Business Information Systems

Krzysztof Węcel

# Big, Open and Linked Data

## Effects and Value for the Economy

Springer

# Business Information Systems

Currently, virtually every organization is using business information systems integrating information technology, business processes, and people. With the advent of new technologies and new business opportunities that arise (e.g. electronic markets), designing and developing information systems that would be in line with the long term strategy of companies and that would efficiently support their core business processes, becomes even a more complex and difficult task. In addition, business information systems in modern companies need to evolve and adapt to changes occurring both inside as well as outside an organization. Therefore, a broad range of issues such as new business requirements (including societal and legal aspects), new market mechanisms and business models, as well as new paradigms and system architectures, need to be considered. The "Springer Series on Business Information Systems" targets high quality research monographs, texts and contributed books covering the vast field of business information processing and applications in various domains. It focuses particularly on fostering the exchange on theoretical and practical aspects of the design and development, implementation and application of business information systems, based on innovative concepts. The book series is not only aimed at researchers and students, but also at information system professionals in industry, commerce, and public administration, who are interested in new ideas on business information systems. Thus, with this series we establish a unique platform for design scienceoriented research and the technology and economically oriented scientific community, trying to bridge the gap between theoretical foundations and real world requirements.

Krzysztof Węcel

# Big, Open and Linked Data

## Effects and Value for the Economy

Springer

Krzysztof Węcel
Department of Information Systems
Poznań University of Economics
and Business
Poznań, Poland

ISSN 2662-1797 ISSN 2662-1800 (electronic)
Business Information Systems
ISBN 978-3-031-07146-1 ISBN 978-3-031-07147-8 (eBook)
https://doi.org/10.1007/978-3-031-07147-8



This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

*To my wife Anna and daughter Kornelia*

# Contents

# List of Figures

# List of Tables

# Chapter 1
# Introduction


## 1.1 Background and Motivation

There is increasing interest in various technologies to support working with the growing data assets collected by enterprises. This is reflected in such concepts as big data, open data, linked data or data science. Research on linked data was a primary motivation; however, it also quickly became necessary to address other types of data: big data and open data.

Big data can be intuitively defined with regard to its size. It is then defined as data that exceeds the capability of a typical database to store and process the data or that which forces us to look beyond known technologies. Big data is a result of the progress in data storage technologies and transfer capabilities. People store more and more data because they hope that one day, they will be able to infer knowledge from that data. Nevertheless, in order to gain value from this data, it is necessary to elaborate upon the methods, techniques, and tools for processing it. Aside from volume, there are also other hallmarks of big data. Data can move fast, change structure very often, and provide ambiguous values. These are the challenges for modern information systems.

Open data is data that can be freely used, modified, and shared by anyone for any purpose. 'Freely' is interpreted as both open—everybody can access it, and free—it is available at no cost. The concept has been popularized by many initiatives on a government level, such as Data.gov and Data.gov.uk. It was also regulated on the European Union level as Public Sector Information (PSI). Open data stirs interest due to its potential to improve the delivery of public services by changing how governments work. It can also empower citizens and create added value for enterprises, and it is also expected to bring about significant benefits to the overall economy. There are reports suggesting that open data can annually unlock $3–5 trillion in economic value.

Linked data is a concept that was coined in comparison to the Web and is based on web technologies. Just as web pages are browsed in the search for information,



K. Węcel, *Big, Open and Linked Data*, Business Information Systems,
https://doi.org/10.1007/978-3-031-07147-8_1

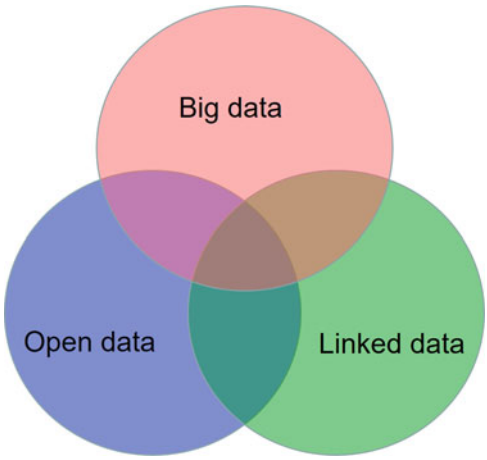



**Fig. 1.1** Intersection of big, open, and linked data

it should be possible to navigate from one dataset to another. With linked data, it is possible to find other related data by exploring the web of data. The initial work introducing the concept of linked data by Berners–Lee did not provide a definition but rather focused on a set of so-called 'linked data principles'. The definition is only implicit by specifying the rules for putting data online. These principles cover technical aspects, but it would be more interesting to study their economic impact.

Nowadays, enterprises encounter all of these forms of data. By the time this work started, it was not common to think about using all of these forms together. Each of the intersecting areas poses specific issues and challenges (see Fig. 1.1). Big data is usually internal data owned by enterprises and typically contains records about the behavior of people, either directly or by the devices used by them. Open data is data owned by the government, being external to any party interested in using it, like enterprises and citizens. While big data is rather structured, open data can very often be weakly structured. Linked data was created as an effort to bring structure to open data. This effort was initially taken on by citizens-enthusiasts, but the concept is slowly being adopted by governments and enterprises. Linked data is not just a technology; it is inherent to open data, a philosophy to make government data available to all interested parties. Thus, it is a whole ecosystem offering new opportunities for data reuse.

## 1.2 Objectives and Research Hypotheses

The idea to combine big, open, and linked data for the benefit of enterprises is based on our confidence that further potential can be released from data by applying advanced analytics to combined proprietary and open knowledge. We have already combined internal data with external information. In the book Abramowicz et al. (2002), we considered a data warehouse as a starting point for defining information

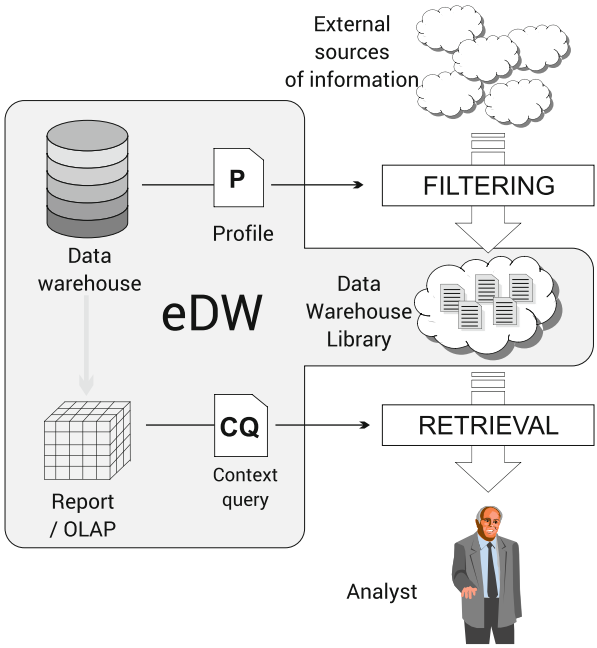



**Fig. 1.2** Transparent information filtering and retrieval. Source: based on Abramowicz et al. (2002)

needs, an idea referred to as *enhanced Data Warehouse*. It realized the following vision: an analyst, looking at a specific data warehouse report, can also browse documents that provide additional context to this report (see Fig. 1.2). By that time, a data warehouse could be interpreted as big data. Information needs were expressed in profiles, which were built and represented using semantic technologies. Translating into current terms, we used linked data to provide a context for big data. As time passed, new technologies and new possibilities emerged and we have identified big, open, and linked data as the new opportunity.

In order to bring open and linked data to enterprises, to support big data, we need to discuss its value. Value can be analyzed on the macroeconomic and microeconomic level. The first is characteristic for governments, and the latter is closer to enterprises and citizens. Governments embraced the concept of open data to improve the overall welfare of citizens. They introduced policies justified by such concepts as transparency and accountability. The goal of governments was to stimulate the innovation activities of enterprises and skilled individuals, and the common belief was that giving them public data will result in new services for citizens and other enterprises, lending to the multiplier effect.

For this to happen, companies need to identify incentives. This is the place where business models can be helpful. A business model is defined as a way a business creates and captures value from services or products and it can also include the rationale how an organization delivers the value. The concept of a business model has already received much attention in literature on management and strategy. There have been a multitude of business models: for the Web, for data alone; however,

there have been no business models for linked data. Identification of this gap encouraged us to carry out detailed studies on value creation and value capture in the context of linked data.

The main objective of this work is to contribute the theoretical foundations, economic justification, and technical framework under which big, open, and linked data can be unified. The following research hypotheses were formulated:

**Hypothesis 1**: Opening and sharing of data makes the economy as a whole more effective.
**Hypothesis 2**: Sharing of data does not lead to losing the competitive advantage.
**Hypothesis 3**: Structuring and contextualization of data and information increase their value.

## 1.3 Structure of the Book

The identified research hypotheses were used to structure the content of this book. The remaining chapters are organized as follows:

Chapter 2 describes the phenomenon of open data. We start with the roots of the open data movement and explain the main principles, and several open data initiatives are described. We also analyze the legal frameworks for open data supply. Adoption of open data is not straightforward; therefore, we also characterize the barriers for adoption. On the motivating side, we analyze what data is available with a specific focus on statistical data, industry classifications, and geographical data.

Chapter 3 provides details about linked data. We start with a literature analysis to come up with the broad definition of linked data. Next, we analyze what linked data assets are available for reuse. One of the sections describes how external information can provide a context for internal data, where ontologies were introduced. This is followed by a discussion on data quality. The chapter concludes with the discoverability of datasets, an issue particularly important in the context of linked data. It also shows the need for cataloging as a consequence of increasing data resources and presents the results of a study on vocabularies used in cataloging.

Chapter 4 covers the organizational challenge stemming from big data. We first start with contemporary solutions used by enterprises for making decisions and then move to the definition of big data, followed by a discussion on various aspects of big data that help in the understanding of this phenomenon. The relations between open data, linked data, and big data provide foundations for the formulation of data unification challenges. We then present our model for linked data-based unification of data. The model is applied to analyze modern enterprise solutions and the potential of leveraging linked data.

