodes
- *Operational data*, covering real-time streaming data, e.g., traffic situation

Governments collect data for a plethora of topics, some of which may look obscure, such as the statistics of people injured by vending machines in the US [70]. Nevertheless, collection of all of the datasets should be justified by their function

[1] http://publicdata.eu/

for fulfiling the requirements of the public task and by their contribution as a source of improvements, such as for increasing the safety of vending machines in the aforementioned example.

2.2 Legal Regulations for Public Sector Information

Public sector information is a subject to jurisprudence based on different sources of law and regulations endowed with legal power. The law relevant for the disclosure of public sector information comes from multiple regulators and as such is a combination of both international law, including conventions or EU directives, and national law [63]. As a result of this state of affairs, the conditions governing public sector information may be composed of rules coming from multiple layers. In effect, the legislation related to public sector information may pose equivocal requirements and ordinances that are difficult to adhere to.

The right to access to public sector information stems from a basic human freedom to seek and impart information. Right to information is enshrined in at least 50 national constitutions [16, p. 62]. Dedicated acts formalizing the right to access to public sector information are established in a large part of countries that acknowledge the freedom of information.

First legal act on the access to public sector information entitled "Freedom of the Press Act" was passed in 1766 in Sweden [16, p. 57]. The right to know what proceedings of the public sector are was recognized as early as 1969 by the Japanese Supreme Court [75]. Other countries followed the suit by establishing the right to know and access to information as a part of the citizen rights. During the following decades the adoption was rather slow and in the middle of 1980s only 11 countries had freedom of information law [15, p. 264]. However, this area experienced a sudden growth of interest paired with an increasing number of countries recognizing the importance of access to information. By 2004 the number of countries that enacted a freedom of information law increased to 59 [15, p. 264].

The prevailing presumption in favour of secrecy has shifted to presumption favouring maximum disclosure and public sector information that is open by default [65, p. 23]. In many countries, the default settings for access to public sector information have changed. Accessing public sector information is no longer perceived as a privilege, it is a right [64, p. 8].

This thesis focuses on the legal situation for public sector information in the European Union and its member countries. The EU legislation is most relevant for the European context, in which the thesis is situated, and which can prove to be a valid model for an official public policy that establishes rules for the domain of public sector information.

2.2.1 Legislation in the European Union

In the EU, public sector information legislation consists of the directives of the European Commission and their local transpositions that weave the directives' regulations into state law of the member countries. A key directive covering public sector information is the EU directive on the re-use of public sector information [36]. The directive *"establishes a minimum set of rules governing the re-use and the practical means of facilitating re- use of existing documents held by public sector bodies"* [36, p. 93]. It prescribes public bodies to provide a mechanism for members of the public to request access to information produced by the bodies. The overarching tenet of the directive is non-discrimination, which manifests itself in stipulations including the prohibition of exclusive arrangements that grant exclusive rights to access to a particular entity, or the recommendation for marginal cost charging.

The planned novelization of the directive [37] extends the scope of public sector information to include the information from the cultural heritage sector, such as libraries, archives, and museums. Furthermore, it strives to conflate the right to access with the right to reuse. It brings about a change in the charging models that declares the marginal cost of reproduction as a new default, while requiring public bodies that continue charge extra price to provide a solid explanation for their behaviour. The novelization also deals with the enforcement of the directive and proposes to set up an independent authority to oversee the compliance with the principles of disclosure.

2.3 Disclosure of Public Sector Information

The regulations require public bodies to take on an obligation of providing access to information they possess. The EU directive on the re-use of public sector information holds the disclosure of public sector information to be a *"fundamental instrument for extending the right to knowledge, which is a basic principle of democracy"* [36, p. 92]. In the light of this assertion, public bodies should ensure wide dissemination and long-term preservation of the information they produce.

2.3.1 Scope of Disclosure

Public sector information is an umbrella term for all content produced by public bodies [93, p. 5]. Nonetheless, there are several exceptions to this rule, when defining the information that should be disclosed.

Public sector information covers any non-personal data held, collected or produced by a public body as a part of the public task, with the exception of the information relating to national security [118, p. 6]. Therefore, disclosure of public sector information should not apply to information that would abrogate individual privacy rights or endanger national security [42]. However, when left

unquestioned, the goal of national security may lead public sector bodies to be overprotective of some data. For example, for some time in the US data about dams were not available due to the fear of misuse for terrorist attacks [65, p. 330].

In the EU, several types of public sector information are exempted from the requirement of disclosure. Public sector information held by cultural heritage institutions, such libraries, museums, and archives, currently falls under a different regime. It often has different qualities than the information from other parts of the public sector. This type of information is mostly static, held as a record, and not directly associated with the pursue of public tasks [110, p. 7]. Similarly, the public broadcasting and research information generated by education institutions is usually exempt from the scope of the definition of public sector information. However, besides the exceptions listed individually, all public sector information is a subject to the requirement of disclosure.

2.3.2 Types of Disclosure

The approaches to disclosure of public sector information are usually categorized either according to the extent of disclosed information or by the activity of the public body.

The information that gets released might be limited a summary of the full information the public body possesses. *Summary disclosure* is used for informing about the decisions made by public bodies. On the other hand, *full disclosure* is used for informing the decisions of the public. For example, in the case of elections, decisions of the members of public are based on information from public bodies.

Based on the distinction of the source of initiative that drives the disclosure, there are two models of information provision in the public sector: *reactive* and *proactive* [41, p. 155].

2.3.2.1 Reactive Disclosure

Reactive disclosure is an on-demand, passive dissemination of public sector information that *"implies an (enforceable) right for a subject to access to information on request"* [101]. It institutes a permission culture of freedom of information requests. Joshua Tauberer criticizes reactive disclosure, because it provides only *"a very narrow view of the public sector that is based on the requested snap-shots of data"* [105]. This model is characterized by a strong information control and a lack of high-level political and bureaucratic support for open government and as such, reactive disclosure is unsuitable for the realization of this vision.

2.3.2.2 Proactive Disclosure

Proactive disclosure is an active dissemination of public sector information that *"means that the information is publicly available on the basis of a direct initiative of the public body"* [101]. This type of disclosure may also be referred to as "suo motu" disclosure, that comes from the Latin "upon its own initiative" [16, p. 69]. Proactive disclosure thus requires a switch from *"presumption of non-disclosure to presumption of openness"* [16, p. 66]. With such presumption, public sector information is thought of as public resource, as something to be shared. This way of disclosure is *"suited for mediators"* [105], that can transform the information and add value to it. An example of a model for proactive disclosure is open data, which is discussed in 3.

2.4 Pricing Models for Disclosure of Public Sector Information

The disclosure of information might be a subject to charge. However, conditioning access to public sector information by prices may constitute a fundamental barrier.

The models for pricing public sector information may be divided into three groups. The first model sees public bodies act as private companies and tries to recover their costs incurred from information production. If public bodies charge only to recover the cost of information provision, they use the marginal cost model. To adopt the third model is to cease charging altogether and not require users of information to pay any price.

2.4.1 Cost Recovery Model

Public sector institutions are usually free to recoup some costs by charging users that access their information [110, p. 11]. When they employ the cost recovery pricing, they essentially behave the same way as for-profit companies.

Aside from the benefit of public bodies being able to sustain themselves, this model introduces a number of challenges. First, it is discriminative for those that cannot afford to pay for the access to information of their interest. For example, full cost recovery may have an adverse effect on small and medium-sized enterprises that do not have the necessary resources to obtain the information they need in order to pursue their business plan. Second, a large part of consumers of public sector infomation is constituted by other public sector bodies. If full cost recovery is demanded from public bodies, it reduces public sector information to an instrument of reallocation of the public funding.

2.4.2 Marginal Cost Model

Marginal cost pricing recoups only the costs of information provision. It is derived from the marginal cost of distribution, that reflects the cost of provision of one further unit. This pricing model is recommended by the EU directive on the re-use of public sector information [36]. If public bodies adopt the marginal cost pricing model and start charging less for their information, they might see a surge of interest for the information that might lead to a greater total income than in the cases when the bodies employ full cost recovery model. The use of the Web brings this pricing model close to the model that applies no prices, because on the Web the marginal cost of distribution covering the reproduction of information is essentially zero.

2.4.3 Open Access Model

In the open access model public body does not require a payment for provisioning of information to the public. This approach entails a significant reduction of friction and administrative overhead associated with each individual transaction of public sector information. It is a non-discriminative model, since it makes access to information to be independent on user's budget.

A common argument for no pricing is that public sector information had been already paid for from the tax revenue and thus there should not be any additional charges [16, p. 55]. Pricing for information is seen as inconsistent with the established way of funding of public sector bodies. Public sector should not run business, and some contend that civil service is too inflexible to do so [5].

Several alternative models to recover partial costs were proposed to substitute for the direct cost recovery. For example, one model suggested imposing a levy on requests for updates of public data, such as in business registers [110, p. 27].

2.5 Reuse of Public Sector Information

A motivation behind the access to information that drives its disclosure is the recognition of the value of reuse. Reuse[2] occurs when information is used for other purpose than the original purpose for which it was created. According to the EU directive on the re-use of public sector information reuse is constituted by the uses for other reasons than to fulfil the public task, for which the information was originally collected [36, p. 90].

[2] In this text "reuse" is used for both noun and verb forms. It is recommended over "re-use" by *New Oxford American Dictionary*. Moreover, it reuses the same spelling adopted by Wikipedia (see http://en.wikipedia.org/wiki/Reuse). To achieve consistency, the form "reuse" was chosen to be used in the text with the exception of titles and quotes, in which the originally used form was preserved.

Access to public sector information opens a wide array of ways how to exploit it. It makes possible many uses that cannot be foreseen in the public sector. Public sector is a provider of unique data that is difficult to replicate, due to reasons such as prohibitively high expenses for data collection. This makes it even more important to allow for unfettered access and reuse of such information, the benefits of which will be discussed in the chapter 5.

2.6 Summary

This chapter established a basic terminology for describing the domain of disclosure of public sector information. It covered related legal frameworks for public sector information, which outline what is possible and legitimate to do with the information. From the practical perspective, the chapter took a look at the models for disclosure and pricing for the information. Finally, the potential for reuse of the information was touched on, preparing the ground to other parts of the thesis, in which the notion will be expanded on. The concepts introduced in the chapter will be build on in the proceeding chapters. The next chapter discusses open data, which can be considered as a proactive model for disclosure of public sector information.

3 Open Data

Open data is a set of practices for publishing data. There is no formally declared and adopted definition of open data and it is not backed by any legal or standards body. Instead, it is essentially a community-driven effort [57, p. 1]. The meaning of the concept of open data stems from a shared discourse in the open data community.

This chapter deals with the core concepts of open data and describes the principles of open data that are guided by these concepts. The following parts cover open data in practice, including general characterizations of policies and implementations of open data in the public sector.

3.1 Concepts

The concept of open data refers, as the name suggests, to two main abstract concepts. It denotes the application of the principles of *openness* to *data*.

3.1.1 Data

Data[1] is the subject of open data. Description of the elusive notion of data may be constructed by juxtaposing its various facets.

One perspective of defining data is through its content. According to the *Suggested Upper Merged Ontology*[2] "data point" or "datum" is *"an item of factual information derived from measurement or research."*[3] Data cover observations, measurements, and records describing the physical or social reality. It may also provide models and conceptualizations of reality for describing other data.[4]

A defining facet that contributes to the meaning of data is its form. First, data

[1] In this text, "data" is used as a singular mass noun. For a detailed discussion on the subject of the correct grammatical use of "data" see http://purl.org/nxg/note/singular-data.

[2] http://sigma.ontologyportal.org:4010/sigma/WordNet.jsp?synset=105816622

[3] Even though the definition refers to the singular form "datum", which has effectively disappeared from use, the definition applies to data as well.

[4] Sometimes referred to as "metadata."

is primarily digital[5], which makes it "computable"; i.e., amenable to automatic computer processing. Data imposes a structure on its content that makes it sufficiently formalized to allow for processing in an automated manner. The perception of this attribute depends on the level of use of data. For example, while a sound recording is usually not thought of as data if it is used to convey words, it may be treated as data if its use is its conversion to a different file format.

Content and format of data determine its affordances; the uses data makes possible. Data is generative, open to a variety of types of use. For example, data is used for preservation, information exchange, or computation.

Not only the types of use shape the prevalent perceptions of data. A common-sense interpretation of data is influenced by the tools we use to work with data; how we store it, represent it, or interface with it. The context in which data is used shapes the mental image of data. For instance, people recall database tables or spreadsheets when they think about data.

Data is characterized by features that make it conducive to be opened. Digital data is easy and inexpensive to copy. Thus, it may be treated as a non-rivalrous, public good [30]. Because data users are working with their own copy, consumption of data does not diminish the ability of others to consume it. In other words, using data does not make it less useful. In fact, quite the opposite is true as using data can make it more valuable. For example, one can extract valuable annotations informing about the ways data is being used, that are based on implicit participations [65, p. 32]. In the light of these properties, openness seems to be innate to data.

3.1.2 Openness

Openness is the intellectual foundation of open data. It is a quality of being open, an absence of restrictions and barriers, the goal of which is to achieve equal access and use.

The principle of openness is transdisciplinary. It is the driving force behind several movements, including open data. For example, Holger Kienle lists several related domains in which the concept of "openness" is applied [62]. They include areas such as "open access", "open content", "open knowledge", and "open source". What these related fields have in common is their concern with an open way of distribution, which is based on the free transfer of digital information on the Web, unencumbered by any common restrictions and barriers of the physical world. In this way, all of these fields, including open data, may be treated as publication models that apply the principles of openness to various domains.

[5] Analogue data carriers, such as cassettes, may also hold data. However, in order for them to be processable with a computer, they need to be digitized, which comprises of sampling through sensors, such as analog-to-digital converters.

Open access[6] focuses on literature, such as articles and pre-prints, that serve as research sources. Open source[7] applies the open publication model to software, with the particular aim to enable free access to software's source code. Open content[8] is a more general framework concerning content of any type of creative work. Content is deemed as open if it allows four types of use: reuse, revision, remix, and redistribution. In the same vein, Open knowledge[9] deals with any type of knowledge, notwithstanding its carrier, that is recognized as open only under the circumstances if anyone is free to use, reuse and redistribute it, requiring at most attribution or sharing it alike, under an analogous licence [84].

3.2 Principles

Principles of open data describe what it means for data to be "open" and what qualities make it open. The principles are the constituent parts that compose the meaning of open data. They are intended to signal the desired state of data. The principles enumerate the attributes of data that are required for it to be recognized as open. In this way, they serve as a benchmark for the open data community to distinguish between open data and data that is not open. The principles offer a pragmatic, non-normative definition of open data that recommends rather than prescribes. However, the principles are used not only to determine the openness of data. They are also used as a tool to communicate the meaning of the concept of open data.

In the discourse surrounding open data a number of principles defining open data were devised. Among the oldest and most referred to are the Eight principles of open government data [1]. An example of principles that take mostly the legal perspective is the Open definition [84] by the Open Knowledge Foundation. In the following part, key requirements for open data were identified based on the study of existing principles.

This examination focused on the requirements that refer to features of open data that are indispensable to its openness. However, it was difficult to separate the principles that cover either data openness or data quality. Some of the principles describe features that are not essential for data openness and that are actually features of a more general good design. Therefore, even though the following part concentrates on the core attributes of open data, some of the attributes closely related to data quality are covered in an independent section.

The following review is also aware that the importance of different attributes of data openness depends on the use case. Having this in mind, no priorities

[6] http://www.earlham.edu/~peters/fos/overview.htm

[7] http://www.opensource.org/osd.html

[8] http://opencontent.org/definition/

[9] http://opendefinition.org/okd/

are imposed on the listed principles. Instead, according to their relations, the principles are clustered into three groups that cover the aspects of legal openness, technical openness, and data quality respectively.

The sections that follow draw significantly from a blog post by the thesis' author [79].

3.2.1 Legal Openness

Legal openness addresses the conditions of use. In other words, it covers what users are allowed to do with the data.

The default conditions of use for open data are declared by law. The main areas of legislation that impact open data include intellectual property rights and database rights [109, p. 138].

The control of intellectual property rights over data depends on the content of data. These rights affect only original creative works. Data, in most cases, does not satisfy this condition. It usually consists of facts and, according to the law, no one can claim ownership of facts. Moreover, data is not a product of creative work [77].

In the case of public sector information, it is a product of the pursue of the public task. In such a case, public data may be explicitly declared to be exempt from copyright, which was proclaimed for the US public data in the 1976's Copyright Law. The baseline here is that, in many cases, data may not be treated as a private property, but more likely as a common good.

The distinction whether there are intellectual property rights associated with data is an important one. The options in this division introduce a completely different default state for data. The assessment of the relation of intellectual property rights is relevant for narrowing down the alternative ways how the rights holders may modify the conditions of use for data.

The impact of database rights on data is restricted by the law that is valid where the data is produced. Of course, local legislations influence intellectual property rights of data as well, however, they tend to be more universal as they are harmonized thanks to a number of international treaties. Sui generis database rights apply especially in the context of the member states of the European Union. In 1996, the EU issued the Directive 96/9/EC on the legal protection of databases [35]. The directive grants rights to the creators of databases, protecting their intellectual contribution to selection and arrangement of the database contents. This directive is now transposed into the legal systems in many EU member countries.

With regard to the described rights, in some cases, open data may be a subject to requirements of both. The content of data may be eligible for intellectual property rights protection and the data as a whole may be entitled to derive its protection from database rights. In such a situation, dual licensing may be applied, providing data content and data structure with different licences that are

more appropriate for the given type of licensed work. However, it may prove to be difficult to find a clear boundary distinguishing between the parts of data to be licensed separately. It also raises the barrier to use of the data, since its users need to know the requirements of both licences. Due to these complexities it may be easier to handle the legal variations with a universal waiver.

Possibilities for opening data may also be limited by implied contracts, such as exclusive licence agreements. Data bounded by contracts may be difficult to work with, because users may be either not aware of their existence or they may be found difficult to interpret and abide with, especially for laypersons. The most usable solution for open data would be to have a single legal document that users need to consult in order to know what the conditions of use are, as explicit and unified rules simplify the use of data.

The legal recommendations found in open data principles usually advise to modify the default conditions under which data is available with a legal instrument that amends the conditions on the basis of contract law, using tools such as a licence or a waiver. Such recommendation serves a number of purposes. First of all, it provides explicit and comprehensive conditions of use that are valid for the data in question, shielding the users from the possibly complex and hard-to-interpret law. The second aim is crucial for open data, because this is the way how a previously restricting conditions may be made more open by renouncing some rights.

There are two main types of legal tools used to amend the conditions of use of data: licences and waivers. Licences redefine how data may be used in accordance with the producer's desires and users' needs. Licences for open data are discussed in the subsequent section.

Waivers serve to waive rights associated with data. The purpose of legal waivers is to reconstruct the conditions of use that applies to the works in the public domain. Yet in some countries, such as the Czech Republic, waiving intellectual property rights is not considered as a valid legal act. In these countries, works may enter into the public domain only naturally and not with a deliberate action. However, licences may be used to emulate the public domain by explicitly setting the same conditions of use.

Both with law, regulations, licences, and waivers data producers are able to accomplish legal openness. Legal openness is a necessary precondition for achieving technical openness. Data that is technically open (e.g., online and in a structured format) but not legally open (e.g., with a prohibitive licence) is not open at all. Most of the data that is legally open can be made open in the technological respect, such as by screen-scraping, a technique that extracts data from web pages. In fact, increasing technical openness of data is an example of reuse that is made possible by open legal conditions of use. On the contrary, there are no ways in which users of data can achieve legal openness of the data, since only data producers can do that.