The next two chapters elaborate upon the value of data, as viewed through macroeconomics and microeconomics. Chapter 5 starts with the macroeconomic aspects, where we study the macroeconomic impact of open data. We then analyze both the direct value of big, open, and linked data, as well as the multiplier effects. The goal of this chapter is to demonstrate the positive impact of sharing data for the national economy (hypothesis 1).

Chapter 6 supplements the microeconomic aspects. We start by introducing the stakeholders within the open data ecosystem and discuss demand and supply issues. We then study the mutual benefits mostly by reference to the sharing economy. Data ownership is also described as a follow-up to sharing. Two specific cases of data sharing are then described: personal data and innovation. The goal of this chapter is to demonstrate that the sharing of data does not lead to a loss of the competitive advantage (hypothesis 2).

Chapter 7 is a transition from theoretical considerations on economics grounds to practical applications. It explains how enterprises create, transfer, and capture value from using data. We start with the overall characteristics of business model research. For this purpose, a structured literature review methodology is applied. We present results of the study in two sections: one for the analysis of business model components, and the other for the analysis of relevant business models. The chapter concludes with a discussion that provides further arguments for hypothesis 2.

Chapter 8 proposes a method for geographical profiling of mobile subscribers. The essence of the method is the combination of big data with linked data. Big data is the internal data of a telecom operator in the form of call detail records, and linked data is crowdsourced geographical information. In the consecutive sections, we characterize the spatial information and bring the mobile data subject closer. Geographical linked data-based profiling is introduced in two variants: to obtain the base transceiver station profiles and the user profiles. The method is then extended to use the notion of latent variables. We show how structuring for big data with concepts provided by linked data makes internal telecom data more useful, thus more valuable (hypothesis 3). The potential was unlocked by applying advanced analytics to the combined data.

## Reference

Abramowicz, W., Kalczyński, P. J., & Węcel, K. (2002). *Filtering the Web to feed data warehouses*. Springer. ISBN: 1-85233-579-3. (pages 2, 3)

# Chapter 2
# Open Data as an Economic, Political, and Technical Phenomenon


## 2.1 Introduction

According to the widely accepted Open Definition, "open data and content can be freely used, modified, and shared by anyone for any purpose."[1] It combines both 'open', as in the Open Source Definition, and 'free' as in the Free Software Definition. The concept is not new, but it was popularized by many initiatives that have their roots, among others, in open government ideas, such as Data.gov and Data.gov.uk. It was later regulated by the European Commission in the Directive 2003/98/EC on the reuse of public sector information (EU, 2003). Public Sector Information (PSI) is "information that public sector bodies collect, produce, reproduce, and disseminate in many areas of activity while accomplishing their public tasks" (Deloitte, 2013, p. 151). These public sector bodies are referred to as a Public Sector Information Holder (PSIH). The above directive has an economic goal to facilitate the development of innovative services and the free exchange of market information.

Open data is a movement raising interest for its potential to improve the delivery of public services by changing how governments work. It can also empower citizens and create added value for enterprises. Reports suggest that open data can annually unlock $3–5 trillion in economic value (Manyika et al., 2013). Further potential can be released by applying advanced analytics to combined proprietary and open knowledge.

Open data also constitutes a cornerstone of the European Single Digital Market. Internet and digital technologies offer new possibilities, which so far are not fully exploited by governments and companies. There is very strong economic motivation, as "tearing down regulatory walls and moving from 28 national markets to a single one [...] could contribute 415 billion euro per year to the [European]

[1] http://opendefinition.org/.



K. Węcel, *Big, Open and Linked Data*, Business Information Systems,
https://doi.org/10.1007/978-3-031-07147-8_2

economy and create hundreds of thousands of new jobs."[2] In most cases, open data means crowdsourced data, i.e., provided by a community of users, but this results in certain disadvantages: quality is mentioned as one of the challenges (Węcel & Lewoniewski, 2015). Data may be incomplete, not up-to-date, inaccurate, or incorrect. One of the approaches to mitigate these deficiencies is to use several sources and then verify the information.

It is inherent to the open data philosophy to make government data available to all interested parties. This creates a whole ecosystem, offering opportunities for data reuse. As data is no longer produced only by specialized entities, the idea changes the value chain, and new business models are defined around open data.

In this chapter, we analyze the phenomenon of open data. We present several initiatives that make use of open data and offer certain benefits to society. We also analyze the open data supply and provide some examples of the available macroeconomic information.

## 2.2 Open Data Movement

In recent years, a number of open data movements have emerged. The milestones were set by the Public Sector Information (PSI) Directive in 2003 in Europe, US President Obama's open data initiative in 2009, the Open Government Partnership in 2011, and the G8 Open Data Charter in 2013 (Attard et al., 2015). As a result of political decisions, several open government data portals were created. In 2009, after enacting the Open Government Directive, `data.gov` was established to serve as a portal for US federal agencies to publish open data. That same year `data.gov.uk` was also launched in closed beta to publish non-personal UK government data as open data. In 2012, the US government published a strategy in which it indicated open data as an important source of support to sustain its twenty-first century e-Government development (US CIO Council, 2012). Since 2013, according to an executive order by Obama, US government information should be, by default, open and machine readable. Based on the Data Act from 2014, data on US federal spending should be open, standardized, and published online.

Since its inception, open data has been considered as the key pillar to sustain open government. The Open Data White Paper released by the UK government also stated the critical role of open data in building a transparent society and unleashing the potential of government data (Open Data White Paper: Unleashing the Potential, 2012). The open data movement is an effort to address the data deficit challenge—the challenge that can only be addressed with certain regulations and agreements. Common resources allow for the strengthening of civil engagement and improving decisions and policy making (World Economic Forum, 2015).

[2] https://ec.europa.eu/priorities/digital-single-market_en.

Nowadays, open data is an important movement among government administrations around the world and has also started attracting researchers to study this field (Yang et al., 2015). The discussion has moved away from dataset formats and portals towards more advanced and broader topics: principles of open data, measurement of impact, common standards, privacy issues, or multilingual data (Enabling the Data Revolution, 2015). We can also observe increased granularity—not only the open data of national governments, but also smaller communities, like cities or even the private sector, are becoming increasingly interested.

The motivations for the development of open data can be divided into three main categories: political—transparency, accountability, and involvement of a society; economic—cost-saving and new opportunities; technical—development in storage, web adoption, and standardization.

### *2.2.1 Open Data Definition*

Definitions of open data usually point to a number of criteria or principles. For example, according to OECD (2005), data is considered open if: (a) access is granted on equal or non-discriminatory terms, and (b) access costs do not exceed the marginal cost of dissemination. Manyika et al. (2013) defined open data as "the release of information by government and private institutions and the sharing of private data to enable insights across industries." According to the definition by the UK government, data is open if it meets the following criteria (Open Data White Paper: Unleashing the Potential, 2012): (a) accessible (ideally via the Internet) at no more than the cost of reproduction, without limitations based on user identity or intent; (b) in a digital, machine readable format for interoperation with other data; and (c) free of restriction on use or redistribution in its licensing conditions.

According to the already mentioned Open Definition, to be truly open, data should be:[3]

- available online to accommodate the widest practical range of users and uses
- open-licensed so that anyone has permission to use and reuse the data
- machine-readable so that large datasets can be analyzed efficiently
- available in bulk so that it can be downloaded as one dataset and easily analyzed by a machine
- free of charge so that anyone can access it no matter their budget.

The central term in the above definitions is non-discriminatory access. Access can be understood in different dimensions: technological barriers, intellectual property rights, and pricing. These three factors determine various degrees of openness, as presented in Fig. 2.1.

[3] http://opendefinition.org/.

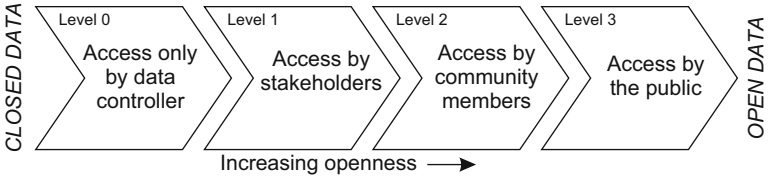



**Fig. 2.1** Data common continuum. Source: based on (OECD, 2015)

The openness is sometimes referred to as liquidity. Datasets can range from completely closed to completely open. Chui et al. (2014) evaluated the liquidity along four criteria: accessibility, machine readability, cost, and rights. They characterize the datasets as follows:

- Completely closed—Data can be accessed only by a limited number of individuals or organizations. They use formats that are not easily retrievable and processable by computers. Moreover, if data is made available, the price is usually high. It is forbidden to reuse, republish, or distribute data.
- Completely open—Everyone can access data. It is also available in formats that can be easily retrieved and processed by computers. No fee is required, and there are no limitations on reuse and redistribution.

Accessibility is not necessarily determined by a data source, i.e., who makes data available. Open data can come from individuals, companies, or governments. Figure 2.2 presents various kinds of data, based on two dimensions: accessibility and data source. Considering personal data, Chui et al. (2014) introduced the notion

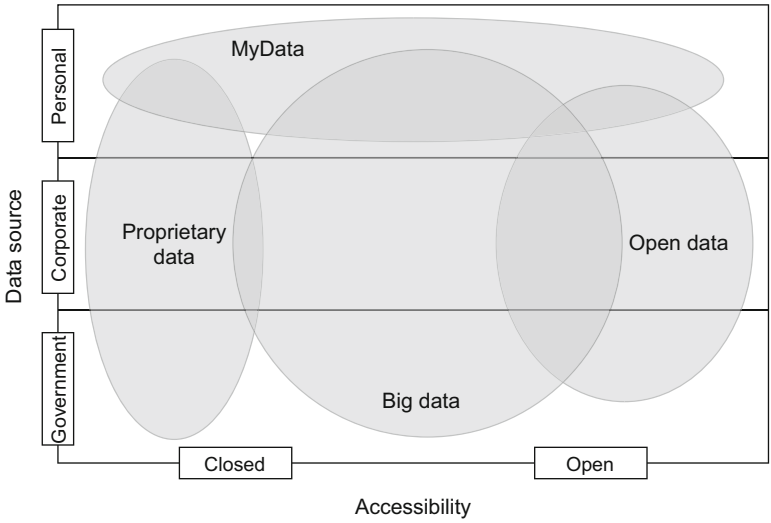



**Fig. 2.2** Data sources vs. accessibility. Source: based on (Chui et al., 2014)

of *MyData*, which is a concept of sharing information collected about an individual with that person. MyData is considered an important factor for leveraging the open data opportunity, especially in education and healthcare. Setting MyData alongside aggregate statistics offers an interesting benchmarking opportunity for individual consumers, for example, some utility companies can show the energy consumption of a consumer compared to his neighbors to encourage conservation (Chui et al., 2014).

Another spectrum, besides the two presented above, was also proposed in (Enabling the Data Revolution, 2015): closed data, shared data, and open data. *Closed data* is data that is foreseen to be used only internally. Personal data is a common example for this category. Many valuable datasets can be derived from data about individuals, but care must be taken to enforce the privacy rights. Personal data must be managed in accordance with best practices and regulations.[4] *Shared data* is data that can be made available to specific entities (people, companies) for particular kinds of reuse. For example, a telecom company can offer researchers access to pseudonymized call detail records to allow them to build epidemiology models. Governments may also share data aggregated from national statistics. Finally, *open data* is data that is provided for unrestricted reuse. Such data is usually released by governments. Data should be available online, in a machine-readable format. Additionally, the license should state that the data is free to be reused by anyone.

### 2.2.2 Open Government Data Principles

Open government data is open data made available by a government. The relation between open data, public sector information, as well as big data, which will be discussed later, is depicted in Fig. 2.3. As the majority of open datasets are offered by governments, it makes sense to review some principles and guidelines to supplement the definition of open data.

In December 2007, during a meeting in California, 30 open government advocates developed a set of eight principles of open government data. Government data can be considered open if it complies with Open Government Data Principles (2007), which is as follows:

1. *Complete*. All public data concerning specific phenomena should be made public. There should be no restriction on availability except when datasets are protected by other regulations, e.g., security, privacy.
2. *Primary*. Data should be made available as collected at the source, with the highest possible level of granularity. It should not be aggregated or otherwise transformed. This task is left to consumers.

[4] For example (GDPR, 2016).

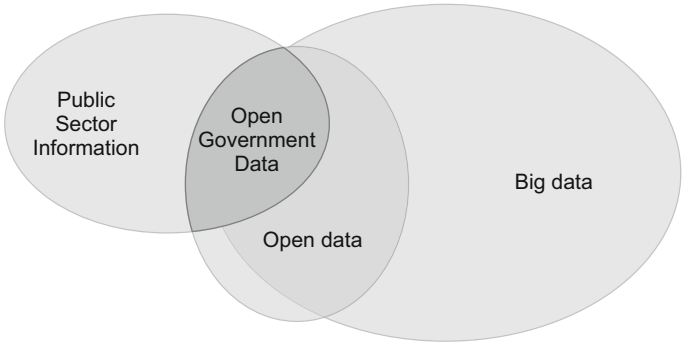



**Fig. 2.3** Boundaries of open data and public sector information. Source: (European Data Portal, 2015)

3. *Timely*. Data should be made available as quickly as necessary to preserve the value of data.
4. *Accessible*. Data should be made available for easy consumption by anyone for any purpose. This also puts some requirements on technical aspects, like publishing in a non-proprietary format.
5. *Machine processable*. Data should be reasonably structured so that automated processing is possible. Scanned documents are not a proper way to publish data, and much more useful are CSV, XML, or XML formats.
6. *Non-discriminatory*. Data should be available to anyone, even anonymously, i.e., without registration.
7. *Non-proprietary*. Data should be available in an open format so that no special software is required.
8. *License-free*. Data should not be subject to any copyright, patent, trademark, or trade secret regulation. The manner and purpose of data reuse should not affect the right to use data.

Another important document containing open data principles is the Open Data Charter, signed in July 2013 by G8 leaders. Initially, the charter contained five core open data principles. It was then heavily discussed and refined within a broad participatory process, engaging the government and civil society.[5] In its current version, ODC (2015) mentions six open data principles:

1. *Open by default*. Government data is of significant value to a society and the economy; therefore, it should be open by default. In the process of opening, citizens' rights to privacy should be respected. When there are legitimate reasons that data cannot be released, it should be justified.
2. *Timely and comprehensive*. Preparation of data for release requires time. Still, open data should be released in a timely manner, without undue delay, and

[5] http://opendatacon.org/toward-an-international-open-data-charter/.

the publication schedule should be known in advance. Data should be comprehensive, accurate, disaggregated, and of high quality, thus valuable for entities planning to use it.
3. *Accessible and usable*. The released data should be easily discoverable and accessible without additional administrative burden. It should help in making better informed decisions, fulfilling the requirement of usability.
4. *Comparable and interoperable*. The released data should be presented in structured and standardized formats so that it is possible to combine it with internal data. Datasets should be properly described and documented. Where possible, common identifiers should be used, and reference to emerging globally agreed standards should be made.
5. *For improved governance and citizen engagement*. Open data should be a foundation for transparency and accountability of governments. By civic participation better policies can be developed, and thus the provision of public services can be enhanced.
6. *For inclusive development and innovation*. The openness is important for stimulating creativity and innovation. The more entities use open data, the greater social and economic benefits are generated. The digital divide should be addressed so that everybody can take advantage of open data.

The international Open Data Charter is gradually being adopted by governments. It provides a common foundation to realize the full potential of open data by implementing the above six principles, thus fulfilling the following mission: "The overarching goal is to foster greater coherence and collaboration for the increased adoption and implementation of shared open data principles, standards, and good practices across sectors around the world." The major expected benefits are stimulation of innovation, increased transparency and the creation of efficiencies in how public money is spent.

Within this initiative, two reports were developed: (Enabling the Data Revolution, 2015) and (Walker & Perini, 2016). Here we could observe a change of focus and further progress of the open data movement. Opening data was just a first step, but then building communities became a priority, and an increasingly sophisticated network of communities started to make data useful in addressing social and economic challenges. Open data is central to many commitments of world leaders, including Sustainable Development Goals (SDGs), the Paris Climate Agreement, and the G20 Anti-Corruption Data Principles. Open data is also an increasingly local issue. Many cities and regional governments implement open data principles to achieve transparency, economic growth, and service delivery in close collaboration with citizens (Walker & Perini, 2016).

### 2.2.3 *Open Data as Infrastructure*

Communities working with open data are shifting their attention from getting access to data to constructing new open data infrastructures.[6] "Just like physical infrastructures such as roads, rail, power lines, and telecom networks that powered development over the last 100 years, datasets have become a crucial part of our modern infrastructure" (Enabling the Data Revolution, 2015). "Data has become a key infrastructure for twenty-first century knowledge economies" (OECD, 2015). There are a lot of programs that work on establishing standards and facilitating the exchange of tools and ideas (see Sect. 2.3). Data infrastructure, as any infrastructure, can be built properly or improperly. Good data infrastructure should care about quality, sustainability, and accessibility. Bad data infrastructure can be recognized by lack of organization, disparate datasets, or barriers excluding certain groups of users (Enabling the Data Revolution, 2015).

Data is also compared to oil, but this comparison stirs the debate. The following is a statement by Neelie Kroes, European Commissioner for Digital Agenda (in office 2010–2014):

> Data really is the new oil. Data is a raw material for information businesses, just as oil is a raw material for fuel and plastics businesses. Data is also everywhere, it is cheap, and it can deliver huge rewards both in terms of services and financial returns.

According to Guo (2014), data is the new natural resource of the twenty-first century.[7] It can be harnessed effectively and used to improve the overall quality of life. According to OECD (2015), data cannot be the 'new oil' as it cannot run out. It is rather a capital good that can be used by a society for various productive purposes. In contrast to oil, the use of data does not exhaust the supply of data. It is a non-rivalrous good and has the potential to meet the demands of all entities interested in the data (OECD, 2015).

## 2.3 Open Data Initiatives

The recent development of open data has its origins in government decisions. It had to be started on a political level, because a government is "the single largest collector, user, holder, and producer of information about citizens, organizations, or public service delivery" (Gonzalez-Zapata & Heeks, 2015). Proper regulations and good examples spurred a plethora of follow-up activities, and interest in the digital government concept is still evolving. Janowski (2015) proposed the Digital Government Evolution Model with four increasingly complex phases in

[6] For detailed discussion on infrastructures, see Sect. 5.2.4.

[7] Yike Guo is a professor in computing science and Director of the Data Science Institute, Imperial College London.

**Table 2.1** Digital government evolution model

| | Application context | Features | | |
|---|---|---|---|---|
| Stage | Technology impacting | Internal government transformation | Transformation affects external relationships | Transformation is context-specific |
| Digitization | Government | No | No | No |
| Transformation | Government organization | Yes | No | No |
| Engagement | Government stakeholders | Yes | Yes | No |
| Contextualization | Sectors and communities | Yes | Yes | Yes |

Source: based on figure from (Janowski, 2015)

the evolution of the concept. Digitization concerns how technology is used in government. Transformation is about how technology impacts the organization of a government (e-government). Engagement extends the impact to government stakeholders, allowing electronic governance. Finally, contextualization makes it possible to impact sectors and communities so that the state achieved is policy-driven electronic governance. Table 2.1 summarizes the evolution model.

### *2.3.1 European Data Economy*

Releasing open data is part of the bigger plan of the European Commission. The first step is definitely data sharing. Neelie Kroes, as Commission Vice President, encouraged public administration to share data with the following statement (European Commission, 2011):

> We are sending a strong signal to administrations today. Your data is worth more if you give it away. So start releasing it now: use this framework to join other smart leaders who are already gaining from embracing open data. Taxpayers have already paid for this information, the least we can do is give it back to those who want to use it in new ways that help people and create jobs and growth.

One statement is particularly interesting—*value of data can be increased by sharing*. The opening of data should contribute to the development of a European data economy, which is part of a Digital Single Market (DSM) strategy (European Commission, 2017). The strategy is made up of three policy areas, also called 'pillars':[8]

[8] https://ec.europa.eu/priorities/digital-single-market_en.