The following part covers the specific topic of licences, the most common legal tools for opening data.

3.2.1.1 Licences

By default, reuse requires permission. Unless there are legal instruments that enforce openness of data by default, there is a need for an explicit, open licence. Licence serves as a legal tool facilitating reuse [44, p. 6].

The licence should state clearly what are the users allowed to do with the data. At the same time, the data should explicitly reference its licence to provide legal certainty. With explicit licences users of data no longer find themselves in a legal vacuum with no clear guidaince on how they can use the data.

However, even though explicit licensing is a fundamental requirement for publishing of both open and non-open data, data producers often neglect to conform with it. For example, 82.16% of data sources in the Linked Open Data Cloud, the diagram overviewing linked open data sources, do not provide any licensing information [17]. Similar situation may be observed for the Czech public sector data, for which the licence is left unspecified in the majority of cases.[10]

An essential goal of open licences is to achieve equal opportunities to access and use of the licensed work. An open licence should thus be non-exclusive, non-dicriminatory, enabling free reuse and redistribution of the licenced data. It should be agnostic of both users and types of use. Therefore, it should not discriminate against any persons or groups, fields of endeavour, or any types of prospective use for the data. Open licences should permit any type of reuse, allowing modifications and creation of derivative data, and any type of redistribution that provides access to data to others.

Access to data must not be restricted by administrative barriers or geography. Limiting access rigths only to citizens of a particular country is unacceptable. On the contrary, enabling access only to a pre-defined group of people is not sufficient. For example, Creative Commons Developing Nations License[11] makes licensed content open only to the citizens in developing countries and as such is not considered to be an open licence.

Even though the primary objective of open licences is to remove obstacles to access and use, licences may stipulate some permissible requirements that the licensees using the licensed content need to comply with. At maximum, an open licence may require attribution to the original author and redistribution with the same or analogous licence.

However, the requirement for attribution can cause difficulties when multiple datasets are reused and combined. This problem is known as "attribution stacking" because the number of parties that have to be attributed increases with the number of datasets that are involved in reuse and come from different authors.

A similar problem to the attribution stacking and spreading arise with share-alike licences that require the same or analogous licence to be used for redistribu-

[10] http://cz.ckan.net/dataset?license_id=notspecified

[11] http://creativecommons.org/licenses/devnations/2.0/

tion. Share-alike licences are "viral" licences, for which the licensed content is their carrier. They may prove to be difficult to work with in cases where data available under the terms of different viral licences are combined and redistributed.

Open data is advised to be equipped with a standard, generic licence. If a custom licence is applied, it makes the use of data more cumbersome, because the user has to first study the unknown licence, instead of relying on terms and conditions of a well-known licence. Thus, the use of a custom licence may imply high transaction costs associated with using the licensed content.

The way users interface with data may be made even more uniform if a single licence is applied. In a controlled setting, such as in the public sector, establishing a unified licence is encouraged to simplify conditions of use, particularly for combining multiple datasets. Nevertheless, data provision under the terms of one licence is unlikely to scale. There are far too many different conditions around data which no single licence can cover.

Open data licences are considered to be those that conform with the Open Definition. Open Definition is a widely established definition of what it means for information to be open. *"A piece of content or data is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and share-alike"* [84]. The definition focuses on the legal aspects of openness and as such it is closely tied to licences that enable open distribution.

Several existing licences conform with the requirements on legal openness of open data. Some of them are the generic licences that may be used regardless of the context.

For example, among the generic licences recommended for open data the commonly applied ones include Creative Commons Zero[12] (CC0) and Open Data Commons Public Domain Dedication and License[13] (ODC PDDL). As a matter of fact, CC0 is not a licence, but a waiver that puts the licensed content in the public domain. As discussed in the previous parts, in some states legislation does not allow content to enter in the public domain by artificial means, such as with a waiver. In such cases, ODC PDDL may be applied because it contains not only a waiver but a licence agreement too, which sets the conditions of use for the licensed content to be the same as for the public domain content.

General-purpose licences may be substituted by licences with a specific purpose. An example of this type of licence is UK Open Government Licence[14] that was designed for releasing open data in the UK public sector in particular.

[12] http://creativecommons.org/publicdomain/zero/1.0/

[13] http://opendatacommons.org/licenses/pddl/1-0/

[14] http://www.nationalarchives.gov.uk/doc/open-government-licence/

3.2.2 Technical Openness

Technical openness is reflected in the recommendations how to make use of technologies to ensure equivalent access and use.

3.2.2.1 Accessibility

The requirements of open data principles covered in this section answer the question how one can obtain the data. Access is important because it necessarily precedes reuse. Making data accessible can be thought of as the next step after making it legally open.

Discoverability In order to be able to access a dataset, you need to discover it. The information that the data actually exists is a necessary prerequisite to data access [32]. Users of open data should be able to discover where the data is and locate where are the parts of data distributed. Essentially, discoverability is the ability to get from a known URI to a previously unknown URI, which may be used to retrieve the data. There are two main approaches to make data discoverable. The URI known to a user may be of a data catalogue or a search engine.

Discoverability is the reason why data should be equipped with a thin layer of commonly agreed descriptive metadata [31, p. 8], such as in a data catalogue or, more broadly, an information asset register [4]. Data catalogue may form a single access interface to data. For instance, PublicData.eu[15] is an example of an unofficial data catalogue of Europe's public data. An official pan-European data portal is planned by the European Commission to be started in 2013 [31, p. 10].

Another way of making data discoverable is to make data accessible to machines, such as search engines, that will index the data and enable it to be found. Machines that index data also profit from access tools, such as descriptive metadata. Indexers may use either the full content of data, if it is machien-readable, or even catalogue records representing the data. However, there are specific types of metadata that can be used to improve discovery by machines, such as site maps[16] that describe information architecture of the way the data is distributed, or `robots.txt`[17] files that police access control for the data.

Accessibility Carl Malamud claims that *"today, public means on-line"* [73]. Open data should be available online on the World Wide Web, retrievable via `HTTP GET` requests. There should be both access interface for human users, such

[15] http://publicdata.eu/

[16] http://www.sitemaps.org/protocol.html

[17] http://www.robotstxt.org/robotstxt.html

as a web site or an application, and access interface for machine users, such as an API or downloadable data dump.

There are a number of ways how to make data accessible online. A common and widely recommended practice is to publish data exports that users may use to download the data in bulk. An option that has lately fallen out of favour is the use of File Transfer Protocol (FTP) to distribute the data dumps. Currently, this option has been replaced by exposing the data via Hypertext Transfer Protocol (HTTP), so that one may retrieve it via `HTTP GET` requests. An efficient alternative to HTTP is to use peer-to-peer file sharing via the BitTorrent protocol instead. However, this technology has not yet received nearly as widespread use as HTTP, particularly in the public sector.

There should be no barriers obstructing access to data, coming from both technological restrictions and policy rules. No party or website should have a privileged or exclusive access to public sector data. There should be no financial cost associated with the use of data, although recovering reasonable marginal costs of data reproduction is acceptable in a limited number of cases in which reproduction of data incurs expenses to its producer.

To safeguard user's privacy and confidentiality any mechanism that identifies users should be prohibited [4] and instead anonymous access without requiring to login should be provided. Protecting user's identity by providing anonymous access is not possible with reactive disclosure that is based on interacting via freedom of information requests. However, proactive disclosure permits users to access data without sharing their identity [16, p. 69]. Users should not be required to register, albeit requesting users to apply for an Application Programming Interface (API) key is reasonable, especially when the data producer needs to control the load on servers hosting the data. There should be no password protection, no strict limit on the number of API calls, and no encryption hindering in access to data.

Permanence Open data should be accessible in the long term. A technical infrastructure needs to be in place to ensure long-term availability of public sector data [31, p. 8]. The overall permanence comprises of the permanence of content, access mechanisms, and software.

Data publishers should have back-up strategies. A common approach to maintaining data permanence is to have data both in exchange formats and preservation formats. Formats employed for storing data for the purpose of preservation should be sustained by a strong community of users or by a standards body, because obsolescence of data format may prevent archival access [104].

To ensure future accessibility of data the data access points, from which the data can be retrieved, should be persistent. Roy Fielding argues that *"the quality of an identifier is often proportional to the amount of money spent to retain its validity"* [39, p. 90]. Identifying resources with persistent access points has the benefit that consumer knowing the identifier does not need to re-discover the

identified resource during each attempt to access it [6]. The sustainability and reliability of data access methods is important especially due to the direct reuse of data, such as in applications built on top of data APIs, or in the cases when the data cannot be copied or it is not efficient to do so. A solution for this requirement may be to introduce indirection by providing a layer redirecting access requests to variable locations of the data, such as with persistent URLs.[18]

Software that implements support for the format of data needs to be preserved as well. Long-term availability of such software is required to preserve the ability to use the data. In this perspective, relying on a single software vendor increases the likelihood of obsolescence and should be thus avoided in favour of data formats that are supported by multiple vendors.

3.2.2.2 Use

This group of principles covers the affordances expected for open data. It highlights the features of data that are deemed to be fundamental in opening data up to a variety of uses. Consequently, it warns of technological choices that may cause unintended usage limitations.

Non-proprietary Data Formats Open data should use data formats over which no entity has exclusive control. Specifications of open data formats should be community-owned, free for all to read and implement, subject to no fees, royalties, or patent rights. Public review should be a part of the decision-making process in the format's development in order to enable participation both from implementers and users of the format. For example, the World Wide Web Consortium's has an open and well-defined process[19] for making standard data formats.

Using proprietary data formats excludes users of a platform or a software that, for the developers of which it is not allowed to implement support for the format. Hence, by using a proprietary format users are confined to acquire software from a single vendor. Data producers risk not being able to change software supplier, experiencing vendor or product lock-in. Relying on proprietary data formats for storing data comes with the risk of them becoming obsolete. These are some of the reasons why it is important to adopt a non-proprietary format for open data. For example, unlike spreadsheets' formats from commercial vendors, Comma-separated values (CSV) is a non-proprietary data format that is more suitable for open data.

Standards Adhering to a set of common standards makes reuse easier as the data can be processed by a wide array of standards-compliant tools. Standards create expected behaviour, enable comparisons, and ultimately lead to superior

[18] http://purl.org

[19] http://www.w3.org/2005/10/Process-20051014/

interoperability. Standards, such as controlled vocabularies and common identifiers, provide better opportunities for combining disparate sources of data. Consistent use of standards leads to "informal" standards encoded as best practices.

For example, standards from the World Wide Web Consortium are appropriate for open data.[20]

Machine Readability Machine readability is a property of the data structure. Machines parse ("read") structures. The more machine-readable the data is, the smaller is the unit that can be read. High level of partitioning in the data structure leads to a greater readability.

For instance, when machines are dealing with scanned documents saved as images in PDF files, the smallest unit they can meaninfully distinguish is the whole file, a blob of data that is opaque to them. On the other hand, when machines read HTML files, the smallest unit that can be read may be one HTML element or even one character.

What is most frustrating is when public servants think it is a good thing if they transform data from a machine-readable format, such as XML, into a format that is not machine-readable, such as PDF [16, p. 27]. While users of the data can convert it from XML to PDF, they cannot convert it from PDF to XML. Tim Koelkebeck writes that *"storing structured information via structureless scanning is the e-government equivalent of burning the files"* [65, p. 278].

The term "machine-readable" is a bit misleading when interpreted strictly. Machines can "read" all digital information. However, some data formats do not leave open many ways how the data may be used. For example, binary formats, such as images or executables, do not lend themselves to other types of use than display or execution, and as such they limit the possibilities of reuse. Therefore, open data should be stored in textual formats (e.g., CSV) with explicit and standard character encoding (e.g., UTF-8).

Open data should be captured in a structured and formalized data format that enables automated processing by software. Daniel Bennett writes that *"structure allows others to successfully make automated use of the data"* [10]. Users should be able not only to display the data, they should also be able to perform other types of automated processing as well, such as full-text search, sorting, or analysis.

Open data should be valid, conforming with its format's specification. Even though, minor errors may be handled by error recovery process of the user's software. For example, web browsers are very tolerant of malformed HTML. However, in general, syntax errors increase the cost of using data, because fixing such mistakes always involves human intervention [105]. Thus, data that contains errors severely violating specification of its data format cannot be considered as machine-readable.

[20] http://www.w3.org/standards/

Machines are users of data too, and thus providing data in a machine-readable format avoids discriminating them. However, *"most government websites weren't designed to share data with other websites"* [65, p. 205]. People view data through machines and machines help them to process it efficiently. For example, one of the main types of data intermediaries are search engines. Therefore, it is important that search engines can access and crawl open data. Another example where machine readability is crucial is big data, since people are not able to process large volumes of data and have to pre-process them with machines first. Machine readability is also important for people who cannot read (e.g., visually impaired, disabled), for whom machines must read (e.g., screen readers).

The connection between the licensed work and the terms of its licence may be made even more explicit by using a machine-readable licence statement. There are several ways how to indicate a licence so that it can be recognized automatically. A widespread method to do this is to embed a qualification of the type of link to the licence.[21] Having the licence attached to data in a way that is meaningful to machines comes with benefits for the users, such as the ability to search for reusable photos under the terms of a particular licence.

Safety Open data should be published in data formats that cannot contain executable content. Such content may contain malicious code harmful to the users of the data. Textual formats, which are recommended for disclosure of open data, are safe to use. On the other hand, Microsoft Office files are not considered to be safe, since they can contain executable macros.

3.2.3 Data Quality

Data quality is complementary to data openness. It is a set of features of data that are not essential for its openness but they are closely related.

3.2.3.1 Content

A primary facet of data quality is the type of content that is included in data. The following group of requirements instructs producers of open data about what should be in their datasets.

Primariness Data is traditionally available in finished products, such as compiled in reports. However, the call for "raw data now" asks rather for disaggregated and un-interpreted data [42]. Open data should thus be made available at the earliest point when it is useful to businesses and citizens [43]. A similar principle is adopted in the open source community, incarnated in the slogan "release early,

[21] For example, with Microformats. http://microformats.org/wiki/rel-license

release often", that emphasizes the importance of a tight loop of gathering and applying user feedback, which steers the released product towards a better quality.

Data should be collected at the source with the highest possible level of granularity to achieve maximum accuracy. It is desirable to strive for high precision of data, because it reflects the depth of information encoded in data [105]. Accuracy then represents the likelihood that the information extracted from the data is correct. For example, publishers of open data should provide fine-grained data with high resolution, with high sampling rate, such as high-definition images or video.

Completeness All public data should be made available, except direct or indirect identifiers of persons, which constitute personally identifiable information, and data that need to be kept secret due to the reasons of national security. The goal of open data principles is make the public sector, not citizens, transparent. Complete datasets should be available to bulk download since whole datasets could be difficult or impossible to retrieve through an API.

Timeliness Essentially, all datasets are snapshots of data streams, capturing the current state of an observed phenomenon. Accordingly, the value of data can decrease over time. For example, weather forecasts lose most of their value after the day for which they predict the weather conditions. What is valid for all types of data is that the value of data decreases as the methods used to capture the data become obsolete.

Usefulness of data may quickly drain out as the data ages. A commercial from IBM[22] stresses the importance of real-time data for decision making. It claims that you would not have crossed a road if everything you had was a five-minute old snapshot of the traffic situation. This is the case of freedom of information requests, the procedure of which is too slow to obtain timely data. The long waiting periods for these requests may result in receiving out of date data.

Having the transient nature of most data in mind, data producers should publish it as soon as possible to preserve its value, such as with live feeds for frequently updated material [16, p. 33]. Preferably, the data should be released to the public at the time of its release for the internal use. In this way, the data can serve to help in achieving real-time transparency and can be treated as a news source.

Integrity To ensure the integrity of open data digital signatures may be used. Signatures serve to guarantee authenticity of data, tracing its digital provenance, and also preserve the integrity of data in course of its transfer to the user. Pub-

[22] http://www.dailymotion.com/video/xdaoae_ibm-commercial-the-road-intelligent_tech

lishing data with the secure HTTPS protocol may decrease the risk of tampering with data during its transmission.

3.2.3.2 Usability

Usability is a quality of data that account for how well the data can be used. Open data that is usable well has a lower cost of use. This section mentions three aspects of open data that contribute to its usability.

Presentation A human-readable copy of data should be available to alleviate the unequal levels of ability to work with raw data. Given the differing data literacy skills among users an effort needs to be taken to provide the largest number of people the greatest benefits from the data and to help them make "effective use" of it.[23] The primary format for human-readable presentation, which is recommended for open data, is HTML [10].

Clarity Open data should communicate as clearly as possible, using plain and accurate language. Descriptions in data should be given in a neutral and unambiguous language that does not skew the interpretation of data. They should avoid jargon or technical language, unless the terminology is well-defined and adds to the clarity of data. Data should employ meaningful scales that clearly convey the differences in data. Data should not contain extraneous information and superfluous padding that might distract users from the important parts of data or confuse them.

To widen the reach of data its descriptive metadata should use a universal language (e.g., English), while the content of the data should be language-independent. This is particularly important to improve the prospects of cross-country reuse.

Documentation An aspect that greatly contributes to usability of data is availability and quality of documentation. Providing documentation is important for users because it helps them understand the data. Tim Davies makes the point that *"data is also only effectively open if any code-lists and documentation necessary to interpret it (e.g., details of the units of measurement used etc.) is also made openly available"* [24, p. 1]. Documentation should require only general knowledge and should not presuppose knowledge of internal practices of the agency that produced the dataset. For example, documentation might explain how a dataset is structured and what abbreviations are used in it.

The need for explanatory descriptions of data may be demonstrated on Comma-separated values (CSV) data format. It is exactly the simple structure of CSV without any schema descriptions that makes interpretation of data in this format

[23] As dubbed by Michael Gurstein in [48].

difficult without an accompanying "codebook", domain knowledge, and manual data inspection [66].

3.3 Policies

Principles describe goals that cover what should be achieved. The goals need to be linked to ways how to accomplish them. It needs to be clear how to implement goals and thus translate principles into action.

For this purpose, policies are made. They represent a pragmatic use of principles, prescribing requirements for behaviour and resulting actions. Priciples need to be distilled into policies, in order to provide direct guidance and practical steps to be taken by their implementers.

Policies should supplement principles with motivations. They should explain their objectives along with their prospective outcomes. The motivations should be underlaid with benefits to be yielded or sanctions to be imposed on those disobeying the policy. The examples of benefits of open data that may be used for this purpose are covered in 5.

Another motivation to make public data more accessible and usable was presented in the research paper *Government data and the invisible hand* [91]. The proposal suggested that there should be a policy requiring public bodies to access their data in the same way the public may access them: *"The policy route to realizing this principle is to require that federal government Web sites retrieve their published data using the same infrastructure that they have made available to the public"* [91, p. 170].

Compliance with policies must be reviewable. Control mechanisms, such as performance indicators or tests, should be designed in order to determine if sanctions should be applied. A contact person must be designated to respond to people trying to use the data and address the complaints about violations of the principles embodied in open data policies.

Open data policies were generally made in the last few years, however, the term "open data" appeared in a policy context several years before. Harlan Yu reports the earliest "open data policy" to be from the 1970s [119, p. 8]. It was a US science policy that insisted on NASA partners to have an *"open-data policy comparable to that of NASA [...] particularly with respect to the public availability of data"*.