- Improving access to digital goods and services—helping to make the EU's digital world a seamless and fair marketplace to buy and sell.
- An environment where digital networks and services can prosper—designing rules that match the pace of technology and support infrastructure development.
- Digital as a driver for growth—ensuring that Europe's economy, industry, and employment take full advantage of what digitalization offers.

The EU data economy was estimated at €272 billion in 2015 (annual growth of 5.6%) and could employ 7.4 million people by 2020. However, the EU is currently not using their full data potential. Establishment of the Digital Single Market could contribute €415 billion per year to the European economy. It is necessary to address barriers to free movement of data, whether localization, technical, or legal, and if existing restrictions are removed, gross domestic product (GDP) can gain up to €8 billion per year (Bauer et al., 2016). Data needs to be used to bring growth to the economy. Moreover, before it can be used, it also needs to be available and analyzed.

The next steps for achieving DSM for Europe concern, among others, the data economy, and several issues are raised here, e.g., localization of data, liability, and standardization. There is an interest in data valuable for research, innovation, and new business opportunities.

### *2.3.2 International Activities*

The biggest 'political' initiative is run by the United Nations. In the report "World That Counts", the United Nations Secretary-General's Independent Expert and Advisory Group on Data Revolution for Sustainable Development recommended to establish a United Nations World Data Forum on Sustainable Development Data (UN World Data Forum) (Independent Expert Advisory Group, 2014). The first such forum was held in January 2017 in South Africa.[9] The second took place in October 2018 in the United Arab Emirates. The UN World Data Forum is foreseen to be a vehicle for facilitating cooperation with various professional groups, such as data scientists, policy-makers, geospatial information managers, end-users, as well as civil society stakeholders.

Considering the community-driven initiatives, OD4D (Open Data for Development) claims to be "the leading global partnership to advance the creation of locally-driven and sustainable open data ecosystems around the world."[10] Its mission is to promote the local supply of quality open data. They have established various initiatives to monitor how leaders and innovators use data in government, civil society, and business. The flagship examples of impact monitoring are the Open Data Impact Map, the Open Data Barometer, and the Open Data Index.

[9] https://undataforum.org/.

[10] http://od4d.net/.

The Open Data Impact Map[11] is a public database of organizations that use open data from around the world. It has been developed by the Center for Open Data Enterprise (CODE[12]) in cooperation with OD4D. As of October 2017, it listed 1777 organizations from 96 countries. It is an interesting source of inspiration on how open data can be used, and data about use cases is collected manually by a network of supporters, assuring the high quality of this database.

CODE maintains the map but also prepares periodic reports. Center for Open Data Enterprise (2016) presented findings from 1534 open data use cases across 87 countries, and the main findings are as follows. Open data is gaining importance as a resource for greater public participation. Higher income countries tend to have a greater share of the private sector in the use of open data. The most common sectors using open data are: governance, data/information technology, and research and consulting. Five sectors account for over half of the organizations using open data founded in the last decade. The most common data types used are government operations, geospatial, demographic and social, and weather data.

Four types of data use (Center for Open Data Enterprise, 2016):

1. Organizational optimization: efficiency gains, market intelligence. Large organizations typically use open data to optimize their operations.
2. Development of new products and services: data as a resource, data as information and analytics. New open data products and services are developed mostly by data/information technology and geospatial companies.
3. Advocacy: efficient allocation of resources, better policy making.
4. Research: industry research, business, investigation, and data journalism.

The Open Data Barometer (ODB[13]) is another initiative of OD4D and is maintained by the World Wide Web Foundation (WWWF[14]). It aims to uncover the impact of open data activities in various countries. The latest third edition compared the contextual data, technical assessments, and secondary indicators in 115 countries. The Barometer ranks countries based on: (a) readiness for open data initiatives, (b) implementation of open data programs, and (c) impact that open data is having on business, politics, and civil society. The data is available online,[15] and as of October 2021, there were 115 jurisdictions represented in the Open Data Barometer.

The Open Data Index (ODI[16]) is an effort to rank countries according to their maturity in open data development. It has a similar objective to the Barometer but is maintained by Open Knowledge.[17] These two projects differ in methodology

[11] http://opendataimpactmap.org/.

[12] http://www.opendataenterprise.org/index.html.

[13] http://opendatabarometer.org/barometer/.

[14] https://webfoundation.org/.

[15] http://opendatabarometer.org/3rdEdition/data/.

[16] http://index.okfn.org/.

[17] https://okfn.org/.

and data used. The Open Data Index measures the state of open government data through crowdsourced surveys to assess the openness of specific government datasets. Datasets are assigned to one of 15 categories, and ODI takes a citizen's perspective. There were 94 countries represented in the recent Open Data Index published in 2016.

## 2.4 Open Data Supply

Open data can be supplied by different kinds of entities. The main sources of open data (or external data) are: public sector activities, private sector activities, individual activities, and environment description and research (Buchholtz et al., 2014). They provide data to alleviate the data deficit by closing one or more possible gaps (World Economic Forum, 2015). The 'data gap' is data needed by the development community that simply does not exist. The 'access gap' is data that cannot be accessed due to a lack of capacity, resources, or agreements. The 'governance gap' is the absence of legal, ethical, and regulatory frameworks that regulate the use of data. Finally, the 'usability gap' is when data is available, collected, but it is not used for making decisions. We further discuss the various open data supply aspects in relation to the above gaps.

### *2.4.1 Obligation to Provide Information*

There are certain regulations that oblige authorities to provide public information. They are defined on the European as well as on a national level.

The right for information is a foundation of democracy. Electors have the right to be informed about the activities of authorities, especially the ones that influence the daily life of electors. Restriction of access to information should only be enforced in certain conditions, for example, concerning safety of the country or economic interest.

In principle, the institutions of the European Union do not have competencies to harmonize and unify the laws of Member States in the area of ensuring the right of access to public information. Areas of competency are defined in Art. 5 of the European Treaty (EU, 2008). Art. 11 only empowers institutions to regulate access to documents provided by themselves. Therefore, there are no binding regulations concerning access to public information. However, EU institutions have managed to enact legislation that is within the powers conferred and which is obligatory for Member States.

Such an act is Directive 2003/98/EC on the reuse of public sector information (EU, 2003). It is based on Art. 26 of EU Treaty, which regulates the internal market. "The Union shall adopt measures with the aim of establishing or ensuring the functioning of the internal market, in accordance with the relevant provisions of

the Treaties" (EU, 2008, Art. 26). This directive has an economical goal to facilitate the development of innovative services and free exchange of market information.

The directive concerns documents that are possessed by public administration institutions and how such documents can be reused both for commercial and non-profit purposes. It is forbidden to discriminate similar categories of information reuse, and public institutions should also avoid making exclusive agreements for using certain documents.

Member States should, according to the standards defined in the directive, specify:

- the form in which documents are made available,
- the scope of obligation to process the data,
- the maximum prices for allowing access to a document,
- the possibility to introduce conditions to access documents by means of licensing (provenance, responsibility, proper use of documents, guarantee of non-alteration, obligation to refer to a source).

The European Commission, in 2013, amended Directive 2003/98/EC with Directive 2013/37/EU (EU, 2013) that introduced a general rule, according to which all documents made available by public administration institutions can be reused for any purpose: commercial or non-profit, unless they are protected by the intellectual property rights of third parties. Moreover, most data should be made available for free, except for in a few well-justified cases. Public administration institutions should not charge fees higher than the cost of processing the application forms (and such costs shall be negligible). According to the directive, data should be available in a format that makes it easy to reuse, so no proprietary formats shall be used. In order to monitor the reuse of information, a regulatory institution should be established. It is also important that the directive classified additional resources as open data—the scope of the directive was extended to libraries, museums, and archives.

### *2.4.2 Open Government Data Publishing*

Open data can also be supplied by a government on a voluntary basis. Attard et al. (2015) outlined three main reasons for opening government data: transparency, releasing value, and participatory governance. *Transparency* means that citizens and other stakeholders can monitor government initiatives and their legitimacy. Not only is access to the data necessary, but the interested entities should also be able to use, reuse, and distribute data, as transparency increases citizen social control. *Releasing social and commercial value* stems from the possibility to use the data for a number of purposes which are different from the ones originally envisaged. A government is the largest producer and collector of data in many domains (Alexopoulos et al., 2014). The publishing of open data encourages stakeholders to expand upon it (for example, to create new services), and therefore, all data can have a social and commercial value. *Participatory governance* means that citizens can be actively

involved in governance processes, e.g., make decisions or contribute to policies. This is in opposition to a traditional approach where citizens express their interest in elections only occasionally.

Taking the technical perspective, the W3C eGov Interest Group developed a set of steps for publishing open government data. They urge the use of standards and methodologies when publishing government data so that it can be easily used by the society (Attard et al., 2015):

1. Identify—data can be found and consumed more easily if we use permanent, patterned, and discoverable URIs.
2. Document—documentation helps to understand data as well as facilitates data discovery. Some formats, like XML or RDF, are self-documenting.
3. Link—data should be linked to other data and documentation, providing context.
4. Preserve—datasets should be versioned so that tracking of changes is possible. Users can also refer to a specific version when linking or citing. Versioning also allows the documentation of changes between versions.
5. Expose interfaces—published data should be both human- and machine-readable. For this purpose, interfaces can be designed. It is preferred to separate published data and interfaces, and external parties should have direct access to raw data.
6. Create standard names/URIs for all government objects—each object should have a unique identifier. This reduces ambiguity and improves discoverability as well as reuse.

Data is published by the government using specialized data portals. Such a portal is based on a dedicated software that makes it easy to work with datasets in various formats. Data should be available in non-proprietary formats. For tabular data, CSV and JSON are preferred, and for spatial data—Shapefile (ESRI, 1998) and GeoJSON. Both JSON and GeoJSON are often made available via API. The supplementary file formats are XML and Excel for tabular data, and GML and KML for spatial data. The most popular data portal solutions are CKAN and Socrata.

CKAN[18] is an open-source project, developed by Open Knowledge. It is usually self-hosted and provides tools for publishing, sharing, finding, and using data. Users can use its faceted search features to browse and find the data they need. Data can be previewed using tables, charts, and maps. The solution is used by `data.gov` in US, the government of Canada, and the European Data Portal, among others.

Socrata[19] is a software-as-a-service platform that provides a cloud-based solution for open data publishing and visualization. Sometimes it is also referred to as a data-as-a-service platform. Customer sites are hosted on Amazon Web Services and the dedicated product is Publica Open Data,[20] which focuses on three aspects: *discoverability*—data from government publishers should be easy to find, *context*—

[18] http://ckan.org.

[19] http://socrata.com.

[20] https://socrata.com/publica-open-data/.

using visualizations, data can be framed within a narrative explaining the details of government activities, and *sharing*—information can be shared through social media or embedded in other websites.

Both CKAN and Socrata are often integrated in various ways with content management systems. For example, DataPress[21] is an integration of WordPress and CKAN and is used by `data.gov`, as well as The Guardian, Telegraph, and ODINE. DKAN[22] is an open-source Drupal plugin and integrates CKAN features into Drupal. It is used by `data.gov.uk` and almost 90 other sites around the world.

Data is best made available with API so that machine-to-machine interaction is possible. Besides JSON, there is also another standard emerging for data-services—OData. OData[23] (Open Data Protocol) is an ISO/IEC approved, OASIS standard that defines a set of best practices for building and consuming RESTful APIs. Its goal is to simplify the querying and sharing of data between multiple stakeholders across disparate applications and platforms (enterprise, mobile, and cloud). The standard, in version 4.0 as of October 2017, consists of three parts: protocol, URL conventions, and Common Schema Definition Language (CSDL) as an OData modeling language.

Sometimes open data supply cannot be realized, and the full potential of open data is not achieved due to many barriers. These include technical, legal, economic, organizational, and cultural issues (Conradie & Choenni, 2012).

One of the technical obstacles is the heterogeneous nature of data formats used by public administration, ranging from document scans through PDF files, through CSV or Excel files to better structured XML files (Attard et al., 2015). This issue concerns both data providers and data consumers. Even though data is available in a structured form, there may be differences in the data schemata, and thus the aggregation and analysis of data is hindered. Moreover, partly to cope with these issues, there is a variety of tools, which further makes the access to data problematic.

An example of the legal barrier for data opening are regulations concerning personal data. Indeed, not all data should be open, as usage of social networks or mobile phone activities provides a lot of indicators about individuals (see Sect. 6.5, page 163).

UN Global Pulse (2012) introduced the concept of "data philanthropy" (cf. open algorithm in Sect. 6.4.3, page 161). In this concept, the private sector shares own data to address certain commercial or societal challenges. Data philanthropy can have two flavors: data commons—some data is shared publicly after adequate anonymization and aggregation; digital smoke signals—where sensitive data is analyzed by companies, but the results are shared with governments.

[21] https://datapress.com/.

[22] https://getdkan.org/.

[23] http://www.odata.org/.

## 2.5 Adoption of Open Data

Regarding the adoption of linked data, Tinholt (2013) pointed at the importance of user involvement. Potential consumers need to be encouraged to use open data. The intent should be to motivate users to actively search for more datasets that can be useful for addressing their problems.

One of the issues about the adoption of open data is the discrepancy between the demand for data and its supply (more on the topic, cf. Sect. 6.2.2). Unsatisfied demand decreases the efficiency of the economy as a whole. Yang et al. (2015) noticed that only 10% of the datasets account for 90% of network traffic load. Public bodies, which observe low interest in their data, are not particularly interested in providing more data, and their engagement in the adoption of open data may be reduced. We can assume that there are also consumers who are not able to find the required data. Such users are also not interested in the adoption of open data.

Zuiderwijk et al. (2015) investigated several factors related to the intention to use and accept open data technologies. The hypotheses, along with their verification status, are enumerated below:

1. Performance expectancy is *positively* related to the behavioral intention to use and accept open data technologies (H1). The hypothesis is supported ($p < 0.001$).
2. Social influence is *positively* related to the behavioral intention to use and accept open data technologies (H3). The hypothesis is supported ($p < 0.001$).
3. Effort expectancy is *negatively* related to the behavioral intention to use and accept open data technologies (H2). The hypothesis is supported ($p < 0.005$).
4. Voluntariness of use is *negatively* related to the behavioral intention to use and accept open data technologies (H5). The hypothesis is supported ($p < 0.005$).
5. Facilitating conditions are *positively* related to the behavioral intention to use and accept open data technologies (H4). The hypothesis is not supported ($p > 0.005$).

To summarize the findings, the expected increase in performance and social pressure from other people or companies is positively related to the intention to use open data. The expected effort necessary to use open data is negatively related to the intention to use open data. Additionally, the more voluntary the use of open data is, the lower the intention to use open data. Hypothesis H4 would be interesting, but unfortunately, it is not supported. Combined with H5, certain regulations are necessary to increase the use of open data. Nevertheless, Zuiderwijk et al. (2015) confirmed that their model is not satisfactory, as it explains only 45% of the variance, which means that a large part of the variance in the use of open data has not yet been explained.

Particularly critical about the adoption of open data were Dawes et al. (2016), who complained about the small number of sustained commercial applications. The successful applications were created in government-sponsored application contests or challenges. Open data programs were predominantly evaluated using only simple

measures, e.g., the number of participating governments or the number of datasets released or downloaded. Loutas et al. (2012) earlier found out that the majority of open government data applications and services were built by individuals, freelancers, and researchers mainly for mobile devices using a single static dataset. They are offered for free, so no commercial value can be attributed.

In the following sections, we discuss why open government data is not always successful and provide some reasons for this.

### *2.5.1 Open Data Complexity*

A higher complexity of data complicates its use and challenges the process of finding patterns and trends in large amounts of data (Zurada & Karwowski, 2011). Yang et al. (2015) explored the complexity of open data initiatives from four perspectives: technology, organization, legislation and policy, and the environmental context. For each perspective, they identified influential factors of open data initiatives.

The technological perspective concerns data format and metadata, information system outsourcing, and level of informatization. The first area should answer the question if data is available in a structured and machine-readable format within organizations. Moreover, we should know if there is a unified data format and a single metadata schema. There should be a strategy about what data to open and what metadata to adopt. The second area analyzes the outsourcing potential. In many government agencies, outsourcing can be the assumed strategy to develop information systems. Government bodies usually lack information by professionals; therefore, outsourcing contractors become important players in opening data. Finally, in the area of informatization, we found that government agencies are not equally advanced in IT development. In some institutions, paper-based processes can still be used and the data to be made available is not even digitized. Various LOD-related standards require higher-level computing skills.

The organizational perspective was split into: organizational culture, authority involvement, perceived effort, perceived liability, perceived benefit, and perceived loss. Organizational culture answers the question if government bodies have the vision how to use open data. Open data is not a core business of government agencies; therefore, it is important that people believe in the mutual benefit of open data. Agencies are usually conservative, and people are reluctant to change. The involvement of higher-level authorities is critical to the success of open data initiatives, and support should also be offered from central to local governments. The implementation of open data necessitates additional effort. First, government agencies need to dedicate time and resources to identify and prepare datasets. Second, there may be an increase in demand for government services stemming from the use of data by more aware consumers. In the area of perceived liability, we need to accept that agencies are afraid of being liable for losses incurred from the use of data of poor quality. They also have concerns about data misuse, where a

negative effect of opening can be observed. Finally, opening data can lead to privacy infringement by exposing personal data. No less important are the areas of perceived benefit and loss. Usually, there is little or no incentive for government agencies to participate in open data initiatives, and many agencies may perceive data as part of their assets and therefore may be reluctant to publish this data. Nevertheless, government agencies need to identify interesting datasets to distribute the budget efficiently. The value of particular open datasets has to be judged by users, while some agencies also charge fees for making data available to requesting users.

The legislation and policy perspective emphasizes that before data is made available, it has to be checked against many regulations, including acts concerning the public sector, copyright law, and personal information. Early regulations may prevent agencies from data opening. Another important issue is the design of a good licensing strategy: regulations should protect public bodies while encouraging data reuse by consumers.

Finally, the environmental perspective encompasses additional effects influencing the use of open data. Media and public opinion is an important factor in deciding which datasets should be open and which should be closed. Open data advocates can influence government agencies in making specific decisions, and pressure for opening can not only be felt from civil society but also from other governments (the peer effect). Government agencies can benchmark each other and therefore may be willing to open more data to stay in a leading position (the driving effect).

Based on their research in Taiwan, Yang et al. (2015) discovered that legislation and policy have the most significant impact on agencies' participation in open data initiatives, and technological factors turned out to be relatively easy to tackle.

### *2.5.2 Barriers for Adoption*

Janssen (2011) discussed the role of the European Directive on the reuse of public sector information (EU, 2003) in shaping the trend towards opening up government data. She identified the confusion between information access and reuse. The impact of the PSI directive remained unclear and might be more limited than previously assumed. What did work was combining PSI with freedom of information access legislation.

Besides public organizations, there are also semi-public organizations, such as cultural heritage foundations and public transport organizations, which also need to observe similar regulations. The motivation of such organizations for opening data is different, as they also aim at realizing commercial gains with their data. They need to balance public and commercial goals. Such organizations were studied, among others, by van Veenstra and van den Broek (2013). They reviewed literature and conducted a case study of open data in a semi-public organization in the Netherlands to identify drivers, enablers, and barriers of open data. The findings are presented in Table 2.2.