Policies may be issued at different levels of the public sector, either at the level of state government or by local administrations. An example of an open data policy is the *Open Government Directive* from Barack Obama's administration in the US, which ordered all agencies in the public sector to publish their non-classified datasets on the Web [86].

3.4 Open Data for Public Sector Information

Like data in general, public sector information seems to be predisposed to be opened. The key argument in favour for opening up public sector information is that this information belongs to the public. Joseph Stiglitz, a noted economist, writes: *"[...] Who owns the information? Is it the private province of the government official, or does it belong to the public at large? I would argue that information gathered by public officials at public expense is owned by the public – just as the chairs and buildings and other physical assets used by government belong to the public"* [103, p. 7]. Collection and maintenance of public sector data is paid for from public funds derived from tax incomes. Therefore, the data should be treated as a public good, which enables equal levels of access and use not only to the public sector officials, but to every citizen as well. In other words, paraphrasing an Internet meme, *"All your data are belong to us"* [65, p. 241].

The public owns the public sector data and demands it to be openly available [5]. In 2010, survey by Socrata showed that there was a strong support for open data in the public sector [100]. It showed that 92.6% of civil servant would commit to open data and that 67.2% of citizens agreed with opening up of public sector data.[24]

The interest of citizens in data from the public sector may also be illustrated by the existence of community alternatives to public sector data[40]. For example, the demand for geo-spatial data may demonstrated by the projects like OpenStreetMap,[25] for which volunteers are "re-engineering" the data that should have been provided by the public sector.

Given the predispositions of public sector information to being opened, the demand for it, and the technologies that make it possible to be opened, one may expect an increase in activity in this domain. Open data in the public sector went from being a niche cause to being pervasive in the whole world. Now, there is over a hundred initiatives opening up data in the public sector world-wide [25], building up to a global, networked data infrastructure.

3.4.1 Data Infrastructure

Information infrastructure is a necessary prerequisite for all information-demanding services. In his treatment on networks Yochai Benkler describes the need for a shared infrastructure. *"To flourish, a networked information economy rich in social production practices requires a core common infrastructure, a set of resources necessary for information production and exchange that are open for all to use. This requires physical, logical, and content resources from which to make*

[24] The surveyed sample in the study contained 1000 adults, out of which 300 were self-identified government employees.

[25] http://www.openstreetmap.org/

new statements, encode them for communication, and then render and receive them" [9, p. 470]. Ursula Maier-Rabler ties these insights to the public sector. *"The prerequisite for the functioning of networks is a common infrastructure. The role of government is to provide that infrastructure"* [72, p. 187].

In the current state of affairs, there are multiple fragmented infrastructures that the performance of public functions depends on. Moreover, it is common that these infrastructures are available to dedicated applications only, while being closed to applications from other parts of the public sector, let alone the ones created by members of the public. These information infrastructures are neither shared nor open.

Open data may serve as a data infrastructure of the public sector. By definition, it constitutes a fundamentally open and shared infrastructure, that is in line with the Benkler's vision. Such infrastructure not only enables public services to run; but, because it is open to everyone, it also enables private services to run. Building such infrastructure is the goal of open data initiatives and policies.

3.4.2 Government as a Platform

Open data infrastructure is the gist of the concept of "government as a platform" formulated by Tim O'Reilly [65, p. 11]. O'Reilly expands on the notion of open data by demanding governments to expose not only raw open data, but also open web services. Government as a platform is a provider of services built on open data. The services, accessible to anyone, offer ways of interfacing with data on which they are based, allowing to perform basic operations on that data. In this way, these open services form an API for the public sector.

This line of thinking sees the public sector as an enabler rather than an implementer, focusing more on creating an open environment rather than delivering end-user services. In contrast to government that works as a platform, current governments may be described rather as "vending machine governments" [65, p. 13]. In such governments citizens pay taxes and expect services in return. If no services are provided or the obtained services are not satisfactory, citizens protest, which is like shaking the vending machine.

If we get on a more metaphorical level of the "government as a platform" concept, as Carl Malamud does, we can see law as the operating system of society [65, p. 45]. Law provides rules that govern society, similar to operating systems governing the allocation of system resources. For an open and democratic society not only an unfettered access to its underlying infrastructure is necessary, it is crucial to guarantee equal access to law as well. As Malamud puts it, *"if a document is to have the force of law, it must be available for all to read"* [65, p. 46]. Law, the operating system of society, has to be made open source.

What is important on the government as a platform is that this idea needs generative data. Jonathan Zittrain defines generativity as the *"system's capacity to produce unanticipated change through unfiltered contributions from broad and*

varied audiences" [120, p. 70]. It is a property that describes the ability of users of the system to produce new content unique to that system without any input from the system's creators. The generativity of a system is based on its affordances, *"the possible actions that exist in a given environment"* [120, p. 78].

Platforms balance control with generativity. Open infrastructures favour generativity and loose control mechanisms. Open data model incentivizes peer production of applications based on the data [54, p. 331]. Jonathan Zittrain claims that *"generatively-enabled activity by amateurs can lead to results that would not have been produced in a firm-mediated market model"* [120, p. 84]. This is the essence of the Many minds principle that asserts that *"the coolest thing to do with your data will be thought of by someone else."*[26]

Bill Schrier writes that *"governments should provide services which are difficult or impossible for the public to provide for themselves, or which are hard to purchase from private businesses"* [65, p. 305]. The rest of the services should be catered for by the public, by businesses or civic associations. What contributes to this approach is the recognition that *"the needs of today's society are too complex to be met by government alone"* [83]. Ultimately, *"if the private sector can make downstream products more cheaply or meet consumer demands in other ways, then the public sector body should consider pulling out of the market"* [43, p. 38]. The solution is to open up the data infrastructure that the public sector works on and invite third parties to build on it. In this way, exposing public sector data within an open infrastructure enables to complement government-provided services with citizen self-service.

Although the government as a platform principle is still in an early stage of realization, there are several places in which the public sector opened up its infrastructure to others. To give an example of this principle in action, the Global Positioning System (GPS), that the US government made publicly available for full commercial use a decade ago, may be considered [65, p. 44]. Built on geospatial data, this system provides geolocation services that are open to anyone to access, free of charge.

Highly successful, yet short-lived, were the occasions in which the public sector opened its data for application challenges. In these competitions public bodies released some of their data and offered prizes for the best applications developed with that data. The challenges proved to have a high return on investment. Not only they created a value in applications that significantly exceeded the original investment in prizes, but the application contests also delivered tangible examples of what data can do. Application challenges, such as the founding Apps for Democracy, that took place in Washingtion D.C. in 2009,[27] were a source of inspiration for others to follow their lead.

Finally, there already is software being built for creating open data infras-

[26] http://assets.okfn.org/files/talks/xtech_2007/

[27] http://www.appsfordemocracy.org/

tructures. An example of such software is the aptly-named *Open Government Platform*[28] dataset management system that is jointly developed by the US and India.

3.5 Summary

This chapter summarized the key concepts behind open data, its essential features, and principles that data have to comply with in order to be recognized as "open". After enacting open data, the chapter delved into its practical side including policies for translating open data principles into prescribed actions and the possibilities of applying open data in the public sector to build a common data infrastructure. In this way, the chapter prepared the ground for the following part that discusses the application of linked data to implement the requirements on technical openness of data.

[28] https://github.com/opengovtplatform/opengovplatform-DMS

4 Linked Data

Linked data is a publication model for structured data on the Web. The term "linked data" was coined by Tim Berners-Lee in 2006 [12] in a note in which he described the Linked Data Principles introduced further in this chapter 4.2.1.

An essential feature of linked data is materialization of relationships. Linked data makes implicit relationships between the described things explicit by materializing them as data [7, p. 94]. Reified relationships expressed as links thus become a part of machine-readable data amenable to automated processing. Traditionally, relationships in data are kept implicit as a part of the background knowledge, documentation, or software. In such cases, integration of data is done on the application level with a custom-crafted code or queries and the effort of discovering relationships in disparate datasets is left to application developers and other data consumers. Materialization of relationships in linked data shifts this integration effort to the data level.

While the current Web turned out to be mostly a web of documents, linked data leads to a growth of a web of data. This web of data may describe not only documents but may also include data, abstract ideas, or physical objects, along with their materialized relationships. In this way, linked data offers a seamless integration of the web of documents and the web of things into the web of data. Marko Rodriguez supposes that *"the web of data may emerge as the de facto medium for data representation, distribution and ultimately, processing"* [92, p. 38].

Linked data is a fundamentally distributed publishing model that locates data in heterogeneous data spaces. Unlike the current data stores that may be likened to silos or terminal nodes, linked data spaces are mutually connected via hyperlinks, through which disparate data sources may be defragmented and integrated into a single, virtual global data space. For linked data, relationships with other data expressed via links are of fundamental value. In his note about linked data Tim Berners-Lee claims that *"the value of your own information is very much a function of what it links to"* [12].

This chapter introduces the core concepts of linked data and argues how linked data may satisfy the requirements on technical openness of open data presented in the previous chapter (3). The following sections present an overview of the involved technologies, principles, and practices, concluding with a discussion of the application of linked data for open data in the public sector.

4.1 Technologies

Linked data may be seen as a pragmatic implementation of the vision of the so-called "semantic web", that is the web that communicates meaning in a way machines can operate on. Linked data has a mature and well-understood technology stack [52] comprised of the semantic web technologies. Most of these technologies are developed and standardized at the World Wide Web Consortium (W3C).[1] In the following section the key technologies for linked data will be introduced: Uniform Resource Identifier for identification of data, Hypertext Transfer Protocol for interaction with data, and Resource Description Framework for data representation.

4.1.1 Uniform Resource Identifier

Uniform Resource Identifiers (URI) offer an extensible, federated naming system for universal and global identification [90, p. 6]. Thanks to URI's universality, resource identified with a URI may be anything, including web sites, ideas, and real-world objects.

URIs and Uniform Resource Locators (URLs) are different. URI needs not to locate the resource it identifies. Location of a resource is described by a URL, that in addition to identifying the resource provides a way of addressing it. In some cases, a resource may have the same URL as URI. This is true for information resources that may be retrieved via the Web. However, resources that may not be retrieved via the Web, such as physical objects, have a URI but do not have any URL, since they cannot be located in that way.

Resource needs not to be identified with a single URI because linked data adopts the non-unique name assumption allowing equivalent resources to have multiple URIs. This approach lowers the start-up barriers for data modelling since it lets linked data publishers to assign resources with their own URIs instead of making the effort to find the URIs that already exist for such resources.

4.1.2 Hypertext Transfer Protocol

Linked data uses URIs with the `http` scheme that are handled by the Hypertext Transfer Protocol (HTTP), *"an application-level protocol for distributed, collaborative, hypermedia information systems"* [89, p. 1]. HTTP is the default interaction protocol for linked data that is used for data exchange, querying, updates, and so forth. Linked data uses HTTP in accordance with the constraints of the Representational State Tranfer architectural style that is described in the next section.

[1] A dedicated page for standards relevant to the semantic web may be found at http://www.w3.org/standards/semanticweb/.

4.1.2.1 Representational State Transfer

The resource-oriented architecture of linked data may be considered as a style that builds on Representational State Transfer (REST). REST is an architectural style defining a stateless communication protocol for distributed client-server applications, such as the World Wide Web. Roy Fielding, the author of REST, defines architectural style as a *"coordinated set of architectural constraints that has been given a name for ease of reference* [39, p. xvi]. In his doctoral dissertation Fielding defines four interface constrains for REST:

- identification of resources with URIs
- manipulation with resources through their representations
- self-descriptive messages
- hypermedia as the engine of application state

Linked data interfaces adopt these constraints and they build onto them further constraints based on the Linked Data Principles, that are described in the section 4.2.

4.1.2.2 Dereferencing

Dereferencing is a basic mechanism built on REST that linked data employs for interaction with URIs. By minting a URI in a namespace, namespace owners *"enter into an implicit social contract with users of their data"* [59] and should be therefore aware that *"there are social expectations for responsible representation management by URI owners"* [60]. The expectation the users of URIs have is that there are dereferencing mechanisms implemented for the URIs, which work in a predictable manner.

Dereferencing is an idempotent operation on URI that exchanges reference to a resource for the resource. HTTP agent (e.g., a web browser) that dereferences a URI issues an HTTP GET request for the resource's reference (i.e., a URI) and the HTTP server administering this reference replies with a response containing the resource or its representation. The response should be accompanied by a correct HTTP `Content-type` header indicating the data format of the response encoded with the Multipurpose Internet Mail Extensions (MIME). Dereferencing can be indirect as redirects may be employed, which is a common practice especially for persistent URIs and non-information resources.

According to the Architecture of the World Wide Web [60] there are two kinds of resources, information resources and non-information resources, for which different dereferencing mechanisms apply. Information resource is *"a resource which has the property that all of its essential characteristics can be conveyed in a message"* [60], and so it may be trasferred via HTTP (e.g., HTML or PDF

files).[2] Non-information resources are those resources that cannot be transferred via HTTP, such as physical objects or abstract notions.[3] Since the owner of a URI of a non-information resource cannot serve the user requesting the URI with the identified resource, a recommended, yet widely disputed practice suggests to reply with the `HTTP 303 See Other` status code redirecting users to a URI of a representation of the non-information resource [52].

4.1.2.3 Content Negotiation

Content negotiation is a way how to decide on an appropriate response format based on the content of HTTP request's headers. HTTP clients can send HTTP headers along with the requested URI to provide context, stating what format of representation of the requested resource they prefer.

A common HTTP header that is used for this purpose in the linked data publication model is the `Accept` header that contains an enumeration of the preferred MIME types for the representation of the requested resource. This pattern allows the client to negotiate with a server on the format of the server's response that is appropriate for the actual communication context. In practice, this is a way how the server may distinguish between human and machine traffic and serve either a human-readable (e.g., HTML) or a machine-readable (e.g., XML) representation of the requested resource.

Principles of content negotation offer a generic approach to communication of the client's preferences. A widespread use of content negotiation may be demonstrated on the `Accept-Language` header that may be used to indicate preferred language of the response. A novel use of this method is the datetime content negotiation that allows the client to access different time snapshots of data using the `Accept-Datetime` header, which is implemented in the Memento software.[4]

There are multiple ways and levels on which content negotation may be implemented. A common way to do it is by configuring HTTP server, such as with the Apache HTTPD's `mod_rewrite`. A recommended way to enable discovery of the supported types of representations is to use the `link` HTML element with a link typed `"alternate"` and the `type` attribute describing a MIME type that the server is capable of responding with.

---

[2] For example, http://dbpedia.org/page/Czech_Republic is a URI of an information resource identifying a page about the Czech Republic.

[3] For example, http://dbpedia.org/resource/Czech_Republic is a URI of a non-information resource representing the Czech Republic.

[4] http://www.mementoweb.org/

4.1.3 Resource Description Framework

Resource Description Framework (RDF) is a standard format for data interchange on the Web. RDF is a generic graph data format that has several isomorphic representations. Any given RDF dataset may be represented as a directed labelled graph that may be broken down into a set of triples, each consisting of subject, predicate, and object.

Triples are items that RDF data is composed of. Subject of a triple is a referent, an entity that is described by the triple. Predicate-object pairs are the referent's characteristics.

RDF is a type of entity-attribute-value with classes and relationships (EAV/CR) data model. EAV/CR is a general model that may be grafted onto implementations spanning relational databases or object-oriented data structures, such as JSON. In the case of RDF, *entities* are represented as subjects, which are instances of classes, *attributes* are expressed as predicates that qualify relationships in data, and objects account for *values*.

In terms of the graph representation of RDF, subjects and objects form the graph's nodes. Predicates constitute the graph's vertices that connect subjects and objects. The graph's nodes and vertices are labelled with URIs, blank nodes (nodes without intrinsic names), or literals (textual values).

4.1.3.1 Serializations

RDF is an abstract data format that needs to be formalized for exchange. To cater for this purpose RDF offers a number of textual serializations suitable for different host environments. A side effect of RDF notations being text-based is that they are open to inspection as anyone can view their sources and learn from them. Now we will describe several examples of the most common RDF serializations.

N-Triples[5] is a simple, line-based RDF serialization that is easy to parse. It compresses well and so it is convenient for exchanging RDF dumps and executing batch processes. However, the character encoding of N-Triples is limited to 7-bit and covers only ASCII characters.[6]

Turtle[7] is a successor to N-Triples that provides a more compact and readable syntax. For instance, it has a mechanism for shortening URIs to namespaced compact URIs. Unlike N-Triples, Turtle requires UTF-8 to be used as the character encoding, which simplifies entry of non-ASCII characters.

RDF serializations based on several common data formats were developed, such as those building on XML or JSON. XML-based syntax of RDF[8] is a W3C

[5] http://www.w3.org/TR/rdf-testcases/#ntriples

[6] Other characters have to be represented using Unicode escaping.

[7] http://www.w3.org/TR/turtle/

[8] http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/

recommendation from 2004. With regard to JSON, there are a number of proposed serializations, such as JSON-LD, an unofficial draft for representing linked data.[9] However, these serializations suffer from the fact that their host data formats are tree-based, whereas RDF is graph-based. This introduces difficulties for the format's syntax as a result of "packing" graph data into hierarchical structures. For example, the same RDF graph may be serialized differently with no way of determining the "canonical" serialization.

Several RDF serializations were proposed to tie RDF data with documents, using document formats as carriers that embed RDF data. An example of this approach is RDFa[10] that allows to interweave structured data into documents by using attribute-value pairs. It is a framework that can be extended to various host languages, of which XHTML has a specification of RDFa syntax that reached the status of an official W3C recommendation.[11]

4.1.3.2 Vocabularies and Ontologies

While RDF is a common data model for linked data, RDF vocabularies and ontologies offer common way of describing various domains. Their role is to provide a means of conveying semantics in data. RDF vocabulary or ontology covers a specific domain of human endeavour and distills the most reusable parts of the domain into *"an explicit specification of a conceptualization"* [47, p. 1]. Conceptualization is thought of as a way of dividing a domain into discrete concepts.

The distinction between RDF vocabularies and ontologies is somewhat blurry. Ontologies provide not only lexical but also intensional or extensional definitions of concepts that are connected with logical relationships, and thus are thought of as more suitable for the tasks based on logic, e.g., reasoning. RDF vocabularies offer a basic "interface" data for a particular domain and as such as better suited for more lightweight tasks. Most of linked data gets by with using simple RDF vocabularies, that are in rare cases complemented with ontological constructs.

Having data described with a well-defined and machine-readable RDF vocabulary or an ontology enables to perform inference on the data. Inference is a type of inductive reasoning for materializing data implied by the rules defined in RDF vocabularies and ontologies, through the means of which the data is expressed. W3C standardized two ontological languages that may be used to create RDF vocabularies and ontologies: RDF Schema (RDFS)[12] and Web Ontology Language (OWL)[13].

[9] http://json-ld.org/spec/latest/json-ld-syntax/

[10] http://www.w3.org/TR/rdfa-core/

[11] http://www.w3.org/TR/rdfa-syntax/

[12] http://www.w3.org/TR/2004/REC-rdf-schema-20040210/

[13] http://www.w3.org/TR/owl2-overview/

There are countless RDF vocabularies and ontologies available on the Web. However, a great deal of them is used only in the dataset, for which they were defined, and only a few of them reached a sufficient popularity in order to be treated as de facto standards for modelling of the domains they cover. An example of a general and widespread RDF vocabulary is Dublin Core Terms,[14] which provides a basic set of means for expressing descriptive metadata. With regards to the public sector, some of the RDF vocabularies and ontologies covering this domain may be found in the Government vocabulary space[15] of the Linked Open Vocabularies project.

4.2 Principles

Linked data principles govern the use of the semantic web technologies described in the previous sections. Unlike the technologies, the principles are not backed by any standards body, such as the World Wide Web Consortium. Instead, they are community-driven and their sole enforcement mechanism is peer pressure. Nevertheless, this may turn out not to be the case in the near term future if the principles get incorporated into official policies and regulations, such as the ones that govern public sector institutions.

Linked data principles provide a guidance both for data publishers and consumers. For publishers, they offer the best practices that they have to comply with in order for their data to be recognized as linked data. From consumers' perspective, the principles prescribe behaviour patterns that they can expect when working with linked data, such as what happens when linked data URIs are resolved in the course of content negotiation.

Compared with the principles of open data, there are fewer instances of the linked data principles. The original Linked Data Principles [12] drafted by Tim Berners-Lee form a strong core that any other, and mostly derivative, linked data principles tend to cite or relate to.

4.2.1 Linked Data Principles

Linked Data Principles, written by Tim Berners-Lee in 2006, effectively define what is linked data [12]. The principles set a touchstone that may be used to determine if datasets qualify for being described as "linked data", by covering all the necessary conditions that datasets need to fulfil in order to earn that label. These conditions are encapsulated in four succinct principles.

1. Use URIs as names for things.

[14] http://dublincore.org/documents/dcmi-terms/

[15] http://labs.mondeca.com/dataset/lov/details/vocabularySpace_Government.html

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).

4. Include links to other URIs, so that they can discover more things.

Berners-Lee, the inventor of the World Wide Web, sees the principles as natural for the Web. He recounts that in writing down the principles he only captures his intentions that were already a part of his original architecture for the Web [12].

After the creation of the principles they were modified to a small extent, clarifying certain issues and making some parts more explicit. For example, the original version from 2006 did not explicitly mention what technologies should be used for achieving the prescribed behaviour of the data. This was amended later, making it clear that the technologies that were intended to be used were RDF and SPARQL.

4.2.2 Five Stars of Linked Open Data

Four years after the inception of the original Linked Data Principles Tim Berners-Lee proposed a more iterative take on publishing linked data in his Five Stars of Linked Open Data scheme [12]. It contains five commandments for data producers explaining how to proceed with improving the way how their data is published.

★ Publish data on the Web under an open licence (e.g., in PDF).

★★ Publish data in a structured format (e.g., in Excel).

★★★ Publish data in a non-proprietary format (e.g., in CSV).

★★★★ Use URLs to identify data, so that it is linkable (e.g., in RDF).

★★★★★ Link your data to other data to provide context.

A major change in this scheme is the recognition of the importance of open access to data, which is already required in order to earn the first star. The scheme emphasizes that adoption of linked data principles creates a space for continuous improvement. Data producers can start publishing data with a low up-front cost and consequently continue investing more resources towards the goal of joining the pool of linked open data.

There are several renditions of the Five Stars of Linked Open Data scheme besides the one done by Tim Berners-Lee himself [12]. For example, Ed Summers was among the first to publish the scheme[16] and Michael Hausenblas illustrated

[16] http://inkdroid.org/journal/2010/06/04/the-5-stars-of-open-linked-data/

the scheme with some examples along with associated costs and benefits for each of the steps described by the scheme.[17]

4.3 Linked Open Data

Due to its innate features linked data is considered as an appropriate technology for publishing open data. The way of opening data of the public sector using linked data is seen among the best ways to do so. For example, Tim Berners-Lee argues that *"each of the purposes of government data is best served by using linked data techniques"* [13] and Bernardatte Hyland posits that *"Linked Data is seen as critical means to satisfy emerging government mandates for transparency and accountability"* [114, p. 52]. At the same time, nothing prevents data producers to publish linked data with closed access, such as on enterprise intranets.

4.3.1 Technical Openness

The following sections review how linked data can support the publication of open data and how well it addresses the requirements posed by the principles of technical openness and data quality described in 3.2.2. The evaluation of linked data from the perspective is complemented with a discussion of comparative advantages of linked data for publishing open data as compared with the other state-of-the-art approaches.

4.3.1.1 Accessibility

Accessibility of linked data boils down to the ability to find URIs and the ability to dereference them.

Discoverability If we define discoverability as the ability to get to a previously unknown URI from a known URI, then this ability depends on the in-bound links from known URIs to unknown URIs. In particular, it depends on the quantity of in-bound links, how likely it is that the users will follow them, and discoverability of their referring URIs.

Linked data fulfils the basic requirement of being linkable by using static and persistent URIs. Moreover, guidelines on URI construction for linked data recommend using human-readable URIs that are easier to communicate [21, p. 4]. To increase the interconnectedness of data services were developed that take into account out-bound links as well, such as the PSI BackLinking Service for the Web of Data.[18]

[17] http://5stardata.info/

[18] http://backlinks.psi.enakting.org/

Dereferencing URIs serves as a way to discover more data. Self-describing resources of linked data *"promote ad hoc discovery of information"* [76]. The representations of resources the users obtain by dereferencing their URIs may contain links to other resources. This allows for a "follow your nose" link traversal exploration style, recursively navigating through the Web. Since dereferencing mechanisms adhere to a standardized protocol, it enables to automate this type of data discovery, such as with crawlers.

The methods to improve discovery of linked data may be categorized either as passive or active. Passive approaches consist in publishing additional data that makes the published linked data easier to find. To improve data traversal for crawlers Semantic Sitemaps[19] listing all the data access points may be published. Several RDF vocabularies were devised for expressing access metadata that help in data discovery, such as Vocabulary of Interlinked Datasets (VoID).[20] A common solution for keeping a record of available data is to post data description to a data catalogue, such as the Data Hub.[21] To address this purpose, Data Catalogue Vocabulary (DCAT)[22] was created.

Active techniques serve the purpose of notifying linked data consumers about the existence of data. A common way to spread information about data availability is to notify prospective consumers via the ping protocol, such as with web services like Ping the Semantic Web.[23] Submission of data to search engines works in a similar way, such as with the form[24] for notifying Sindice, a search engine for the semantic web.

Linked data also ranks well in regular search engines. For example, Martin Moore reported that in 2010 linked data resources from the BBC's Wildlife Finder appeared high in Google search results for animal names [78].

Accessibility Linked data requires using dereferenceable HTTP URIs that serve as open access points to data. Resolution of linked data URIs may be either implemented by serving static files or by generating resource representations on the fly.

Linked data may be published in static files in one of the RDF serializations described in 4.1.3.1. This approach is used mainly for serving RDF vocabularies and ontologies, transfer of datasets for local batch processing, or for files with embedded RDF. Serving static files is easy to implement, however, their content is fixed and difficult to manipulate and update.

[19] http://sw.deri.org/2007/07/sitemapextension/

[20] http://vocab.deri.ie/void

[21] http://thedatahub.org/

[22] http://www.w3.org/2011/gld/wiki/Data_Catalog_Vocabulary

[23] http://pingthesemanticweb.com/

[24] http://sindice.com/main/submit

To take advantage of the flexible nature of linked data on-demand, dynamically generated RDF representations may be served instead. One option for this approach is to use wrappers for dynamic data extraction from non-RDF data sources. For example, D2R Server[25] allows to expose relational databases as RDF through a pre-defined mapping.

However, to reap the full benefits of RDF a triple store should be used to store the data. Triple store is a database optimized for storage and retrieval of RDF data. To publish data from a triple store SPARQL endpoints are used as the interfaces users interact with. The endpoints expose an interface defined by the SPARQL Protocol for RDF,[26] which allows to query[27] or manipulate data[28] and serves the query results in XML via HTTP.

In order to comply with linked data principles publishers should use front-end applications that implement dereferencing and content negotiation. A common way how to expose RDF as linked data is through lightweight SPARQL wrappers that dereference URIs to concise bounded descriptions [102] of the requested resources, the descriptions of which they retrieve via SPARQL queries. Example implementations of linked data front-ends include Pubby[29] or Graphite.[30]

To ease the transition to the use of linked data for web developers specification of Linked Data API[31] was created. Linked Data API is a framework for more user-friendly APIs interacting with linked data in a way that follows the guidelines of REST and uses simple data formats, such as JSON. Among the example implementations of this framework are Puelia[32] and Elda.[33]

Permanence Linked data principles enforce separation of data and applications, which promotes permanence. Modelling linked data is modelling without a context of use [115, p. 11]. When designing a data model for linked data, its creators abstract away from particular uses the data may get, such as in specific applications. Such design principle results in an application-agnostic data model that is not tightly coupled with any type of use that might be intended for the data. As a result, the data supports a wide range of unintended and unforeseen uses. Given the data is decoupled from applications using it, it needs not to be changed when

[25] http://d2rq.org/d2r-server

[26] http://www.w3.org/TR/rdf-sparql-protocol/

[27] SPARQL Query Language for RDF is defined at http://www.w3.org/TR/rdf-sparql-query/.

[28] The specification for SPARQL 1.1 Update may be found at http://www.w3.org/TR/sparql11-update/.

[29] http://www4.wiwiss.fu-berlin.de/pubby/

[30] http://graphite.ecs.soton.ac.uk/

[31] http://code.google.com/p/linked-data-api/

[32] http://code.google.com/p/puelia-php/

[33] http://code.google.com/p/elda/

the implementation of interfaces mediating it changes. Moreover, the software used for publishing or consuming linked data is in most cases open source and thus needs not to be changed if a vendor providing it changes. Even if there was no support for these open source solutions, data formats used for linked data have open specificiations that may be re-implemented by anyone.

Established design patterns for linked data promote persistent URIs providing long-lasting access points [21, p. 5]. Several of the best practices for minting URIs contribute to their persistence. URIs should not be made session-specific, in which case they cannot be used for re-identifying the requested resources after the session expires. URIs should be made implementation-agnostic because if they depend on an implementation they cannot outlast it. Therefore, URIs should not be cluttered with implementation details, such as file type suffixes (e.g., `.php`). A technique that further decouples URIs from the way they are dereferenced is to introduce a layer of indirection by using a service such as `http://purl.org` to redirect URIs to URLs that serve their representations. However, ultimately the persistence of URIs is proportional to the commitment of institutions maintaining them.

4.3.1.2 Use

The flexible, application-agnostic nature of linked data makes it possible to employ it for a broad spectrum of uses. Linked data does not discriminate according to the type of use as *"Linked Data principles and publishing guidelines are designed to make structured data more amenable to ad hoc consumption on the Web"* [56, p. 13].

Roy Fielding wrote that *"the primary mechanisms for inducing reusability within architectural styles is reduction of coupling (knowledge of identity) between components and constraining the generality of component interfaces"* [39, p. 35]. Fielding's REST, covered in the section 4.1.2.1, is based on uniform interfaces between components and thus abides by this recommendation. However, a trade-off of uniform interfaces is of efficiency because such interfaces are optimized for the general case [39, p. 82]. Since linked data is based on REST it also inherits this trade-off.

Linked data adopts separation of concerns and decouples content from presentation. In this way, it decouples data from upstream (producers) and downstream (consumers) interfaces enabling variability without introducing interoperability costs. Since linked data is not application-specific it may be used to power all kinds of applications.

Modelling of linked data is based on the reuse of existing models provided by RDF vocabularies and ontologies. A common approach to modelling of linked data is to mix various vocabularies and ontologies at will, cherry-picking their components to build a customized model suited for particular data.

Flexibility of the RDF data model enables to query the data and reconfigure

it for a particular use. Semantic web technologies open opportunities for reuse by offering *"query interfaces for applications to access public information in a non-predefined way"* [2]. This is more difficult to achieve for non-RDF data formats. For example, Fadi Maali argues that *"providing the data in a fixed table structure, as in CSV files, makes it harder for consumers to re-arrange the data in a way that best fits their needs"* [71, p. 86].

Together, composing data models of parts of data models already known to applications and the flexibility that allows to rearrange the data model to the application model is facilitative to generic consumption. Such an advantage is particularly manifest when applications combine multiple sources of linked data The applications of this type are referred to as "meshups" since they are built on data sources that mesh with each other [82, p. 321]. Without linked data, this scenario would require manual integration effort on the application level, whereas linked data would be already integrated on the data level.

The following paragraphs provide answers on how linked data meets the concrete criteria on the use of open data.

Non-proprietary Data Formats RDF is a non-proprietary data format and its specifications are open and free for anyone to inspect and implement.

Standards Linked data builds on web standards maintained by the W3C or the Internet Engineering Task Force (IETF).[34]

Machine Readability RDF serializations covered in section 4.1.3.1 are machine-readable. Specifications of RDF serializations have well-defined conformance criteria, which facilitate the development of standard parsers and make it possible for data to be validated for conformance, such as with the W3C RDF Validation Service.[35]

RDF data is well-structured with a high level of granularity. Users of RDF may use it as a graph that may be broken down into individual triples, which allows access to data at a very detailed level.

Linked data makes explicit, machine-readable licensing possible by linking to licences. There are several RDF vocabularies that contain properties to do that, such as the Dublin Core Terms with `dcterms:rights`. For a structured representation of the licences themselves Creative Commons Rights Expression Language[36] may be employed.

[34] For an overview of standard specifications related to linked data see http://linkeddata-specs.info/.

[35] http://www.w3.org/RDF/Validator/

[36] http://wiki.creativecommons.org/CC_REL

Safety RDF cannot include executable content. Serializations of RDF are textual,[37] which promotes inspection and eases safety checks. However, using RDF in adversarial environments with security problems, such as RDF injection or query sanitization, is an area in which little research is conduced.

4.3.2 Data Quality

Data quality is not inherent in technologies but it is a result of the way technologies are used. Apart from the strict limitations of semantic web technologies and linked data principles enforced by peer pressure, there is a body of knowledge about linked data captured in informal design patterns and best practices, that is embodied in resources like Linked data patterns [27] or Cookbook for open government linked data [59]. Among the other aspects these recommendations deal with they propose ways how linked data should be used to achieve the best data quality.

4.3.2.1 Content

The content facet of open data quality metrics tracks if the content of data is primary, complete, timely, and delivered intact.

Primariness A key principle of linked data is to ensure access to raw data. Linked data URIs are required to dereference also to raw, machine-readable data, such as RDF in XML. Besides dereferencing, linked data may implement interfaces for access raw data, such as SPARQL endpoints.

Completeness A common way to arrange for the access to complete data is to provide data dumps exported from a database or a triple store in the back-end. In this way, users are allowed to work with the data as a whole.

RDF offers an inclusive way for representing data of varying degree of structure and granularity. Depending on the modelling style, RDF can capture both highly-structured data and unstructured free-text. Linked data improves this inclusiveness by enabling to link to non-RDF content.

Linked data offers a means for materialization of the types of data that are, for the most part, out of the scope of the other approaches to data representation. For example, it may include explicit relationships between the described resources. From this perspective, linked data may be seen as a more complete representation of a particular phenomenon.

[37] With the exception of the proposed Binary RDF [38].

Timeliness Even though timely release of data is rather a matter of policy and human resources, technologies employed for that task can make it easier. In particular with highly dynamic data that goes through frequent changes it is important to have a flexible update mechanism at hand. Updates of linked data may be automated with SPARQL 1.1 Update that offers a very expressive method for patching data.

Timeliness is crucial in two areas that are gaining prominence: streaming sensor data and user-generated content. Research on the technological solutions for these areas is in its infancy [95]. However, there already are experiments with streaming linked data or real-time extraction from user-generated content, such as DBPedia Live[38] that captures updates in Wikipedia in a near real-time.

Integrity The stack of the semantic web technologies, which linked data builds on, includes both digital signature and encryption as a part of the so-called Semantic Web Layer cake.[39] For ensuring the content of data is not tampered with during its transmission secure HTTPS connections should be employed. An example of semantic web technology that builds on digital signatures is WebID,[40] that may be used to authenticate data publishers.

4.3.2.2 Usability

Usability may be perceived as the weakest point of linked data. In most cases, raw, disintermediated linked data is not intended for direct consumption. This is the result of the separation of concerns that linked data employs. For example, consider working with a SPARQL endpoint that, even though it is a powerful way of interacting with data for applications, may be baffling for the regular users. Linked data should be rather mediated through end-user interfaces of web applications, that present the data in a more usable and visually-appealing manner. However, there are still aspects in which raw linked data excels when compared to other types of data.

Presentation Intelligible presentation of linked data should be arranged for by the implementation of mechanisms for dereferecing URIs, which should be able to serve a human-readable resource representation, such as in HTML. However, representations of linked data resources are usually generated into generic templates in an automated fashion, which impedes custom adaptation of representations for different resource types.

[38] http://live.dbpedia.org/

[39] http://upload.wikimedia.org/wikipedia/commons/4/47/W3c-semantic-web-layers.svg

[40] http://www.w3.org/wiki/WebID

Clarity RDF has a well-defined way how to convey semantics through the use of RDF vocabularies and ontologies, the workings of which are described in 4.1.3.2. RDF vocalabularies and ontologies make thorough data modelling feasible, which increases the fidelity and clarity of the way representations of RDF resources are modelled.

Documentation Linked data is self-describing data. Since the *"consumers of Linked Data do not have the luxury of talking to a database administrator who could help them understand a schema"* [59], all the information necessary to interpret the data, including RDF vocabularies and ontologies used by the data, should be stored on the Web and should be possible to retrieve via the mechanism of dereferencing by issuing HTTP GET requests and recursive following of links.

While the representations of resources should be self-documenting, there is no such requirement on the linked data URIs. URIs may be opaque since *"the Web is designed so that agents communicate resource information state through representations, not identifiers"* [60].

4.3.3 Linked Open Data in the Public Sector

Having reviewed the theoretical foundations for technical openness and data quality of linked data, this section turns to the ways in which linked open data is used in practice in the public sector. Contrary to the popular belief, linked open data is not any more confined to the research institutes producing pilots and prototypes. It is used in practice, and the public sector is one of the central areas in which linked data is being adopted.

To find out about the role of public sector data in the ever-increasing web of data, the Linked Open Data Cloud[41] diagram may be consulted. This diagram depicts the connections between the existing linked data sources that are published under the terms of an open licence. Progressive changes made to this diagram over time illustrate the growth of the web of data that now contains more than a billion triples.[42] The cloud is partitioned in broad subject categories that include a category for "government". According to the *State of the LOD Cloud* [17] survey from September 2011 the datasets in this category represented 42.09% of triples in the cloud. However, these datasets accounted only for 3.84% of outbound links to external datasets.

The Linked Open Data Cloud features datasets from the public sector of a number of countries. The U.S. is represented by their pioneering Data.gov[43] project started by the Obama administration in May 2009. In the United Kingdom,

[41] http://lod-cloud.net

[42] According to LODStats (http://stats.lod2.eu/), the available datasets constituting the Linked Open Data Cloud contained 1,174,474,890 triples as of April 25^{th}, 2012.