**Table 2.2** Summary of organizational drivers, enablers and barriers to open data

| | Drivers | Enablers | Barriers |
|---|---|---|---|
| Information technology | Linked data | • Usefulness of the databases<br>• Discoverability of the data | • Poor data structures<br>• Legacy systems<br>• Fragmented databases<br>• Limited data quality<br>• Lack of standardization |
| Organizational and managerial | Efficiency and budget cuts | • Data stewardship<br>• Clear implementation strategy | • Complexity of the changes to be made<br>• Lack of business case for generating revenue from reuse<br>• Embedding open data in the strategy and work processes |
| Legal and regulatory | • PSI directive<br>• Law enforcement | | • Privacy and data protection<br>• National security |
| Institutional and environmental | • Transparency and accountability<br>• Enabling reuse | • Political leadership<br>• Value for users | • Closed culture of government<br>• Lack of support of user feedback |

Source: (van Veenstra & van den Broek, 2013)

Countries can be at different levels of development regarding opening data. For example, Tinholt (2013) classified countries into the following groups: trend setters, followers, and beginners. Countries within the most advanced group put emphasis on releasing extensive amounts of data and updating it at regular intervals. Data is characterized by a significant breadth and granularity. Moreover, blogs and the possibility to discuss dataset-related issues drive engagement among users. World Economic Forum (2015) went even further and explicitly divided countries into two groups: Global North and Global South. The first experiences 'data deluge' Margetts (2014), and the latter is affected by a relative 'data deficit,' because there are significant constraints on the creation, collection, and use of data.

Not all countries or jurisdictions are successful. Various initiatives may not be adopted because stakeholders see no added value in open data. Janssen et al. (2012) distinguished several categories of adoption barriers that hamper the publication of data. The first category concerned institutional barriers: lack of policy, resources or knowledge, questionable quality of data, and no added value of published data. The second category stressed the complexity of the task: data cannot be easily discovered, unknown meaning or quality, complex data formats, and no tool support.

Janssen et al. (2012) also compiled *five myths of open data*. The first myth suggests that the publicizing of data will automatically yield benefits. It is similar to the third myth, which states that obtaining effects is a matter of simply publishing public data. The reality is that giving access to data is not enough, and the publisher should also provide a means to process data. So far, too much attention was focused on the supplier side and user support was neglected. Publication should not happen without additional activities, e.g., improvement of quality, and public administration should also reduce the complexity of data before publication.

The second myth states that all information should be unrestrictedly publicized. On the contrary, the dataset selection for publication should be cautious. Personal data should not be publicized, and regulations may specify other forbidden datasets. Moreover, data collected for one purpose might not be reused for another purpose. Another limit on published datasets may be enforced by limited resources. Low-quality datasets should also be kept closed, as governments are accountable for published data, as data of worse quality means less transparency. Finally, some organizations make profit from publishing data, and opening requires changes in their business models.

The fourth myth asserts that each citizen can make use of open data. Governments assume that potential consumers of their data have the necessary resources, expertise, and capabilities to use the data. Sometimes a deeper understanding of data is necessary, e.g., knowledge of the relations. Although data is available to anybody, knowledge is not; time and effort may be necessary to acquire it. Sometimes more sophisticated statistical techniques have to be used, and a phrase attributed to Herbert G. Wells "statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write"[24] is particularly true for open data. Only people with appropriate skills are able to make sense of data and understand the implications. Browsing single open datasets is supported by tools, but linking and combining data by users requires sophisticated knowledge. There is a risk that "opening data might further contribute to the digital divide, as the use of data might be limited to certain groups" (Janssen et al., 2012).

The last myth is related to the belief that disclosure of data will automatically result in an accountable and transparent government, and data analyzed by various parties may indeed result in different conclusions. A problem with the consistency of open data sources may also be observed. System theory requires a feedback mechanism in order to close the loop between the government and citizens. Concluding, open government is not created by only opening data.

The difference between opening data and simply making it available is also referred to as 'openwashing' (Villum, 2014). Related to government, this term means "providing selective information without having an environment where citizens can freely use that data" (Brandusescu, 2016). In a broader sense, data publishers can claim their data as open even though it is not, i.e., it does not meet the full open criteria, e.g., by restricted use. Openwashing can also occur when 'vacuous data' is published, i.e., data not relevant to anybody, and which is incorrect or outdated. Brandusescu (2016) is also skeptical about the impact of open data—there have been case studies published claiming the success and impact of open government data, while the impact can be observed over a longer period.

A more fundamental problem than openwashing is openwishing. This refers to big claims concerning the social and individual impact of open data (Dickinson,

[24] Quote from the presidential address of Samuel S. Wilks to the American Statistical Association, who was paraphrasing H.G. Wells from his book *Mankind in the Making*. Found in (Wilks, 1951).

2016). Open data projects should not overpromise, otherwise they provide arguments for undermining the whole idea of open data.

Speaking of the adoption, we also need to consider some risks related to open data—from misinterpretation of data to violating privacy. Zuiderwijk and Janssen (2014) conducted interviews with public sector officials and data archivists to identify the negative effects of open data. They have identified sixteen categories of effects grouped into three classes: the dark side of open data, implementation of open data efforts, and management of open data efforts.

The first group of negative effects concerns open data efforts. There may be difficulties with data ownership, or opening data may violate legislation in another way, e.g., privacy can be violated. Moreover, published data can be biased, or data can be misinterpreted, and thus data may be misused due to poor quality. One of the arguments for opening is transparency, but this transparency may have negative consequences for the government. Finally, there may be problems with the timely publishing of data.

The second group is related to the implementation of open data efforts. Opening data should be well planned. Currently, technology allows for the rapid publishing of data, but the focus should be on institutionalizing the opening of data. There is also a risk of little attention being paid to public value and solving societal problems. Certain data might be hard to find. There is also unclear responsibility for open data, for example, who is responsible for incorrect use or interpretation of data or low quality. Finally, citizens may be the main beneficiaries of open data because data usually does not answer the questions of ordinary people.

The third group is about the dark side related to the management of open data efforts. For example, resources may be wasted because of the publishing of invaluable data. Data publication is not prioritized, and it is very often difficult to identify the data publication policy.

Summarizing, the dark side can be the motivation for risk-adverse decision-making concerning data opening.

## 2.6 Macroeconomic Information

A significant part of open data published by governments and international organizations concerns economic indicators and demographics (Tambouris et al., 2015). Generally, government data can be divided into the following groups (Deloitte, 2016):

- Statistical data—raw data collected directly or indirectly by public entities. It is processed by national statistical offices to provide a background for government decisions.
- Information about individuals and businesses—data collected as part of government activities and services. It usually contains information about the relation

between entities and the state. Certain parts of data can be released to the public to facilitate other business activities.
- Data about phenomena—data collected from various forms of automatic measurements, e.g., weather information, air pollution, and maps, as a result of data analysis.

### *2.6.1 Statistical Data*

Very popular among statistical data are global economy indicators. For example, the World Bank publishes World Development Indicators.[25] This is a collection of the current global development indicators, compiled from official international sources. There are over 800 indicators covering more than 150 economies. The collection includes accurate data on a national level and regional level, as well as global estimates. The DataBank[26] allows for the preparation of visualizations.

As this data is open, the same indicators are provided by a larger number of entities. Charts can be prepared in Google Public Data Explorer,[27] which offers access to over 100 datasets.[28] Besides World Development Indicators,[29] it also contains Human Development Indicators,[30] Global Competitiveness Report,[31] and OECD Factbook.[32]

Statistical data is often published as data cubes using RDF Data Cube Vocabulary (Cyganiak & Reynolds, 2014). The values contained in the cells described by measures are organized by dimensions (e.g., country, time).

A significant portion of open data concern statistics. For example, the OpenCube project[33] once estimated that 6875 out of 7682 (89.5%) datasets of the EU Open Data Portal were of a statistical nature.[34] Moreover, statistical data is often organized as data cubes—each cell contains a measure which is described by a number of dimensions (according to SDMX[35]).

In order to fulfill their tasks, analysts need to combine statistical data from multiple datasets. Unfortunately, the data is usually in different sources (files,

[25] https://data.worldbank.org/products/wdi.

[26] http://databank.worldbank.org/data/reports.aspx?source=world-development-indicators.

[27] https://www.google.com/publicdata/explore.

[28] Number of datasets is 113 as of October 2017.

[29] https://www.google.com/publicdata/explore?ds=d5bncppjof8f9_.

[30] https://www.google.pl/publicdata/explore?ds=ife8n327iup1s_.

[31] https://www.google.pl/publicdata/explore?ds=z6409butolt8la_.

[32] https://www.google.pl/publicdata/explore?ds=ltjib1m1uf3pf_.

[33] http://opencube-project.eu/.

[34] As of December 2014, according to https://www.slideshare.net/OpenCubeProject/orebro-2015-tambouris.

[35] Statistical Data and Metadata eXchange, https://sdmx.org/.

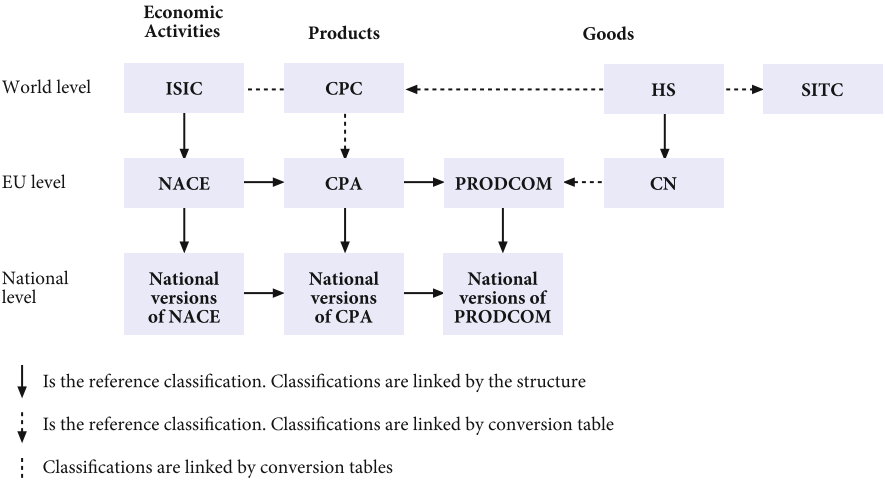



**Fig. 2.4** NACE in the context of world statistical systems. Source: (Eurostat, 2008, p. 13)

databases) and formats, placed in so-called data silos. This was addressed, among others, by the OpenCube project, which proposed a process for the integration of statistical data cubes.