[43] http://data.gov

the adoption of linked open data in the public sector was kick-started by research projects, such as AKTivePSI [3] at the University of Southampton. The research activity quickly developed into an official part of work of the public sector and gave rise to Data.gov.uk,[44] one of the most comprehensive and progressive government data catalogues to-date. Aside from the other countries, initial experiments with linked open data for the data produced in the public sector are also conducted in the Czech Republic by an un-official initiative OpenData.cz.[45]

The thriving growth of linked open data activities in the public sector pointed to a need for coordination and development of standards and best practices. The W3C has taken the lead and established the Government Linked Data Working Group[46] to help guide the adoption of linked open data in the public sector. The group is scheduled to run until 2013, but it already published several recommendations, such as the Cookbook for open government linked data [59].

4.4 Summary

This chapter investigated whether the linked data publishing model is in accordance with the publishing practices of open data. It was discussed how linked data reaches compliance with the requirements on technical openness and data quality posited by the principles of open data. After an introduction of the underlying technologies borrowed from the semantic web stack, the agreed-on principles establishing the rules for linked data were presented. A review of the techniques for publishing linked data revealed a close similarity to the features deemed essential for open data. The final section dedicated to linked open data went through the requisites for open data and suggested ways how linked data fulfils them, concluding with an overview of the implementations of linked open data in the public sector.

[44] http://data.gov.uk

[45] http://opendata.cz/

[46] http://www.w3.org/2011/gld/wiki/Main_Page

5 Impact and Challenges

Current network society is a complex system. Outcomes of introducing change in such a system cannot be controlled by any single body. The complexity of this system implies that no one is able to predict the effects of system-wide changes that propagate through its network.

Applying linked open data for public sector information is an example of that kind of a change. The topic of open data is still rather new and the existing research covering it, as noted by Jonathan Gray from the Open Knowledge Foundation, is a combination of evidence and expectations [46]. Even though the decisions behind this change are motivated by the expected positive results, its impact is hard to predict. A number of challenges for adoption of linked open data in the public sector may be identified, some of which can be effectively addressed in public policy, while others wait for further research to provide at least provisional solutions.

5.1 Impact

Rufus Pollock from the Open Knowledge Foundation argues that *"open data is a means to an end, not an end in itself"* [87]. Open data alone has no impact, as its impact is triggered by its use. Thus, no impact is guaranteed by the intrinsic properties of open data.

Open data discourse contains a vision that promises a better society in the offing. It is a vision that stems from the belief in transformative effects of open data principles and information technologies that are entrusted to deliver this vision. However, this vision will not be put into practice by releasing open data. Its the use of open data that puts the transformation into motion.

Rhetoric of open data advocates emphasizes the positive side of open access to public sector data. Moreover, it is often presented as an asymptomatic and strictly apolitical issue. However, it would be short-sighted to assume it is a neutral, technological change. We need to admit that there are both positive and negative impacts of open data, bringing both benefits and repercussions.

Distinguishing between the target of open data impacts, a rough categorization can be drawn classifying impacts either as internal, if they affect data producers, or as external, if they influence others.

5.1.1 Internal Impact

Internal impact, which affects the producers of public sector data, is based largely on data about the public sector. The data describing the public sector is a record of its activity that may be used and scrutinized to improve the workings of the public sector. An open and better performing public sector is among the key objectives of the open data movement. Ultimately, open data paves the way to an open and more efficient government.

Open data disrupts existing workflows that are established in the public sector. It subjects the public sector to a greater transparency, which enables to held civil servants accountable, and establishes conditions under which the public sector may function in a more efficient way.

5.1.1.1 Transparency

Transparency of the public sector reflects the ability of the public to see what is going on. David Weinberger declares that transparency is the new objectivity [112], a change that he claims to stem from the transformation of the currect knowledge ecosystem to one that is inherently network-based. Transparency replaces the role of the long-discredited objectivity in that aspects that it is used as a source of veracity and reliability [111].

Transparency serves for fraud prevention. It puts the public sector under a peer pressure based on the fact that anybody can inspect the its public proceedings. The peer supervision makes it more difficult for civil servants to profit from the control they have and abuse of the powers vested in them. By increasing the risks of exposure of venal activities, it lowers the systemic corruption [11, p. 9]. In effect, members of the public may hold civil servants accountable for corruption, illegal takeover of subsidies, or plain budgetary waste [16, p. 80].

An illustrative example of the self-regulating effects of transparency was presented in [65, p. 110]. In 1997, restaurants in the Los Angeles county were ordered to post highly visible letter grades on their front windows. The grades (A, B, C) were based on the results of County Department of Health Services inspections probing hygiene maintenance in the restaurants. The ready availability of evidence on insanitary practices in food handling made it easier for people to make better choices about restaurants and helped them to avoid restaurants that were deemed unsafe to eat at. The introduction of this policy proved to have a significant impact both on the restaurants and their customers. Revenues at C-grade restaurants dropped, while those of A-grade restaurants increased, leading over time to a growth of the number of cleanly restaurants and a steep decline of the poorly performing ones. The policy also improved health conditions of the restaurants' customers, with a decrease of hospitalizations caused by food-borne illnesses from 20% to 13%.

Transparency has an ambiguous impact on trust in the public sector. While

there is a positive impression of stronger control over the public sector, at the same time more failures are identified, which chips away at the trust in public affairs. Furthermore, transparency makes citizens aware of how vulnerable to manipulation the public sector data is.

Open data shapes the reality it measures [18, p. 3]. When communicating, the sender conveying information modifies its content based on the perceived context of communication. Evaluation of the way of communication, the expected audience, and other circumstances factored into the communication context impacts what messages are sent. Open data establishes a new context with a wider and less defined range of potential recipients and a different set of expectations about the effect of communicated data. Such re-contextualization may affect what gets released and in what form. Data may be distorted in a direction so that it supports only the interpretations data producers expect [61]. As a result, some data may end up withheld from the public, while other data may turn out to be misrepresenting of the phenomena it bears witness to. At the same time, the change brought about by the obligation to disclose data may have positive consequences by forcing public bodies *"to rethink, reorganize and streamline their delivery before going online"* [51, p. 448].

As the control is ultimately in the hands of civil servants, data disclosure may be shaped as required by various interest groups, including politicians or lobbyists. It illuminates the fact that there is no direct causation between open data and open government. *"A government can be an 'open government,' in the sense of being transparent, even if it does not embrace new technology"* [119, p. 2]. Only politically important and sensitive disclosures take government further on its way to open government. *"A government can provide 'open data' on politically neutral topics even as it remains deeply opaque and unaccountable"* [119, p. 2].

This reflects what Ellen Miller from the Sunlight Foundation calls the danger of a mere "transparency theater."[1] This is nothing new in the politics. For instance, questions that politicians get asked may be moderated to include only those that are not sensitive and do not require the interviewee to disclose any delicate facts.

It also indicates that there is a limit to transparency, a limit that Joshua Tauberer entitled the "Wonderlich Transparency Paradox" [105]. It is named after John Wonderlich from the Sunlight Foundation that once wrote that *"How ever far back in the process you require public scrutiny, the real negotiations [...] will continue fervently to exactly that point"* [113]. Some parts of the processes in the public sector are exempted from disclosure to provide a "space to think" [16, p. 74]. However, this paradox shows that no matter how thourough and deep the transparency of the public sector is, the real decision-making processes will always have a chance to elude what is recorded and exposed for public scrutiny.

Everything may be abused and transparency is no different. For example, releasing data about how well are civil servants paid may be used to identify

[1] https://twitter.com/#!/EllnMllr/status/182629508552200192

targets for bribery. Disclosing salaries of politicians helps lobbyists to find a low-paid politician who is an easier target for corruption. A difficult question is also to ask whether terrorist watch list should be made open [65, p. 4].

These examples showcase the unintended consequences of opening data. What these concerns illustrate is that transparency is obviously not a panacea and it would be naïve to think it is. Open data is not an end to itself and transparency by itself is an input, not an output [99].

5.1.1.2 Accountability

Transparency feeds into accountability. *"In the world of big data correlations surface almost by themselves. Access to data creates a culture of accountability"* [22]. Open data enables to hold politicians accountable by comparing their promises with data showing how are their promises put into practice. For example, unfavourable audit results based on open data may cause a politician not being reelected.

Public scrutiny of governmental data may reveal fraud or abuse of public funds. Given the availability of public data everyone may check out, we may see a rise of the so-called "armchair auditing." In the same way, it improves the function of "watchdog" institutions, such as non-governmental organizations dedicated to overseeing government transparency. In this way, open data increases civic engagement leading to a more participatory democracy and better democratic control.

Open data enables to apply crowdsourcing to monitor institutions and their performance, which is described in the data. Rufus Pollock illustrated the opportunities of leveraging citizen feedback by saying that *"to many eyes all anomalies are noticeable,"* in which he paraphrased the quote *"given enough eyeballs, all bugs are shallow"* by Linus Torvalds. Accordingly, releasing data to the public allows to get the data verified or inspected for quality for free.

5.1.1.3 Efficiency

The public sector itself is the primary user of public sector data. Open access to public data thus impacts the way the public sector operates. While the initial costs of opening up data may turn out to be significant, adopting open data promises to deliver cost savings in the long run, enabling the public bodies to operate more efficiently. *"There is a body of evidence which suggests that proactive disclosure encourages better information management and hence improves a public authority's internal information flows"* [16, p. 69]. For instance, open data produces cost savings on cheaper information provision and efficient development of applications providing services to citizens.

For information provision, similarly to health services, prevention is cheaper than therapy [40]. Prevention via proactive disclosure is presumed to be more

cost-efficient than therapy via acting on the demand of freedom of information requests [49, p. 25]. Open data saves the effort spent on responding freedom of information requests by providing the requested data in advance. In this way, the effort of providing data is expended only once, instead of repeating it due to the requests for the same data. Although the initial set-up overhead for open data may be higher, it is supposed to lower the per-interaction overhead.

Open data promotes a new way of information management that may streamline the data handling procedures and curb unnecessary expenditures. By elimination of the costs associated with access to public sector data the adoption of open data removes the expenses on data acquisition from public sector bodies selling their data. In effect, a better interagency coordination is established, which lessens administrative friction. Given the reduced workload, it may lead to destruction of some clerical jobs [40], which will produce savings on labour costs.

A common argument in favour of open data is based on the observation that the public sector is not capable of creating applications providing services to citizens in a cost-efficient way. Commissioning software for the public sector must pass through the protracted process of public procurement. Such procedure is slow to respond to users' demands and the resulting applications may end up being costly. With openly available public sector data, the public sector is no longer the only producer that can deliver applications based on the data. Third parties may take the data a produce applications on their own, substituting the applications subsidized by the public sector. This is how a more cost-efficient means of production of applications may be devised.

The way in which open data makes efficiency of the public sector better is not limited to monenatary savings. The internal impact of open data encompasses that the data quality may be improved by harnessing the feedback from citizens. It may also inform the way the public sector is governed through evidence-based policies.

Opening data enables anybody to inspect it. Feedback from users probing the data puts a pressure on the public sector to improve the data quality. Better quality data enables better quality service delivery, improving the pursue of public task on many levels, such as better responsiveness to citizen feedback. Based on user feedback, collection of less used datasets may be discontinued, leading to a more responsive and user-oriented data disclosure.

Quality of data influences the quality of the policy that is based upon it [80]. It may become a source for a more efficient, evidence-based policy. Public policies may be improved by considering data as an input, as an evidence of the phenomena to be policed, and should be made with publicly available data [80, p. 384], an empiciral data that is open to public scrutiny [96, p. 4], in order to keep the policy creators accountable.

5.1.2 External Impact

External impact of open data affects the demand side of open data. It results chiefly from availability of data about the environment governed by the public sector bodies releasing the data.

A recognized issue with the open data movement is that it lacks focus on the demand side of data. It suffers from unrealistic expectations brought about with the pervasive tendency to pay attention solely to the supply side, which is coupled with a lack of consideration of how the data would be used after its release [74, p. 1]. The public sector should abandon this ill-considered model and instead adopt a user-centric model for data disclosure.

Close attention to the demand side is needed because the power of open data is not in itself, it resides in the ways it can empower people that use the data. Open data empowers citizens to make better decisions. For example, access to crime data may assist city dwellers in finding the safest route home. Information about wheelchair access to public transportation may help persons with reduced mobility to arrange their city transport better.

The effects of open data that impact users of data are covered in the following sections. Among the effects that are discussed is the phenomenon of disintermediation that allows users of data to by-pass intermediaries and the ways in which open data enables citizens to participate in public affairs. Influences of open data on two specific domains are considered. The availability of public sector data is a new potential for the economy. For journalism open data brings about a change that makes it become more data-driven.

5.1.2.1 Disintermediation

Who draws and controls the maps controls how other people see the world [33]. Who interprets data from the public sector controls how other people see the things described in the data. By releasing raw open data the public sector also releases its total control over the interfaces in which the data is presented. In this way, the interpretive dominance of the public sector data is abolished and it no longer controls the way how citizens should see the world described in the data [8]. Civil servants perceive this as a loss of control over the released data, but in fact, it is only a loss of control over interfaces in which the data is presented.

Providing raw data is an example of disintermediation. It reduces the frictions and inherent cognitive biases that come with interpretations by intermediaries. It allows users to skip the intermediaries that stand between them and access to raw data. For example, both civil servants producing reports based on primary data and journalists transforming data into narratives conveyed in articles serve as intermediaries that affect how the public perceives public sector data.

Depending on the type of use mediation may be either a barrier or a help. It is a barrier for those that want to access raw data to interpret them themselves.

However, common perception has it that too few people are interested in raw data [49, p. 71]. Yet one should not make such generalizations as there is evidence that suggests otherwise. For example, after the release of data from the Norwegian meteorological institute,[2] the institute registered more data downloads (14.8 million) than page views (4.5 million).[3] In general, it is the case that raw data receives relatively few downloads, yet access to raw data is vital to build new applications on top of the data.

Disintermediation creates a demand for reintermediation. Mediation helps users that need to get user-friendly translations of data in order to reach understanding. Applications mediating data in ways that are accessible and compelling, such as visualizations, may attract a lot of attention proving the demand for public sector data. For instance, this has happened in the case of the UK crime statistics, the visualization of which crashed under the weight of 18 million requests per hour at the time it was released [107].

5.1.2.2 Participation

Open data enables better interaction between citizens and governments through the Web [2]. It redresses the information asymmetry between the public sector and citizens [42] by advocating that everyone should have the same conditions for use of public sector data as the public body from which the data originates. Sharing public data facilitates universal participation since no one is excluded from reusing and redistributing open data [26].

Open data opens the possibility of citizen self-service. It makes the public more self-reliant, which reduces the need for government regulation [104]. It enables to tap into the cognitive surplus and improve public services with the crowdsourced work of the public. One of the main benefits of open data consists in third-party developed citizen services [69, p. 40]. Citizens may thus become more involved in public affairs, which ultimately leads to a more participatory democracy.

5.1.2.3 Business Potential

There is no direct return on investment on open data. As a matter of fact, economic impact of releasing open data is difficult, if not impossible, to anticipate and quantify beforehand, prior to the publication date. The causal chain connecting open data as a cause with its economic effects is particularly unrealiable. However, it seems to be feasible to recount the effect on business after the moment data is made accessible. For instance, an analyst may consider the number of uses by

[2] http://www.yr.no/

[3] These numbers were given by Anton Eliassen, the institute's director, during the first plenary on the revised public sector information directive at the ePSI Platform Conference 2012. http://vimeo.com/38804207

businesses comparing how it changed before and after the data was opened [85]. Accordingly, the economic value of open data can be rather considered as indirect.

Given the way open data affects economy, estimates of the market size for public sector data are based on methodologies that are insufficient to come up with accurate numbers. For example, most of the studies evaluating economic impact of opening up data in the public sector were based on extrapolations from research conducted on a smaller scale. In his study for the European Commission, Graham Vickery assessed the aggregate volume of the direct and indirect economic impacts of opening public sector information in the EU member countries to be EUR 140 billion annually [110, p. 4]. In contrast with this number, estimates of the direct revenue based on selling public sector information were much lower, and Vickery quantified it to EUR 1.4 billion [110, p. 5].

Open data opens new opportunities for private businesses. It allows new business models to appear, including crowdsourced administration of public property by services such as FixMyStreet.[4] Another example of a business that is based on public sector data is BrightScope[5] that delivers financial information for investors. An area that may benefit the most from availability of public sector data are location-based services. The EU Directive on the reuse of public sector information was reported to have the strongest impact on the growth of the market of geospatial data that is essential for such services to be operated [110, p. 20].

The opportunities offered by open data are particularly important for small and medium enterprises. These businesses are a prime target for reuse of open data since they usually cannot afford to pay the charges to public bodies for data that is not open. Stimulation of economic activities may result in new jobs being created. Availability of public data may give rise to a whole new sector of "independent advisers", that add value to the data by making it more digestible to citizens [53]. More businesses eventually generate more tax revenue, which ultimately promises to return the investment in open data back to the budget from which the public sector is funded.

Open data fosters product and service innovation. It affects especially the areas of forecasting, prediction, and optimization. For example, European Union makes its official documents available in all languages of the EU member states. This multilingual corpus is used as a training set for machine translation algorithms in Google Translate leading to an improvement in quality of its service [26].

At the same time, open data disrupts existing business models that are based on exclusive arrangements for data provision by public sector bodies to companies. This is how businesses that thrive on barriers to access to public data are made obsolete. Open data weeds out companies that hoard public data for their benefit and establishes an environment, in which all businesses have an equal opportunity to reuse public sector data for their commercial interests.

[4] http://www.fixmystreet.com/

[5] http://www.brightscope.com/

5.1.2.4 Data-driven Journalism

The availability of data and data processing tools gives birth to a new paradigm in journalism that is commonly referred to as data-driven journalism. It refers to the practice of basing journalistic articles on hard data, which allows to back up claims with well-founded evidence.

Unlike in journalism that is driven by data, unverified claims abound in traditional journalistic practice. To address this deficiency, data-driven journalism may employ open data sources to cross-verify the claims. Data triangulation combining disparate sources may establish validity of the verified claims.

If data-driven journalists strive to draw closer to objectivity, they need to share their sources to achieve transparency. Sharing the underlying data is an imperative of data-driven journalism, so that others can see what lead to insights conveyd in articles. In the light of such transparency, claims made by journalists may be verified by third parties and trust may be established.

The best known examples of data-driven journalism include the Guardian's Datablog[6] or Pro Publica.[7]

5.2 Challenges

Open data not only opens new opportunities, it also opens new challenges. These challenges point to the limits of openness and to shortcomings of the approaches used to put linked open data in practice in the public sector.

The top 10 barriers and potential risks for adoption of open data in the public sector, which were compiled by Noor Huijboom and Tijs van den Broek [58, p. 7], comprise of the following.

- closed government culture
- privacy legislation
- limited quality of data
- limited user-friendliness/information overload
- lack of standardisation of open data policy
- security threats
- existing charging models
- uncertain economic impact

[6] http://www.guardian.co.uk/news/datablog

[7] http://www.propublica.org/

- digital divide
- network overload

Some of these challenges will be discussed in detail in the following parts of the thesis. In particular, this section will cover the difficulties that may be encountered during implementation of linked open data, information overload and the problems of scalable processing of large, heterogeneous datasets, usability of raw data, issues for protection of personal data, deficiencies in data quality, adverse effects of open data on trust in the public sector, and finally the unresolved question of opening data obtained via public procurement.

5.2.1 Implementation

Data publishers may perceive adoption of linked open data to have daunting entry barriers. In particular, they are aware of the high demands on expertise for publishing linked data, which is esteemed to have a steep learning curve. Linked data publishing model poses requirements that may seem to be difficult to meet. The Frequently Observed Problems on the Web of Data [55] testify to that.

Therefore, *"it is vital to follow a realistic, practical and inexpensive approach"* [3]. Fortunately, linked data facilitates an incremental, evolutionary information management. Its deployment may follow a step by step approach, adopting iterative development for continuous improvement. For example, before a switch of the database technology linked data publishers could start by caching given legacy databases into triple stores. Another way how to cushion the demands of linked data adoption is to minimise their ontological commitment by creating small ontologies that may be gradually linked together.

Two implementation challenges collocated with the adoption of linked open data in the public sector will be dealt with in detail; resistance to change in the public sector and maturity of the linked data technology stack.

5.2.1.1 Resistance to Change

Rhetoric of open data supporters puts an emphasis on bureaucracy as a major barrier to opening data in the public sector. There is a tendency to frame the politics of access to data as a struggle between the public sector, that has an inbreed attachment to secrecy, and members of the public, which are depicted rather as individuals than groups [74, p. 7].

While this view seems to be biased, the institutional inertia may pose a challenge to adoption of open data, which may require a *"cultural change in the public sector"* [45]. The transition from the status quo may be significantly hindered by the established culture in the public administration. *"A major impediment is an entrenched closed culture in many government organisations as a result of the fear of disclosing government failures and provoking political escalation and*

public outcry" [108]. The intangible problem of the closed mindset prevailing in the public sector proves to be difficult to resolve. And so, in many ways, the adoption of open data *"isn't a hardware retirement issue, it's an employee retirement one"* [28].

Resistance to change is not the only barrier hindering in the adoption of open data. A hurdle that is commonly encountered by open data advocates is that civil servants perceive open data as an additional workload that lacks clear justification [49, p. 70]. Unlike citizens that are allowed to do everything that is not prohibited, public servants are allowed to do only what law and policies order them to do. Voluntary adoption of open data at the lower levels of public administration is thus highly unlikely. It requires a policy to push open data through.

However, it might be for the existing policies that the change is made difficult. In general, the public sector is a subject to special obstacles that impede adoption of new technologies. For example, the combination of strict data handling procedures and constricted possibilities due to limited budget resources may effectively stop any technological change [2]. That is why there must by a strong commitment to open data on the upper levels of the public sector in order to put through the necessary amendments to existing data handling policies.

5.2.1.2 Technology Maturity

Semantic web technologies underlying linked data were for a long time thought of as not being ready for adoption in the enterprise settings and in the public sector. In 2010, linked data technology stack was not perceived to be ready for large-scale adoption in the public sector. John Sheridan reports three key things missing [98]:

- Repeatable design patterns
- Supportive tools
- Commoditization of linked data APIs

At that time, standards were mature enough, but their translation to repeatable design patterns applicable in practice was lacking. This has changed since. Several sources recommend established design patterns (e.g., [27], [52], [59]), supportive tools were developed and packaged (e.g., LOD2 Stack[8]), and frameworks for developing custom APIs based on linked data were created (e.g., Linked Data API mentioned in 4.3.1.1). Linked data has matured progressively in the recent years and so it may be argued that it is ready to be implemented at the level of the public sector.

[8] http://stack.lod2.eu/

5.2.2 Information Overload

As more and more data is released in the open there is a growing danger that irrelevant data might flood the data that is important [40]. Only few of the available datasets contain "actionable" information and there is no effective filtering mechanism to track them down. With open data *"we have so many facts at such ready disposal that they lose their ability to nail conclusions down, because there are always other facts supporting other interpretations"* [112].

The sheer volume of the existing open data makes it difficult to comprehend. At such scale there is a need for tools that make the large amounts of data intelligible. Edd Dumbill writes that *"big data may be big. But if it's not fast, it's unintelligible"* [28].

While human processing does not scale, machine processing does. Thus, the challenge of information overload highlights the need for machine-readable data. Big, yet sufficiently structured data may be automatically pre-processed and filtered to "small data" that people can manage to work with. For example, linked data may be effectively filtered with precise SPARQL queries harnessing its rich structure.

Scaling the processing of large amounts of machine-readable data with well-defined structure may be considered solved. However, the current challenge is to deal with the heterogeneity of data from different sources.

5.2.2.1 Heterogeneity

Not only is there a perceived information overload, there is also an overload of different and incompatible ways of representing information. What we have built out of different data formats or modelling approaches seems to be the proverbial "Tower of Babel". In this state of affairs, the data available on the Web constitutes a highly dimensional, heterogeneous data space.

Nonetheless, it is in managing heterogeneous data sources where linked data excels. Linking may be considered as a lightweight, pay-as-you-go approach to intergration of disparate datasets [52]. Semantic web technologies also address the intrinsic heterogeneity in data sources by providing means to model varying levels of formality, quality, and completeness [97, p. 851].

5.2.2.2 Comparability

A key quality of data that suffers from heterogeneity is comparability. According to the SDMX content-oriented guidelines comparability is defined as *"the extent to which differences between statistics can be attributed to differences between the true values of the statistical characteristics"*. [94, p. 13] It is a quality of data that represents the extent to which the differences in data can be attributed to differences in the measured phenomena.

Improving comparability of data hence means minimizing unwanted interferences that skew the data. Influences leading to distorsion of data may originate from differences in schemata, differing conceptualizations of domains described in the data, or incompatible data handling procedures. Elimination of such influences leads to maximization of evidence in data, which reflects more directly on the observed phenomena.

The importance of comparability surfaces especially in data analysis tasks. Insights yielded from analyses then feed into decision support and policy making. Comparability also supports transparency of public sector data because it clears the view of public administration. It supports easier audits of public sector bodies due to the possibility to abstract from the ways used to collect data. On the other hand, incomparable data corrupts monitoring of public sector bodies and imprecise monitoring thus leaves an ample space for systemic inefficiencies and potential corruption.

The publication model of linked data has in-built comparability features, which come from the requirement for using common, shared standards. RDF provides a commensurate structure through its data model that linked data is required to conform to. The emphasis on reuse of shared conceptualizations, such as RDF vocabularies, ontologies, and reference datasets, provides for comparable data content.

In the network of linked data the "bandwagon" effect increases the probability of adoption of a set of core reference datasets, which further reinforces the positive feedback loop. Core reference data may be used to link other datasets to enhance their value. Such datasets attract most in-bound links, which leads to emergence of "linking hubs". In this case, these de facto reference datasets derive their status from their highly reusable content. An example of this type of datasets is DBpedia[9], which provides machine-readable data based on Wikipedia. Its prime condition may be illustrated by the Linked Open Data Cloud, in the center of which it is prominently positioned, indicating the high number of datasets linking to it.

In contrast to these datasets, traditional reference sources are established through the authority of their publishers, which is reflected in policies that prescribe to use such datasets. Datasets of this type include knowledge organization systems, such as classifications or code lists, that offer shared conceptualizations of particular domains. For instance, a prototypical example of an essential reference dataset is the *International System of Units* that is a source of shared units of measurement. In contrast with the linking hubs of linked data, traditional reference datasets are, for the most part, not available in RDF and therefore not linkable.

The effect of using both kinds of reference data is the same. The conceptualizations they construct offer reference concepts that make data referring to them

[9] http://dbpedia.org/

comparable. A trivial example to illustrate this point may be the use of the same units of measurement, which enables to sort data in an expected order.

Data might need to be converted prior to comparison with other datasets. In this case, there is a need for comparability on the level of the data the incomparable datasets refer to. Linked data makes this possible through linking; the same technology it applies to data integration. With the techniques, such as ontology alignment, mappings between reference datasets may be established to serve as proxies for the purpose of data comparison. Ultimately, machine-readable relationships in linked data make it outperform other ways of representing data when it comes to the ability to draw comparisons.

5.2.3 Usability

Considering usability as a property of interfaces, raw data provides a difficult one. Largely, data is too unwieldy to be used by most people. For example, 50% of the respondents in the Socrata's open data study said that the data was unusable [100]. Alternatively, poor usability may be correlated with the low level of use most open data sources receive.

The requirements on usability of open data reviewed in 3.2.3.2 prove to be difficult to satisfy. The usability barrier may be especially high when dealing with linked open data as was reported in the section 4.3.2.2. Yet it is important not to compromise the generative potential of open data to low usability of the underlying technologies.

The challenge of usability requires data producers to refocus on the view of user-centric perspective. This section highlights the increased need for data literacy, which is necessary for interacting with open data, and warns of the dangers of incorrect interpretations drawn from data.

5.2.3.1 Data Literacy

Even though open data bridges the data divide between the public sector and members of the public, it might be introducing a new data divide that separates those with resources to make use of the data and those who do not. Despite the fact that open data virtually eliminates the cost of data acquisition, the cost of use remains *"sufficiently high to compromise the political impact of open data"* [74, p. 11].

An oft-cited quote attributed to Francis Bacon claims that *"knowledge is power"*. If data is a source of knowledge, then opening it up creates a shift in access to a source of power. However, equal access to data does not imply equal use, nor equal empowerment, as transforming data into power requires not only access. Letting aside the concerns of unequal access addressed by the agenda of the digital divide, while the principles of open data lead to the removal of barriers to access, they do not remove all barriers to use. In this respect, it is vitally

important to distinguish between the "opportunity" and the actual "realization" of use of open data [48]. Even though everyone may have equal opportunities to access and use open data, only someone is able to achieve "effective use" [48]. In the light of this assertion, open data empowers only the already empowered; those that have access to technologies and computer skills that are necessary to make use of the data.

The belief in transformative potential of open data is based on optimistic assumptions about the citizens' data literacy. The technocratic perspective with which open data principles are drafted takes high level of skills necessary for working with data for granted. Thus, the open data initiatives are in a way exclusive as they are limited mostly to technically inclined citizens [15, p. 268].

The minimalist role of the public sector, withdrawn into the background to serve as a platform, proceeds of the supposition that members of the society have all the necessary ingredients to make effective use of open government data, such as high level of information processing capabilities [42]. Even though ICT penetration and internet connectivity may be sufficient to access open data, it is not enough to make use of it. What is also needed are the abilities to process and interpret the data. However, open data released in a raw form may not be easily digestible without a substantial proficiency in data processing. Therefore, it should not be underestimated that users are required to possess technical expertise to process the data.

The bottom line is that access to data may in fact increase the asymmetry in society. If all interest groups have equal access to public sector information, then we can expect that the better organized and well-equipped groups to make better use of it [99]. The asymmetry may stem from the fact, that the interest groups that are able to take advantage of the newly released information will prosper at the expense of groups that cannot do that.

On the other hand, this type of unequality is in a sense natural. Such state of affairs should not be considered as a final one, but rather as a starting point. David Eaves compares the challenge of increasing data literacy to increasing literacy in libraries and reminds us that *"we didn't build libraries for an already literate citizenry. We built libraries to help citizens become literate"* [29]. In the same way, we do not publish open data expecting everyone will be able to use it. The data are released since access is a necessary prerequisite for use. Direct access to data by the empowered, technically-skilled infomediaries may become a basis for an indirect access for many more [105]. Coming from this perspective, the most effective uses of open data can be thought of as those that let others make effective use of the data.

5.2.3.2 Misinterpretation

Another argument pointing at the potential risks in disclosure of public data was presented by Lawrence Lessig in an article titled *Against transparency* [67], in

which he draws attention to adverse effects of misinterpretation of public data. He highlights the issues that arise when monopoly on interpretation is removed and members of the public are provided with raw, uninterpreted data [23, p. 2]. Disintermediation causes decontextualization of public sector data that may lead to highly divergent interpretations of the same data [61]. Such change may be perceived as a loss of control the civil servants used to have. Instead of an "official" interpretation of open data this would potentially lead to a plurality "competing" and possibly conflicting interpretations, some of which may be driven by malicious interests.

Lessig claims, paying respect to the alleged shortening attention spans of members of the public, that it is easier to come up with an incorrect judgement based on public data than one that is based on solid understanding [67]. The ability to correctly interpret data is largely prevalent only among people with suffiecient expertise and data literacy skills. Moreover, Archon Fung and David Weil argue that the way open data is disclosed is conducive to pessimistic view of the public sector. They claim that *"the systems of open government that we're building - structures that facilitate citizens' social and political judgments - are much more disposed to seeing the glass of government as half or even one-quarter empty, rather than mostly full"* [65, p. 107]. Such conditions may also make users of data susceptible to apophenia, a phenomenon of seeing patterns that actually do not exist [18, p. 2]. In fact, Lessig writes, encountered with the vast amounts of available public data, ignorance is a rational investment of attention [67]. Without a significant time investment and data literacy skills people will usually come to shallow and premature conclusions based on their examination of public data. Unfounded conclusions may be quickly adopted and spread by the media, which may cause significant harm of reputation of public sector bodies, civil servants, or politicians, until these assertions are re-examined and proven to be false. For example, unverified oversimplifications may be yielded from public data to support political campaigns. Open data can be misused for skewed interpretations supporting political actions, casting suspicion on public image of politicians that are the target of discreditation campaigns.

Misinterpretations may increase distrust in the public sector. Thus, Lessig makes the case for disclosing a limited amounts of public data prone to misinterpretation [67]. Even though, he is not completely opposing the transparency initiatives, he warns that careful considerations should be given when releasing sensitive information that may be misused for defamation.

Unrestricted access to communication channels provided by new media gives strong voice to all competing interpretations, unhindered by the filtering mechanisms of traditional publishing. This state of affairs results in unfounded claims and rumours to amplify and spread with an impact that was previously impossible to achieve, causing harm to personal reputations and the public image of government. Fortunately, the self-repairing properties of communication networks eventually lead to the rebuttal of misinformation. The openness of public data

thus brings not only a greater control of the public sector, but indirectly also a better control of unproven claims.

5.2.4 Privacy

In the pursuit of the public task public sector bodies collect personal data as well. Such data does not fall under the scope of open data. Principles of open data explicitly exclude personal data from being released and suppose it to be left closed in well-secured databases.

A complaint that is heard with regard to privacy is that the public sector collects more personal information than the minimum it needs. An example where public data collection posed a potential privacy breach comes from Finland [26]. A Finnish travel system logged all instances when a travel card was scanned by reader machine on different public transport lines. Since travel cards can be traced to individual persons, in this arrangement the travel system had location data for a large number of people, which was perceived as a violation of privacy. Ultimately, based on the data protection legislation, the travel card data was ceased to be collected.

However, in most cases personal data is not collected at an excessive rate and is governed by an access regime that is strictly limited to authorized users from the public sector to prevent accidental leaks of private data. In line with this observation, Marco Fioretti notes that privacy issues of open data have almost always been a non-issue [40].

Nonetheless, a new privacy risk is being recognized in the danger of statistical re-identification. This privacy threat is inflicted by the availability of large amounts of machine-readable data, that contains indirect personal identifiers, and the technologies allowing to combine it.

So far, privacy was guaranteed by the "practical obscurity" [97, p. 867]. It existed chiefly due to the difficulty of obtaining and combining data. In many cases, personal data was not logged down at all. Under such conditions, the right to privacy was akin to the right to be forgotten [40]. However, this assumption loses ground when confronted with the ever-increasing amount of data that is currently being recorded and stored.

Data anonymization that is based on removal of direct identifiers, such as identity card numbers, is insufficient on its own. A subject may be identified and linked to sensitive information through a combination of indirect identifiers [117, p. 8]. Indirect identifier is a data item that narrows down the set of persons who might be described by the data. An example of an indirect identifier that works this way is gender. When enough indirect identifiers are combined, they may narrow down the set of subjects they might identify to a single person.

There are established techniques for protecting personal privacy in data by limiting the risks of re-identification by statistical methods. Chris Yiu lists several

of them, most of which have adverse impact on data quality and openness [118, p. 26].

- *Access and query control*, e.g., filtering and limiting size of query results to samples
- *Anonymisation*, or *deidentification*, such as stripping personal information from data
- *Obfuscation*, that may, for example, reduce precision in data by replacing values with ranges
- *Perturbation*, introducing random errors into data
- *Pseudonymisation*, including replacing persons' names with identifiers

Fortunately, both direct and indirect personal identifiers are rare in public sector data. Most of the data tracked by the public sector consists of non-identifiers. Moreover, the data is usually available in aggregated forms and not as microdata that results directly from data collection. Therefore, in most cases, data quality and openness do not need to be compromised due to the requirements of privacy protection.

5.2.5 Data Quality

Data quality is required for data that may be depended upon. Yet public sector data may be mired in errors and suffer from unintentional omissions that may markedly decrease usability of data. For example, Michael Daconta [20] identified ten common types of mistakes in datasets at the U.S. data portal Data.gov.[10]

- *Ommission errors* violating data completeness, missing metadata definitions, using code without providing code lists
- *Formatting errors* violating data consistency, syntax errors not fulfilling requirements of the employed data formats' specifications
- *Accuracy errors* violating correctness, errors breaking range limitations
- *Incorrectly labelled records* violating correctness, for example, some datasets misnamed as CSV even though they are just dumps from Excel files that do not meet the standards established in the specification the CSV data format
- *Access errors* referring to incorrect metadata descriptions, for example, not linking to the content described by the link's label

[10] http://www.data.gov/

- *Poorly structured data* caused by improper selection of data format, using formats that are innapropriate for the expected uses of data
- *Non-normalized data* violating the principle of normalization, which attempt to reduce redundant data by, e.g., removing duplicates
- *Raw database dumps* violating relevance and providing raw database dumps that are hard to interpret and use correctly
- *Inflation of counts* that is a metadata quality issue having an adverse impact on usability, for instance, when datasets pertaining to the same phenomena are not properly grouped and thus difficult to find
- *Inconsistent data granularity* violating expected quality of metadata, such that datasets use widely varying levels of data granularity without their explicit specification

Linked data principles impose a rigour to data that may improve its consistency and quality. At the same time, linked data is more susceptible to corruption caused by "link rot" and the issues that arise when links no longer resolve.[11] The reliance on URI makes it even more important for linked data to adopt URIs that are stable and persistent.

5.2.6 Trust

Transparency brought about by the adoption of open data affects the trust in the public sector. Current governments experience a crisis of legitimacy[65, p. 58] and lack the trust of citizens. Improved visibility of the workings of public sector bodies established by the open access to their proceedings enables to track their actions in detail and improves the trust citizens put in the bodies. Nevertheless, the release of open data may reveal many fallacies of public sector bodies, which may produce a temporary disillusion, distrust in government, and loss of interest in politics [40].

The initial assumption of most open data advocates is that the data made in the public sector may be relied on. However, the public sector data cannot be treated as neutral and uncontested resource. *"Unaudited, unverfied statistics abound in government data, particularly when outside parties-local government agencies, federal lobbyists, campaign committees-collect the data and turn it over to the government"* [65, p. 261]. False data may be fabricated to provide alibi for corruption behaviour. For instance, Nithya Raman draws attention to an Indian dataset on urban planning in which non-existent public toilets are present, so that the spending, that supposedly goes for the toilets' maintenance, may be

[11] For example, in 2006 it was found that 52% of links from the official parliamentary record of the UK were not functional [16, p. 20].

justified [88]. Another example that demonstrates how false data is contained with the public sector data is the exposure of errors in subsidies awarded by the EU Common Agricultural Policy. The data shows that the oldest recipients of these funds, coming from Sweden, were 100 years old, though both dead [16, p. 85].

In the light of such facts, it is important to acknowledge that *"public confidence in the veracity of government-published information is critical to Open Government Data take-off, essential to spurring demand and use of public datasets"* [42]. If the data is regarded as manipulated instead of being recognized as trustworthy, the impact of open data will be significantly diminished.