### *2.6.2 Industry Classifications*

Statistical offices use codes to classify various phenomena and activities. For example, for the classification of activities, NACE (Nomenclature Statistique des activités économiques) is the standard in Europe. NACE is part of an integrated system of statistical classifications, developed mainly under the auspices of the United Nations Statistical Division. Figure 2.4 presents the relations between statistical systems on global, EU, and national levels. Detailed data is provided by RAMON, Eurostat's Metadata Server.[36]

In 1989, the Statistical Commission of the United Nations proposed a set of classifications for activities, goods, and services. The following are the main components (Eurostat, 2008):

- ISIC—the activity side of the system,[37]
- CPC—the central instrument for classifying goods and services,[38]

[36] http://ec.europa.eu/eurostat/ramon/index.cfm?TargetUrl=DSP_PUB_WELC.
[37] https://unstats.un.org/unsd/cr/registry/isic-4.asp.
[38] https://unstats.un.org/unsd/cr/registry/cpc-21.asp.

- SITC—the aggregated classification of transportable goods for international trade statistics for comparison,
- BECs—the classification of goods according to Broad Economic Categories for the purposes of economic analysis.

CPA, Statistical Classification of Products by Activity,[39] is the European version of the CPC. The classification is designed to categorize products that have common characteristics. CPA has a hierarchical structure with six levels, with 21 sections at the first level. Product categories are related to activities as defined by NACE. For customs tariffs, another classification is used—CN Combined Nomenclature.[40] For providing statistics by products, Prodcom[41] is used. Prodcom uses product codes and distinguishes about 3900 different types of manufactured products, identified by an 8-digit code. The first four digits are the classification of the producing company given by NACE, and the first six correspond to CPA. The remaining digits provide more details about the product. Most product codes correspond to one or more Combined Nomenclature (CN) codes (except for industrial services).

The classification crucial for public procurement in Europe is CPV, Common Procurement Vocabulary.[42] The recent version was adopted by Regulation (EC) No. 213/2008 and has been in use since September 2008 (hence the reference CPV2008). CPV is used to fill the notices of calls for contract and to search for business opportunities in TED, as well as in national public contract notices. It consists of a main vocabulary for defining the subject of a contract and a supplementary vocabulary for adding further qualitative information. The first two digits identify the divisions (XX000000-Y), followed by groups, classes, and categories. The use of CPV is mandatory in the European Union.

### *2.6.3 Open Geographical Data*

Open geographical data in Europe is regulated by the Inspire Directive (EU, 2007), which came into force in May 2007. The objective was to improve the availability, access, and sharing of spatial information across EU Member States. Implementation is divided into stages, and full implementation was required by 2019. Information under consideration was divided into 34 ‘themes’.

INSPIRE is based on a number of common principles.[43] First of all, data should be collected only once and kept where it can be maintained most effectively. It can then be made available to those who need it under a good governance process.

[39] http://ec.europa.eu/eurostat/web/cpa-2008.

[40] https://ec.europa.eu/taxation_customs/business/calculation-customs-duties/what-is-common-customs-tariff/combined-nomenclature_en.

[41] http://ec.europa.eu/eurostat/web/prodcom.

[42] https://simap.ted.europa.eu/cpv.

[43] http://inspire.ec.europa.eu/inspire-principles/9.

Spatial information, once represented in a common standard, can be combined across datasets and shared with many users and applications. Information collected at one level (e.g., city) can be shared with other levels (e.g., country). When necessary, the analyst should be able to move from a general level to a detailed level. Geographic information should be easy to find along with information about permitted use.

## 2.7 Summary

Open data is still a developing phenomenon. Thus far, the focus is on the government side, and phrases like 'transparency' and 'accountability' are prevailing. The principles of official statistics can be put into practice for all stakeholders of open data ecosystems.

Sometimes opening data can cause disappointment. First, citizens do not use data, and transparency is not always the primary goal of open government data; it can be 'stimulating innovation.' Second, datasets opened to the public are not necessarily improving the internal organizational processes of public bodies, especially when the feedback loop is missing. There is also significant criticism concerning the quality and accessibility of data.

It is necessary to close the open data capacity gap. Entities can be supported not only in data release but also in making use of data. This can result in better collaboration between governments, businesses, and civil society.

New incentives are necessary to catalyze private sector innovation with open data, particularly in developing countries. Governments and the private sector should discuss their priorities concerning opening datasets. Addressing the data deficit can strengthen the involvement of potential beneficiaries and the overall open data movement. Governments should avoid the temptation of openwashing—the number of datasets is not important, but their value for economy is.

Not only should public data be open, but commercial companies can also consider the benefits of opening up their own data. New services can be created based on open government data, where users add value to initially released data. Organizations can also combine open data with their internal data, developing new commercial products and generating even more value.

In order to generate new insights and extract more value from data, we need a cross-referenced analysis of different datasets, i.e., a linking.

## References

Alexopoulos, C., Zuiderwijk, A., Charapabidis, Y., Loukis, E., & Janssen, M. (2014). Designing a second generation of open data platforms: Integrating open data and social media. In *Electronic Government: 13th IFIP WG 8.5 International Conference EGOV 2014, Dublin,*

*Ireland, September 1–3, 2014. Proceedings* (pp. 230–241). Berlin Heidelberg: Springer. https://doi.org/10.1007/978-3-662-44426-9_19 (page 19)

Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. *Government Information Quarterly, 32*(4), 399–418. (pages 8, 19, 20, 21)

Bauer, M., Ferracane, M. F., Lee-Makiyama, H., & Van der Marel, E. (2016). *Unleashing internal data flows in the EU: An economic assessment of data localisation measures in the EU member states*. Research rep. 03/2016. ECIPE—European Centre for International Political Economy. (page 16)

Brandusescu, A. (2016). *#openwashing . . . anyone?* World Wide Web Foundation. https://webfoundation.org/2016/10/openwashing-anyone/ (visited on 2017-10-27). (page 26)

Buchholtz, S., Bukowski, M., & Śniegocki, A. (2014). *Big and open data in Europe. A growth engine or a missed opportunity?* (p. 114). Warsaw: demosEUROPA. ISBN: 978-83-925542-1-9. (page 18)

Center for Open Data Enterprise. (2016). *Open data impact map*. (page 17)

Chui, M., Farrell, D., & Jackson, K. (2014). How government can promote open data and help unleash over 3 million$ in economic value. In *Innovation in local government. Open data and information technology* (pp. 4–23). McKinsey&Company. (pages 10, 11)

Conradie, P., & Choenni, S. (2012). Exploring process barriers to release public sector information in local government. In *Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance - ICEGOV '12* (p. 5). ACM Press. ISBN: 978-14-503120-0-4. https://doi.org/10.1145/2463728.2463731 (page 21)

Cyganiak, R., & Reynolds, D. (Eds.) (2014). *The RDF data cube vocabulary*. *W3C recommendation*. https://www.w3.org/TR/vocab-data-cube/ (page 28)

Dawes, S. S., Vidiasova, L., & Parkhimovich, O. (2016). Planning and designing open government data programs: An ecosystem approach. *Government Information Quarterly, 33*(1), 15–27. ISSN: 0740-624X. https://doi.org/10.1016/j.giq.2016.01.003 (page 22)

Deloitte. (2013). *Market assessment of public sector information*. BIS/13/743. London: Department for Business Innovation & Skills. (page 7)

Deloitte. (2016). *The value of DDI (data driven innovation)*. (page 27)

Dickinson, A. (2016). *Openwishing data*. https://medium.com/@digidickinson/openwishing-data-bd33850c7b58 (visited on 2017-10-27). (page 27)

Enabling the Data Revolution. (2015). Conference Report. Open Data for Development. (pages 9, 11, 13, 14)

ESRI. (1998). *ESRI shapefile technical description*. https://www.esri.com/library/whitepapers/pdfs/shapefile.pdf (visited on 2017-10-26). (page 20)

EU. (2003). 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, 46*(L 345), 90–96. (pages 7, 18, 24)

EU. (2007). Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). *Official Journal of the European Union, L 108*, 1–14. (page 30)

EU. (2008). The treaty on the functioning of the european union. *Official Journal of the European Union, 51*(C 115), 47–199. (pages 18, 19)

EU. (2013). Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information. *Official Journal of the European Union, L 175*. (page 19)

European Commission. (2011). *Digital agenda: Turning government data into gold*. http://europa.eu/rapid/press-release_IP-11-1524_en.htm (visited on 2017-10-25). (page 15)

European Commission. (2017). *Commission outlines next steps towards a European data economy*. http://europa.eu/rapid/press-release_IP-17-5_en.htm (visited on 2017-10-25). (page 15)

European Data Portal. (2015). *Creating value through open data: Study on the impact of re-use of public data resources* (p. 112). Luxembourg: Publications Office of the European Union. ISBN: 978-92-79-52791-3. https://doi.org/10.2759/328101 (page 12)

Eurostat. (2008). NACE Rev. 2. Statistical classification of economic activites in the European Community. Luxembourg. (page 29)

GDPR. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC*. European Parliament. http://eur-lex.europa.eu/eli/reg/2016/679/oj (visited on 2017-09-27). (page 11)

Gonzalez-Zapata, F., & Heeks, R. (2015). The multiple meanings of open government data: Understanding different stakeholders and their perspectives. *Government Information Quarterly, 32*(4), 441–452. ISSN: 0740-624X. https://doi.org/10.1016/j.giq.2015.09.001 (page 14)

Guo, Y. (2014). *Data is the new oil or new soil?* Imperial College London. http://www3.imperial.ac.uk/newsandeventspggrp/imperialcollege/engineering/computing/eventssummary/event_28-2-2014-9-58-2 (visited on 2017-10-25). (page 14)

Independent Expert Advisory Group. (2014). *A world that counts. Mobilising the data revolution for sustainable development*. United Nations. (page 16)

Janowski, T. (2015). Digital government evolution: From transformation to contextualization. *Government Information Quarterly, 32*(3), 221–236. ISSN: 0740-624X. https://doi.org/10.1016/j.giq.2015.07.001 (pages 14, 15)

Janssen, K. (2011). The influence of the PSI directive on open government data: An overview of recent developments. *Government Information Quarterly, 28*(4), 446–456. ISSN: 0740624X. https://doi.org/10.1016/j.giq.2011.01.004 (page 24)

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management, 29*(4), 258–268. ISSN: 1058-0530. https://doi.org/10.108/10580530.2012.716740. arXiv: arXiv:1011.1669v3. (pages 25, 26)