5.2.7 Procured Data

The public sector is not only considered to be unable to deliver applications in a cost-efficient way, it may also lack the abilities to collect some data. There are several kinds of data, including geospatial surveys, that are difficult to gather using the means available in the public sector. The solution that public bodies adopt for such cases is to outsource data collection to private companies. Using the standard procedures of public procurement, the public bodies contract a provider to produce the requested data.

The challenge starts to appear when commercial data suppliers recognize the value of the procured data and become aware of the possibilities for reuse of such data that might generate revenue for them. Hence the suppliers offer the data under the terms of licences that prevent public sector bodies to share the data with the public, since releasing the data as open data would hamper the suppliers' prospects to resell it. Should the public sector require a licence that allows to open the procured data, it would markedly increase the contract price.

Privatisation of collection of public sector data might be a way to achieve a better efficiency [118], yet without a significant investment it prohibits releasing the data as open data. It leaves open the question asking if public sector bodies should buy in expensive data to share it with others or if the infrastructure of the public sector should be enhanced to cater for acquisition of data that would be difficult to collect without such improvements.

5.3 Summary

Open data creates opportunities that may end up being missed if the challenges associated with them are left unaddressed. This chapter raised some of the questions the open data "movement" would have to face and resolve in order not to lose these opportunities and restore the faith in the transformative potential of open data.

Open data agenda is biased by its prevailing focus on the supply side of open data and its negligence of the demand side that gets to use the data. A significant

part of the challenges associated with open data stems from a narrow-minded view of open data as a technology-triggered change that might be engineered. Although open data brings a change in which technology plays a fundamental role, it is important not to fail to recognize its side effects and the issues that cannot be solved by better engineering.

It is comfortable to abstract away from these issues at hand. So far, the challenges of open data are in most cases temporarily bypassed. While the essential features of open data are described thoroughly, its impact is left mostly unexplored. In fact, open data advocates frequently substitute their expectations for the effects of this relatively new phenomenon. The full implications of open data still need to be worked out. This chapter can be thus read as an outline of some of the areas in which further research may be conducted and case studies may be commissioned.

6 Conclusions

This thesis considered the application of the principles of open data and linked data in the domain of public sector information. It provided both an overview of the required steps and modalities for this application and estimated the outcomes of such a transition, including the advantages, possible side effects, and imminent shortcomings.

The starting chapter defined legal rules for public sector information and options for its disclosure. It prepared the ground for further parts of the thesis by defining the key concepts for the domain of the public sector.

The following chapter introduced open data and considered it as a model for proactive disclosure of public sector information. It covered the implications of the application of the fundamental tenets of openness and transparency to data. The implications were sorted into three categories encompassing legal openness regarding the necessary legal actions, technical openness informing the choices of technology, and data quality concerning the best practices of data maintenance. The chapter reviewed the principles of open data that define the features that are required of data to be recognized as open data.

After a thorough examination of the multifaceted topic of open data the practice of publishing data as linked data was presented. This chapter reviewed the technologies linked data borrows from the semantic web stack and the linked data principles governing their use and implementation practices. It went through the ways how linked data reaches compliance with the publication model prescribed by the principles of open data, which was described in the preceding chapter.

The final chapter delved into the consequences of applying linked open data to public sector information. It provided a critical review of the expected impact resulting from the translation of the principles for linked open data into action, in the course of building an open and generative data infrastructure of the networked public sector. The chapter extrapolated from the current state of affairs and identified several challenges that might have detrimental effect undermining the positive outcomes of the adoption of linked open data in the public sector.

The key contribution of the thesis consisted in a combination of disparate research and activism in the areas of public sector information, open data, and linked data. The application of linked open data to public sector information was thought through to its potential consequences. Linked data was argued to be

among the most appropriate technology choices for publishing open data. In many respects, it was demonstrated to have a head start on comparable technologies considered for publication of open data.

The thesis described both the expected benefits of open data, that play into the hands of open data advocates, and its drawbacks, that hinder its adoption. It identified that a large part of the challenges for open data stems from an almost exclusive focus on the supply side of public sector information, while disregarding to pay attention to the demand side and the issues that arise with the use of the data. The expected merits of open data suffer from the fallacy of the narrow-minded view of technological determinism prevalent in the open data community that purports that the transformative effects of open data are driven by technology. Failing to acknowledge many other dimensions of the complex change towards open data in the public sector is a serious shortcoming abound in open data initiatives. The thesis argues that if left unaddressed, these challenges may compromise the positive impact of linked open data.

7 Bibliography

[1] *8 principles of open government data* [online]. December 7 – 8^{th}, 2007 [cit. 2012-04-07]. Available from WWW: https://public.resource.org/8_principles.html

[2] ACAR, Suzanne; ALONSO, José M.; NOVAK, Kevin (eds.). *Improving access to government through better use of the Web* [online]. W3C Interest Group Note. May 12^{th}, 2009 [cit. 2012-04-06]. Available from WWW: http://www.w3.org/TR/egov-improving/

[3] ALANI, Harith; CHANDLER, Peter; HALL, Wendy; O'HARA, Kieron; SHADBOLT, Nigel; SZOMSZOR, Martin. Building a pragmatic semantic web. *IEEE Intelligent Systems*. May-June 2008, vol. 23, iss. 3, p. 61 – 68. Also available from WWW: http://webscience.org/publications/pragmatic_semantic_web.pdf. ISSN 1541-1672. DOI 10.1109/MIS.2008.42.

[4] American Library Association. *Key principles of government information* [online]. Chicago, 1997 – 2012 [cit. 2012-04-07]. Available from WWW: http://www.ala.org/advocacy/advleg/federallegislation/govinfo/key_principles

[5] ARTHUR, Charles; CROSS, Michael. Give us back our crown jewels. *Guardian* [online]. March 9^{th}, 2006 [cit. 2012-03-09]. Available from WWW: http://www.guardian.co.uk/technology/2006/mar/09/education.epublic

[6] *AusGOAL qualities of open data* [online]. 2011 [cit. 2012-04-07]. Available from WWW: http://www.ausgoal.gov.au/ausgoal-qualities-of-open-data

[7] AYERS, Danny. Evolving the link. *IEEE Internet Computing*. January/February 2007, vol. 11, no. 1, p. 94 – 96. ISSN 1089-7801.

[8] BARNICKEL, Nils; HÖFIG, Edzard; KLESSMANN, Jens; SOTO, Juan. Organisational and societal obstacles to implementations of technical systems supporting PSI re-use. In *Share-PSI Workshop: Removing the Roadblocks to a Pan-European Market for Public Sector Information Re-use* [online]. 2011

[cit. 2012-03-08]. Available from WWW: http://share-psi.eu/submitted-papers/

[9] BENKLER, Yochai. *The wealth of networks: how social production transforms markets and freedom*. New York: Yale University Press, 2006. ISBN 978-0-300-11056-2.

[10] BENNETT, Daniel; HARVEY, Adam. *Publishing open government data* [online]. W3C Working Draft. September 8th, 2009 [cit. 2012-04-07]. Available from WWW: http://www.w3.org/TR/gov-data/

[11] BERLINER, Daniel. The political origins of transparency. In In HAGOPIAN, Frances; HONIG, Bonnie (eds.). *American Political Science Association Annual Meeting Papers, Seattle, Washington, 1 – 4 September 2011* [online]. Washington (DC): American Political Science Association, 2011 [cit. 2012-04-29]. Also available from WWW: http://ssrn.com/abstract=1899791

[12] BERNERS-LEE, Tim. *Linked data: design issues* [online]. Last changed 2009-06-18 [cit. 2011-11-05]. Available from WWW: http://www.w3.org/DesignIssues/LinkedData.html

[13] BERNERS-LEE, Tim. *Putting government data online* [online]. Last changed 2009-06-30 [cit. 2012-04-06]. Available from WWW: http://www.w3.org/DesignIssues/GovData.html

[14] BERNERS-LEE, Tim; SHADBOLT, Nigel. Our manifesto for government data. *Guardian Datablog* [online]. January 21st, 2010 [cit. 2012-04-07]. Available from WWW: http://www.guardian.co.uk/news/datablog/2010/jan/21/timbernerslee-government-data

[15] BERTOT, John C.; JAEGER, Paul T.; GRIMES, Justin M. Using ICTs to create a culture of transparency: e-government and social media as openness and anti-corruption tools for societies. *Government Information Quarterly*. July 2010, vol. 27, iss. 3, p. 264 – 271. DOI 10.1016/j.giq.2010.03.001.

[16] *Beyond access: open government data & the right to (re)use public information* [online]. Access Info Europe, Open Knowledge Foundation, January 7th, 2011 [cit. 2012-04-15]. Available from WWW: http://www.access-info.org/documents/Access_Docs/Advancing/Beyond_Access_7_January_2011_web.pdf

[17] BIZER, Chris; JENTZSCH, Anja; CYGANIAK, Richard. *State of the LOD Cloud* [online]. Version 0.3. September 19th, 2011 [cit. 2012-04-11]. Available from WWW: http://www4.wiwiss.fu-berlin.de/lodcloud/state/

[18] BOYD, Danah; CRAWFORD, Kate. Six provocations for big data. In *Proceedings of A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, 21 – 24 September 2011, University of Oxford*. Oxford (UK): Oxford University, 2011. Also available from WWW: http://ssrn.com/abstract=1926431

[19] CLARKE, Paul. *There's data and there's data* [online]. June 4, 2010 [cit. 2012-04-06]. Available from WWW: http://paulclarke.com/honestlyreal/2010/06/theres-data-and-theres-data

[20] DACONTA, Michael. 10 flaws with the data on Data.gov. *Federal Computer Week* [online]. March 11th, 2010 [cit. 2012-04-10]. Available from WWW: http://fcw.com/articles/2010/03/11/reality-check-10-data-gov-shortcomings.aspx

[21] *Designing URI sets for the UK public sector : a report from the Public Sector Information Domain of the CTO Council's Cross-Government Enterprise Architecture* [online]. 2009 [cit. 2012-02-26]. Available from WWW: http://www.cabinetoffice.gov.uk/media/301253/puiblic_sector_uri.pdf

[22] Data, data everywhere. *Economist*. February 25th, 2010. Also available from WWW: http://www.economist.com/node/15557443

[23] DAVIES, Tim. *Open data, democracy and public sector reform: a look at open government data use from data.gov.uk* [online]. Based on an MSc Dissertation submitted for examination in Social Science of the Internet, University of Oxford. August 2010 [cit. 2012-03-09]. Available from WWW: http://practicalparticipation.co.uk/odi/report/wp-content/uploads/2010/08/How-is-open-government-data-being-used-in-practice.pdf

[24] DAVIES, Tim. *Linked data in international development: practical issues* [online]. Draft 0. 1. September 2011 [cit. 2011-11-07]. Available from WWW: http://www.timdavies.org.uk/wp-content/uploads/1-Primer-Introducing-linked-open-data.pdf

[25] DAVIES, Tim; BAWA, Zainab Ashraf. The promises and perils of open government data (OGD). *Journal of Community Informatics* [online]. 2012 [cit. 2012-04-12], vol. 8, no. 2. Available from WWW: http://ci-journal.net/index.php/ciej/article/view/929/926. ISSN 1712-4441.

[26] DIETRICH, Daniel; GRAY, Jonathan; MCNAMARA, Tim; POIKOLA, Antti; POLLOCK, Rufus; TAIT, Julian; ZIJLSTRA, Ton. *The open data handbook* [online]. 2010 – 2012 [cit. 2012-03-09]. Available from WWW: http://opendatahandbook.org/

[27] DODDS, Leigh; DAVIS, Ian. *Linked data patterns* [online]. Last changed 2011-08-19 [cit. 2011-11-05]. Available from WWW: http://patterns.dataincubator.org

[28] DUMBILL, Edd (ed.). *Planning for big data: a CIO's handbook to the changing data landscape* [ebook]. Sebastopol: O'Reilly, 2012, 83 p. ISBN 978-1-4493-2963-1.

[29] EAVES, David. *Learning from libraries: the literacy challenge of open data* [online]. June 10th, 2010 [cit. 2012-04-11]. Available from WWW: http://eaves.ca/2010/06/10/learning-from-libraries-the-literacy-challenge-of-open-data/

[30] EAVES, David. *UK adopts open government license for everything: why it's good and what it means* [online]. October 1st, 2010 [cit. 2012-04-02]. Available from WWW: http://eaves.ca/2010/10/01/uk-adopts-open-government-license-for-everything-why-its-good-and-what-it-means/

[31] European Commission. *Open data: an engine for innovation, growth and transparent governance* [online]. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Brussels, 2011 [cit. 2012-03-15]. Available from WWW: http://ec.europa.eu/information_society/policy/psi/docs/pdfs/opendata2012/open_data_communication/opendata_EN.pdf

[32] European Commission. *Digital agenda: Commission's open data strategy, questions & answers* [online]. MEMO/11/891. Brussels, December 12th, 2011 [cit. 2012-04-11]. Available from WWW: http://europa.eu/rapid/pressReleasesAction.do?reference=MEMO/11/891

[33] ERLE, Schuyler; GIBSON, Rich; WALSH, Jo. *Mapping hacks: tips & tools for electronic cartography*. Sebastopol: O'Reilly, 2005, 568 p. ISBN 978-0-596-00703-4.

[34] The Council of the European Communities. Council Directive 93/37/EEC of 14 June 1993 concerning the coordination of procedures for the award of public works contracts. *Official Journal of the European Communities*. August 9th, 1993, vol. 36, L 199, p. 54 – 84. Also available from WWW: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31993L0037:EN:PDF. ISSN 0378-6978.

[35] EU. Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases. *Official Journal of the European Union*. 1996, vol. 15, L 77, p. 20 – 28. Also avail-

able from WWW: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:1996:077:0020:0028:EN:PDF. ISSN 1725-2555.

[36] EU. Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information. *Official Journal of the European Union*. 2003, vol. 46, L 345, p. 90 – 96. Also available from WWW: http://ec.europa.eu/information_society/policy/psi/docs/pdfs/directive/psi_directive_en.pdf. ISSN 1725-2555.

[37] EU. *Proposal for a Directive of the European Parliament and of the Council amending Directive 2003/98/EC on re-use of public sector information* [online]. Brussels, December 12th, 2011 [cit. 2012-04-30]. COM (2011) 877. 2011/0430/COD. Available from WWW: http://ec.europa.eu/information_society/policy/psi/docs/pdfs/directive_proposal/2012/en.pdf

[38] FERNÁNDEZ, Javier D.; MARTÍNEZ-PRIETO, Miguel A.; GUTIERREZ, Claudio; POLLERES, Axel. *Binary RDF representation for publication and exchange (HDT)* [online]. W3C Member Submission. March 30th, 2011 [cit. 2012-04-24]. Available from WWW: http://www.w3.org/Submission/2011/SUBM-HDT-20110330/

[39] FIELDING, Roy Thomas. *Architectural styles and the design of network-based software architectures*. Irvine (CA), 2000. 162 p. Dissertation (PhD.). University of California, Irvine.

[40] FIORETTI, Marco. *Open data, open society: a research project about openness of public data in EU local administration* [online]. Pisa, 2010 [cit. 2012-03-10]. Available from WWW: http://stop.zona-m.net/2011/01/the-open-data-open-society-report-2/

[41] FRANCOLI, Mary. What makes governments 'open'?: sketching out models of open government. *eJournal of eDemocracy and Open Government* [online]. 2011 [cit. 2012-03-15], vol. 3, no. 2, p. 152 – 165. ISSN 2075-9517. Available from WWW: http://www.jedem.org/issue/view/5

[42] GIGLER, Bjorn-Soren; CUSTER, Samantha; RAHEMTULLA, Hanif. *Realizing the vision of open government data: opportunities, challenges and pitfalls* [online]. World Bank, 2011 [cit. 2012-04-11]. Available from WWW: http://www.scribd.com/WorldBankPublications/d/75642397-Realizing-the-Vision-of-Open-Government-Data-Long-Version-Opportunities-Challenges-and-Pitfalls

[43] GRAVES, Antoinette. The price of everything the value of nothing. In UHLIR, Paul F. (rpt.). *The socioeconomic effects of public sector information on digital networks: toward a better understanding of different access and reuse policies:*

workshop summary. Washington (DC): National Academies Press, 2009. Also available from WWW: http://books.nap.edu/openbook.php?record_id=12687&page=37. ISBN 0-309-13968-6.

[44] GRAY, Jonathan; HATCHER, Jordan; HEGGE, Becky; PARRISH, Simon; POLLOCK, Rufus. *Unlocking the potential of aid information* [online]. Version 0.2. December 2009 [cit. 2012-04-08]. Available from WWW: http://www.unlockingaid.info/

[45] GRAY, Jonathan. The best way to get value from data is to give it away. *Guardian Datablog* [online]. December 13th, 2011 [cit. 2011-12-14]. Available from WWW: http://www.guardian.co.uk/world/datablog/2011/dec/13/eu-open-government-data

[46] GRAY, Jonathan. *Interview for University of Southhampton open data study* [online]. December 6th, 2011 [cit. 2012-04-06]. Available from WWW: http://jwyg.okfn.org/2011/12/06/interview-for-university-of-southhampton-open-data-study/

[47] GRUBER, Thomas R. A translation approach to portable ontology specifications. *Knowledge Acquisition*. 1993, vol. 5, iss. 2, p. 199 – 220. Also available from WWW: http://tomgruber.org/writing/ontolingua-kaj-1993.htm

[48] GURSTEIN, Michael. Open data: empowering the empowered or effective data use for everyone? *First Monday* [online]. February 7th, 2011 [cit. 2012-04-01], vol. 16, no. 2. Available from WWW: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3316/2764

[49] HALONEN, Antti. *Being open about data: analysis of the UK open data policies and applicability of open data* [online]. Report. London: Finnish Institute, 2012 [cit. 2012-04-05]. Available from WWW: http://www.finnish-institute.org.uk/images/stories/pdf2012/being%20open%20about%20data.pdf

[50] HARTIG, Olaf; BIZER, Christian; FREYTAG, Johann-Christoph. Executing SPARQL queries over the web of linked data. In *Proceedings of the 8th International Semantic Web Conference, Chantilly (VA), USA, 25 – 29 October, 2009*. Berlin, Heidelberg, New York: Springer, 2009. Lecture notes in computer science, vol. 5823. Also available from WWW: http://www.dbis.informatik.hu-berlin.de/fileadmin/research/papers/conferences/2009_iswc_hartig_preprint.pdf. ISBN 978-3-642-04929-3.

[51] HAZLETT, Shirley-Ann; HILL, Frances. E-government: the realities of using IT to transform the public sector. *Managing Quality Service*. 2003, vol. 13, iss. 6, p. 445 – 452. ISSN 0960-4529. DOI 10.1108/09604520310506504.

[52] HEATH, Tom; BIZER, Chris. *Linked data: evolving the Web into a global data space*. 1st ed. Morgan & Claypool, 2011. Also available from WWW: http://linkeddatabook.com/book. ISBN 978-1-60845-430-3. DOI 10.2200/S00334ED1V01Y201102WBE001.

[53] HIRST, Tony. *So what's open government data good for? Government and "independent advisers", maybe?* [online]. July 7th, 2011 [cit. 2012-04-07]. Available from WWW: http://blog.ouseful.info/2011/07/07/so-whats-open-government-data-good-for-government-maybe/

[54] HÖCHTL, Johann; REICHSTÄDTER, Peter. Linked open data: a means for public sector information management. In ANDERSEN, Kim Normann; FRANCESCONI, Enrico; GRÖNLUND, Åke; VAN ENGERS, Tom M. (eds.). *Electronic Government and the Information Systems Perspective: proceedings of the second international conference, Toulouse, France, August 29 – September 2, 2011.* Heidelberg: Springer, 2011, p. 330 – 343. Lecture notes in computer science, vol. 6866. DOI 10.1007/978-3-642-22961-9_26.

[55] HOGAN, Aidan; CYGANIAK, Richard. *Frequently observed problems on the web of data* [online]. Version 0.3. November 13th, 2009 [cit. 2012-04-23]. Available from WWW: http://pedantic-web.org/fops.html

[56] HOGAN, Aidan; UMBRICH, Jürgen; HARTH, Andreas; CYGANIAK, Richard; POLLERES, Axel; DECKER, Stefan. An empirical survey of linked data conformance. In *Journal of Web Semantics* [in print]. 2012. Also available from WWW: http://sw.deri.org/~aidanh/docs/ldstudy12.pdf. ISSN 1570-8268. DOI 10.1016/j.websem.2012.02.001.

[57] HOWARD, Alex. *Data for the public good* [ebook]. 1st ed. Sebastopol : O'Reilly, 2012. ISBN 978-1-449-32976-1.

[58] HUIJBOOM, Noor; VAN DEN BROEK, Tijs. Open data: an international comparison of strategies. *European Journal of ePractice* [online]. March/April 2011 [cit. 2012-04-30], no. 12. Available from WWW: http://www.epractice.eu/files/European%20Journal%20epractice%20Volume%2012_1.pdf. ISSN 1988-625X.

[59] HYLAND, Bernardette; TERRAZAS, Boris Villazón; CAPADISLI, Sarven. *Cookbook for open government linked data* [online]. Last modified on February 20th, 2012 [cit. 2012-04-11]. Available from WWW: http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook

[60] JACOBS, Ian; WALSH, Norman (eds.). *Architecture of the World Wide Web, volume 1* [online]. W3C Recommendation. December 15th, 2004 [cit. 2012-04-20]. Available from WWW: http://www.w3.org/TR/webarch/

[61] KAPLAN, Daniel. *Open public data: then what? Part 1* [online]. January 28th, 2011 [cit. 2012-04-10]. Available from WWW: http://blog.okfn.org/2011/01/28/open-public-data-then-what-part-1/

[62] KIENLE, Holger M. *Open data: reverse engineering and maintenance perspective* [online]. February 8th, 2012 [cit. 2012-03-08]. Available from WWW: http://arxiv.org/abs/1202.1656

[63] KOUMENIDES, Christos L.; SALVADORES, Manuel; ALANI, Harith; SHADBOLT, Nigel R. Global integration of public sector information. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-line, April 26 – 27th, 2010, Raleigh (NC), US*. Raleigh, 2010.

[64] KUNDRA, Vivek. *Digital fuel of the 21st century: innovation through open data and the network effect* [online]. President and Fellows of Harvard College, 2012 [cit. 2012-03-15]. Discussion Paper Series, no. D-70. Available from WWW: http://www.hks.harvard.edu/presspol/publications/papers/discussion_papers/d70_kundra.html

[65] LATHROP, Daniel; RUMA, Laurel (eds.). *Open government: collaboration, transparency, an d participation in practice*. Sebastopol: O'Reilly, 2010. ISBN 978-0-596-80435-0.

[66] LEBO, Timothy; WILLIAMS, Gregory Todd. Converting governmental datasets into linked data. *I-Semantics 2010: proceedings of the 6th International Conference on Semantic Systems, September 1 – 3, 2010, Graz, Austria*. New York (NY): ACM, 2010. ISBN 978-1-4503-0014-8. DOI 10.1145/1839707.1839755.

[67] LESSIG, Lawrence. Against transparency: the perils of openness in government. *The New Republic* [online]. October 9th, 2009 [cit. 2012-03-29]. Available from WWW: http://www.tnr.com/article/books-and-arts/against-transparency

[68] LIENERT, Ian. *Where does the public sector end and the private sector begin?* [online]. June 1st, 2009 [cit. 2012-04-29]. IMF working paper, no. 09/122. Also available from WWW: http://www.imf.org/external/pubs/ft/wp/2009/wp09122.pdf

[69] LONGO, Justin. #OpenData: digital-era governance thoroughbred or new public management Trojan horse? *Public Policy & Governance Review*. Spring 2011, vol. 2, no. 2, p. 38 – 51. Also available from WWW: http://ssrn.com/abstract=1856120

[70] LOVLEY, Erika. The government has a database for most everything. *Politico* [online]. June 24th, 2009 [cit. 2012-04-07]. Available from WWW: http://www.politico.com/news/stories/0609/24118.html

[71] MAALI, Fadi. *Getting to the five-star: from raw data to linked government data.* Galway, 2011. Masters thesis (MSc.). National University of Ireland. Digital Enterprise Research Institute.

[72] MAIER-RABLER, Ursula; HUBER, Stefan. "Open": the changing relation between citizens, public administration, and political authority. *eJournal of eDemocracy and Open Government* [online]. 2011 [cit. 2012-03-15], vol. 3, no. 2, p. 182 – 191. ISSN 2075-9517. Available from WWW: http://www.jedem.org/issue/view/5

[73] MALAMUD, Carl. *By the people* [online]. Government 2.0 Summit. Washington (DC), September 10th, 2009 [cit. 2011-03-23]. Available from WWW: http://public.resource.org/people/

[74] MCCLEAN, Tom. Not with a bang but with a whimper: the politics of accountability and open data in the UK. In HAGOPIAN, Frances; HONIG, Bonnie (eds.). *American Political Science Association Annual Meeting Papers, Seattle, Washington, 1 – 4 September 2011* [online]. Washington (DC): American Political Science Association, 2011 [cit. 2012-04-19]. Also available from WWW: http://ssrn.com/abstract=1899790

[75] MENDEL, Toby. Freedom of information: an internationally protected human right. *Comparative Media Law Journal.* 2003, no. 1. Also available from WWW: http://www.juridicas.unam.mx/publica/rev/comlawj/cont/1/cts/cts3.htm

[76] MENDELSOHN, Noah. *The self-describing web* [online]. W3C TAG Finding. February 7th, 2009 [cit. 2012-04-11]. Available from WWW: http://www.w3.org/2001/tag/doc/selfDescribingDocuments

[77] MILLER, Paul; STYLES, Rob; HEATH, Tom. Open Data Commons, a license for open data. In BIZER, Christian; HEATH, Tom; IDEHEN, Kingsley; BERNERS-LEE, Tim (eds.). *Linked Data on the Web (LDOW 2008): proceedings of the WWW2008 Workshop on Linked Data on the Web, Beijing, China, April 22nd, 2008.* Aachen: RWTH Aachen University, 2008. CEUR workshop proceedings, vol. 369. ISSN 1613-0073.

[78] MOORE, Martin. 10 reasons why news organizations should use 'linked data'. *Idea Lab* [online]. March 16th, 2010 [cit. 2012-04-24]. Available from WWW: http://www.pbs.org/idealab/2010/03/10-reasons-why-news-organizations-should-use-linked-data073.html

[79] MYNARZ, Jindřich. *Principled open data* [online]. March 20th, 2012 [cit. 2012-04-22]. Available from WWW: http://headtoweb.posterous.com/principled-open-data

[80] NAPOLI, Philip M.; KARAGANIS, Joe. On making public policy with publicly available data: the case of U.S. communications policymaking. *Government Information Quarterly*. October 2010, vol. 27, iss. 4, p. 384 – 391. DOI 10.1016/j.giq.2010.06.005.

[81] ØLNES, Svein. Interoperability in public sector: how use of a lightweight approach can reduce the gap between plans and reality. In WIMMER, Maria A.; CHAPPELET, Jean-Loup; JANSSEN, Marijn, SCHOLL, Hans Jochen (eds.). *Electronic Government: 9th International Conference: proceedings, Lausanne, Switzerland, August 29 – September 2, 2010*. Heidelberg: Springer, 2010, p. 315 – 326. Lecture notes in computer science, 6228. ISBN 978-3-642-14798-2.

[82] OMITOLA, Tope; KOUMENIDES, Christos L.; POPOV, Igor O.; YANG, Yang; SALVADORES, Manuel; SZOMSZOR, Martin; BERNERS-LEE, Tim; GIBBINS, Nicholas; HALL, Wendy; SCHRAEFEL, Mc; SHADBOLT, Nigel. Put in your postcode, out come the data: a case study. In AROYO, Lora; ANTONIOU, Grigoris; HYVONËN, Eero; TEN TEIJE, Annette; STUCKENSCHMIDT, Heiner; CABRAL, Liliana; TUDORACHE, Tania (eds.). *The semantic web: research and applications, 7th Extended Semantic Web Conference, Heraklion, Crete, Greece, May 30 – June 3, 2010, Proceedings, Part I*. Heidelberg: Springer, 2010. Lecture notes in computer science, 6088. ISBN 978-3-642-13485-2.

[83] *Open declaration on European public services* [online]. 2009 [cit. 2012-04-07]. Available from WWW: http://eups20.wordpress.com/the-open-declaration/

[84] *Open definition* [online]. Version 1.1. November 2009 [cit. 2012-03-17]. Available from WWW: http://opendefinition.org/okd/

[85] ORAM, Andy. European Union starts project about economic effects of open government data. *O'Reilly Radar* [online]. June 11th, 2010 [cit. 2012-04-09]. Available from WWW: http://radar.oreilly.com/2010/06/european-union-starts-project.html

[86] ORSZAG, Peter R. *Open government directive*. M-10-06. Memorandum for the heads of executive departments and agencies. Washington: Executive Office of the President, December 8th, 2009. Also available from WWW: http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-06.pdf

[87] POLLOCK, Rufus. *Open data: a means to an end, not an end in itself* [online]. September 15th, 2011 [cit. 2012-04-06]. Available from WWW: http://blog.okfn.org/2011/09/15/open-data-a-means-to-an-end-not-an-end-in-itself/

[88] RAMAN, Nithya V. Collecting data in Chennai city and the limits of openness. *Journal of Community Informatics* [online]. 2012 [cit. 2012-04-12], vol. 8, no. 2. Available from WWW: http://ci-journal.net/index.php/ciej/article/view/877/908. ISSN 1712-4441.

[89] RFC 2616. *Hypertext Transfer Protocol: HTTP/1.1* [online]. FIELDING, Roy Thomas; GETTYS, J.; MOGUL, J.; FRYSTYK, H.; MASINTER, L.; LEACH, P.; BERNERS-LEE, Tim. June 1999 [cit. 2012-04-21], 176 p. Available from WWW: http://tools.ietf.org/html/rfc2616. ISSN 2070-1721.

[90] RDF 3986. *Uniform Resource Identifier (URI): generic syntax* [online]. BERNERS-LEE, Tim; FIELDING, Roy Thomas; MASINTER, Larry. January 2005 [cit. 2012-04-23]. 61 p. Available from WWW: http://tools.ietf.org/html/rfc3986. ISSN 2070-1721.

[91] ROBINSON, David G.; YU, Harlan; ZELLER, William P.; FELTEN, Edward W. Government data and the invisible hand. *Yale Journal of Law & Technology*. 2009, vol. 11, p. 160 – 175.

[92] RODRIGUEZ, Marko A. A reflection on the structure and process of the web of data. *Bulletin of the American Society for Information Science and Technology*. August/September, 2009, vol. 35, no. 6. ISSN 1550-836.

[93] SCHELLONG, Alexander; STEPANETS, Ekaterina. *Unchartered waters: the state of open data in Europe* [online]. CSC, 2011 [cit. 2012-04-12]. Public sector study series, 01/2011. Available from WWW: http://assets1.csc.com/de/downloads/CSC_policy_paper_series_01_2011_unchartered_waters_state_of_open_data_europe_English_2.pdf

[94] SDMX. *SDMX content-oriented guidelines. Annex 1: cross-domain concepts.* 2009. Also available from WWW: http://sdmx.org/wp-content/uploads/2009/01/01_sdmx_cog_annex_1_cdc_2009.pdf

[95] SEQUEDA, Juan F.; CORCHO, Oscar. Linked stream data: a position paper. In TAYLOR, Kerri; AYYAGARI, Arun; DE ROURE, David (eds.). *Proceedings of the 2nd International Workshop on Semantic Sensor Networks, collocated with the 8th International Semantic Web Conference, Washington DC, USA, October 26th, 2009*. Aachen: RWTH Aachen University, 2009, p. 148 – 157. CEUR workshop proceedings, vol. 552. Also available from WWW: http://oa.upm.es/5442/1/INVE_MEM_2009_64353.pdf. ISSN 1613-0073.

[96] SHADBOLT, Nigel. *Towards a pan EU data portal – data.gov.uk*. Version 4.0. December 15th, 2010 [cit. 2012-03-10]. Available from WWW: http://ec.europa.eu/information_society/policy/psi/docs/pdfs/towards_an_eu_psi_portals_v4_final.pdf

[97] SHADBOLT, Nigel; O'HARA, Kieron; SALVADORES, Manuel; ALANI, Harith. eGovernment. In DOMINGUE, John; FENSEL, Dieter; HENDLER, James A. (eds.). *Handbook of semantic web technologies*. Berlin: Springer, 2011, p. 849 – 910. DOI 10.1007/978-3-540-92913-0_20.

[98] SHERIDAN, John; TENNISON, Jeni. Linking UK government data. In BIZER, Christian; HEATH, Tom; BERNERS-LEE, Tim; HAUSENBLAS, Michael (eds.). *Linked Data on the Web: proceedings of the WWW 2010 Workshop on Linked Data on the Web, April 27th, 2010, Raleigh, USA*. Aachen: RWTH Aachen University, 2010. CEUR workshop proceedings, vol. 628. ISSN 1613-0073.

[99] SHIRKY, Clay. Open House thoughts, Open Senate direction. In *Open House Project* [online]. November 23rd, 2008 [cit. 2012-04-19]. Available from WWW: http://groups.google.com/group/openhouseproject/msg/53867cab80ed4be9

[100] Socrata. *2010 open government data benchmark study* [online]. Version 1.4. Last updated January 4th, 2011 [cit. 2012-04-07]. Available from WWW: http://www.socrata.com/benchmark-study

[101] SOLDA-KUTZMANN, Donatella. Public sector information: a market without failure? In *Share-PSI Workshop: Removing the Roadblocks to a Pan-European Market for Public Sector Information Re-use* [online]. 2011 [cit. 2012-03-09]. Available from WWW: http://share-psi.eu/submitted-papers/

[102] STICKLER, Patrick. *CBD: concise bounded description* [online]. W3C Member Submission. June 3rd, 2004 [cit. 2012-04-23]. Available from WWW: http://www.w3.org/Submission/CBD/

[103] STIGLITZ, Joseph E. *On liberty, the right to know, and public discourse: the role of transparency in public life*. Oxford Amnesty Lecture. Oxford (UK), 1999. Also available from WWW: http://derechoasaber.org/documentos/pdf0116.pdf

[104] TAUBERER, Joshua. *Open data is civic capital: best practices for "open government data"* [online]. Version 1.5. January 29th, 2011 [cit. 2012-03-17]. Available from WWW: http://razor.occams.info/pubdocs/opendatacivliccapital.html

[105] TAUBERER, Joshua. *Open government data: principles for a transparent government and an engaged public* [online]. 2012 [cit. 2012-03-09]. Available from WWW: http://opengovdata.io/

[106] Transparency Board. *New public sector transparency board and public data transparency principles* [online]. June 25th, 2010 [cit. 2012-04-15]. Available from WWW: http://data.gov.uk/blog/new-public-sector-transparency-board-and-public-data-transparency-principles

[107] TRAVIS, Alan; MULHOLLAND, Hélène. Online crime maps crash under weight of 18 million hits an hour. *Guardian* [online]. February 1st, 2011 [cit. 2012-04-17]. Available from WWW: http://www.guardian.co.uk/uk/2011/feb/01/online-crime-maps-power-hands-people

[108] VAN DEN BROEK, Tijs; KOTTERINK, Bas; HUIJBOOM, Noor; HOFMAN, Wout; VAN GRIEKEN, Stefan. Open data need a vision of smart government. In *Share-PSI Workshop: Removing the Roadblocks to a Pan-European Market for Public Sector Information Re-use* [online]. 2011 [cit. 2012-03-09]. Available from WWW: http://share-psi.eu/submitted-papers/

[109] VAN DER SLOOT, Bart. On the fabrication of sausages, or of open government and private data. *eJournal of eDemocracy and Open Government* [online]. 2011 [cit. 2012-03-15], vol. 3, no. 2, p. 136 – 154. ISSN 2075-9517. Available from WWW: http://www.jedem.org/issue/view/5

[110] VICKERY, Graham. *Review of the recent developments on PSI re-use and related market developments* [online]. Final version. Paris, 2011 [cit. 2012-04-19]. Available from WWW: http://ec.europa.eu/information_society/policy/psi/docs/pdfs/report/psi_final_version_formatted.docx

[111] WEINBERGER, David. *Transparency is the new objectivity* [online]. July 19th, 2009 [cit. 2012-04-25]. Available from WWW: http://www.hyperorg.com/blogger/2009/07/19/transparency-is-the-new-objectivity/

[112] WEINBERGER, David. *Too big to know*. New York (NY): Basic Books, 2012. ISBN 978-0-465-02142-0.

[113] WONDERLICH, John. Pelosi reverses on 72 hour promises? In *Open House Project* [online]. November 7th, 2009 [cit. 2012-04-19]. Available from WWW: http://groups.google.com/group/openhouseproject/msg/94060a876083d86a

[114] WOOD, David (ed.). *Linking enterprise data*. 1st ed. Heidelberg: Springer, 2010, XXVI, 291 p. Also available from WWW: http://3roundstones.com/led_book/led-contents.html. ISBN 978-1-4419-7664-2.

[115] WOOD, David (ed.). *Linking government data*. Heidelberg: Springer, 2011. ISBN 978-1-4614-1766-8.

[116] WRUUCK, Patricia. 2012: the year of big data. *European Public Policy Blog* [online]. Brussels, May 1st, 2012 [cit. 2012-05-01]. Available from WWW: http://googlepolicyeurope.blogspot.com/2012/05/2012-year-of-big-data.html

[117] YAKOWITZ, Jane. Tragedy of the data commons. *Harvard Journal of Law & Technology*. Fall 2011, vol. 25, no. 1. Also available from WWW: http://ssrn.com/abstract=1789749

[118] YIU, Chris. *A right to data: fulfilling the promise of open public data in the UK* [online]. Research note. March 6th, 2012 [cit. 2012-03-06]. Available from WWW: http://www.policyexchange.org.uk/publications/category/item/a-right-to-data-fulfilling-the-promise-of-open-public-data-in-the-uk

[119] YU, Harlan; ROBINSON, David G. *The new ambiguity of "open government"* [online]. Princeton CITP / Yale ISP Working Paper. Draft of February 28th, 2012. Available from WWW: http://ssrn.com/abstract=2012489

[120] ZITTRAIN, Jonathan. *The future of the Internet: and how to stop it*. New Haven: Yale University Press, 2008. Also available from WWW: http://futureoftheinternet.org/static/ZittrainTheFutureoftheInternet.pdf. ISBN 978-0-300-15124-4.