Loutas, N., Varitimou, A., & Peristeras, V. (2012). Unraveling the mystery of Open Government Data Apps. In *Workshop on using open data: Policy modeling citizen empowerment, data journalism (PMOD 2012)*. (page 23)

Manyika, J., Chui, M., Groves, P., Farrell, D., Van Kuiken, S., & Doshi, E. A. (2013). *Open data: Unlocking innovation and performance with liquid information*. McKinsey Global Institute. (pages 7, 9)

Margetts, H. (2014). Data, data everywhere: Open data versus big data in the quest for transparency. In N. Bowles, J. Hamilton, & D. Levy (Eds.), *Transparency in politics and the media: Accountability and open government* (pp. 167–178). London: I.B. Tauris. ISBN: 978-1-78076-675-1. (page 25)

ODC. (2015). *Open data charter principles*. https://opendatacharter.net/principles/ (visited on 2017-09-27). (page 12)

OECD. (2005). *Principles and guidelines for access to research data from public funding*. Paris: OECD Publishing. (page 9)

OECD. (2015). *Data-driven innovation. Big data for growth and well-being* (pp. 1–456). ISBN: 9789264229358. https://doi.org/10.1787/9789264229358-en (pages 10, 14)

Open Data White Paper: Unleashing the Potential. (2012). London: UK Minister of State for the Cabinet Office. (pages 8, 9)

Open Government Data Principles. (2007). *Eight open government data principles* https://public.resource.org/8_principles.html (visited on 2017-09-27). (page 11)

Tambouris, E., Kalampokis, E., & Tarabanis, K. (2015). Processing linked open data cubes. In E. Tambouris, M. Janssen, H. J. Scholl, M. A. Wimmer, K. Tarabanis, M. Gascó, B. Klievink, I. Lindgren, P. Parycek (Eds.), *Proc. 14th IFIP WG 8.5 International Conference EGOV 2015, Thessaloniki, Greece* (pp. 130–143). Springer. ISBN: 978-3-319-22479-4. https://doi.org/10.1007/978-3-319-22479-4_10 (page 27)

Tinholt, D. (2013). *The open data economy. Unlocking economic value by opening government and public data*. Capgemini Consulting. (pages 22, 25)

UN Global Pulse. (2012). *Big data for development: Opportunities & challenges*. United Nations Global Pulse. (page 21)

US CIO Council. (2012). *Digital government: Building a 21st century platform to better serve the American people*. (page 8)

van Veenstra, A. F., & van den Broek, T. A. (2013). Opening moves-drivers, enablers and barriers of open data in a semi-public organization. In *Electronic Government: 12th IFIP WG 8.5 International Conference, EGOV 2013, Koblenz, Germany September 16–19, 2013. Proceedings* (Vol. 8074, pp. 50–61). LNCS. ISBN: 9783642403576. https://doi.org/10.1007/978-3-642-40358-3-5 (pages 24, 25)

Villum, C. (2014). *'Open-washing'—the difference between opening your data and simply making them available*. Open Knowledge. https://blog.okfn.org/2014/3/1/open-washing-the-difference-between-opening-your-data-and-simply-making-them-available/ (visited on 2017-10-27) (page 26)

Węcel, K., & Lewoniewski, W. (2015). Modelling the quality of attributes in wikipedia infoboxes. In W. Abramowicz W (Eds.), *Business information systems workshops* (Vol. 228, pp. 308–320). Lecture Notes in Business Information Processing. Springer. ISBN: 978-3-319-26761-6. https://doi.org/10.1007/978-3-319-26762-3_27 (page 8)

Walker, S., & Perini, F. (2016). *International open data roadmap*. Open Data for Development. (page 13)

Wilks, S. S. (1951). Undergraduate statistical education. *Journal of the American Statistical Association, 46*253, 1–18. (page 26)

World Economic Forum. (2015). *Data-driven development. Pathways for progress*. http://reports.weforum.org/data-driven-development/ (pages 8, 18, 25)

Yang, T. M., Lo, J., & Shiang, J. (2015). To open or not to open? Determinants of open government data. *Journal of Information Science, 41*(5), 596–612. ISSN: 0165-5515. https://doi.org/10.1177/0165551515586715 (pages 9, 22, 23, 24)

Zuiderwijk, A., & Janssen, M. (2014). The negative effects of open government data—investigating the dark side of open data. In *Proceedings of the 15th Annual International Conference on Digital Government Research* (pp. 147–152). ISBN: 978-1-4503-2901-9. https://doi.org/10.1145/2612733.2612761 (page 27)

Zuiderwijk, A., Janssen, M., & Dwivedi, Y. K. (2015). Acceptance and use predictors of open data technologies: Drawing upon the unified theory of acceptance and use of technology. *Government Information Quarterly, 32*(4), 429–440. ISSN: 0740-624X. https://doi.org/10.1016/j.giq.2015.09.005 (page 22)

Zurada, J., & Karwowski, W. (2011). Knowledge discovery through experiential learning from business and other contemporary data sources: A review and reappraisal. *Information Systems Management, 28*(3), 258–274. ISSN: 1058-0530. https://doi.org/10.1080/10580530.2010.493846 (page 23)

# Chapter 3
# Linked Data for Enrichment of Data Assets


*Data is more powerful in the presence of other data (Palmer, 2017)*

## 3.1 Introduction

In the context of open data, we can very often encounter a related concept—linked data. Linked data as a technical concept is relatively well defined. It is based on a simple data model—triples. A triple is a single fact consisting of a subject, a predicate, and an object. It follows the representation proposed within the Resource Description Framework (RDF) standard. As it borrows from web technologies, subjects and properties, and very often, objects are also expressed as Unified Resource Identifiers (URIs), resembling internet addresses.

To recognize some lesser-known aspects of linked data, we carried out a literature review. We searched for the phrase "linked data" in the full text, without any additional restrictions. Table 3.1 presents the number of articles containing the phrase in various digital libraries. As Google Scholar indexes research articles from most major academic publishers and repositories worldwide, ranks results by both relevance and quality, and returns the biggest number of results, we have decided to use this database for review purposes. We have Google Scholar to identify not only publications in renowned journals but also more recent papers presented at conferences. There were about 54,100 results, sorted by relevance.[1] We analyzed abstracts of the first 100 results and selected 14 papers for detailed content analysis towards the definition of linked data. Two additional papers were excluded as not relevant, one of them, coming from 1981, entitled "Garbage Collection of Linked Data Structures" by J. Cohen. The findings are interpreted in the remainder of this section in a structured manner.

[1] As of 2017-02-28.



K. Węcel, *Big, Open and Linked Data*, Business Information Systems,
https://doi.org/10.1007/978-3-031-07147-8_3

**Table 3.1** Number of search results for 'linked data' in various sources

| Database | Number of results |
|---|---|
| ACM Digital Library | 698 |
| Google Scholar | 54,100 |
| IEEE Xplore | 560 |
| ProQuest | 2359 |
| Scopus | 5129 |

Source: own analysis

## 3.2 Linked Data Definition

### *3.2.1 Definition*

A primary publication about linked data by Berners-Lee (2006), still referenced by a majority of authors, provided only a vague idea of what linked data is. It focused on the so-called "linked data principles" rather than the strict definition. The definition was indirect, by specifying rules for putting data online.

What is characteristic for linked data is the ability to establish data-level links, i.e., to set hyperlinks between data items (Bizer, 2009; Hitzler & Janowicz, 2013). Links are said to be typed because they can refer to different relations (Bizer et al., 2009). Such links not only link data items but also combine various sources into a single global data space, which is then called "Web of Data" or "data cloud" (Berners-Lee, 2006; Bizer, 2009; Heath & Bizer, 2011; Jain et al., 2010). Analysis of links between the datasets led to the creation of the Linking Open Data cloud diagram,[2] which represents high-level associations between the sources (Bizer, 2009). In order to make data items unambiguous, the idea of global identifiers from the Web is borrowed and applied to raw data, not documents (Hitzler & Janowicz, 2013). Guerrini and Possemato (2013) highlighted the differences between a traditional web of documents and a web of data. Linked data can be used to establish new relations between data items or to represent the links between data already linked using other methods (Dadzie & Rowe, 2011). A relational database is a good example of the latter approach, where data is exposed in a new manner.[3]

Linked data does not introduce new technology—it builds upon the general architecture of the Web. For linking, HTTP URIs are used for the identification of arbitrary real-world entities (Auer et al., 2011; Bizer et al., 2009). The data itself is represented in the RDF format (Bizer, 2009; Latif et al., 2009). Thanks to RDF, linked data can inherently be used for the realization of AAA—Anybody can say Anything about Anything (Hitzler & Janowicz, 2013). Reuse of HTTP for interlinking of various datasets should enable the navigation between them much in the same way as the browsing between documents is possible (Jain et al., 2010).

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

[3] e.g. D2R server, http://d2rq.org/.

Linked data provides a technical background to realize such linkage. Additionally, the use of web standards facilitates the creation of generic applications that can exploit this global data space (Heath & Bizer, 2011).

As linked data is based on open web standards, there are generic tools that can be used straightway (Bizer, 2009). By tools we mean applications capable of handling semantic web-specific technologies, mostly RDF and SPARQL, which were in play long before linked data.[4] Several dedicated tools were also developed, particularly RDF stores (e.g., Virtuoso) and publishing tools (e.g., Pubby, Triplify) (Bizer et al., 2009). Another class encompasses useful tools for converting between various serialization formats like RDF-XML, Turtle, and RDFa.[5]

### *3.2.2 Features of Linked Data*

Linked data can be characterized by several features, which are explicitly mentioned by Berners-Lee (2009):

- open—expressed in open, non-proprietary formats, available to any application,
- modular—can be combined with any other linked dataset, integration is available out-of-the-box,
- scalable—it is easy to add more linked data, even when data changes over time.

Other authors point to a probably more important feature, being the essence of linked data: self-describing. In fact, the meaning of data is explicitly defined and is machine-processable (Bizer et al., 2009). It can be inferred from various contexts where a data item not only can be linked *to* other external data but can be linked *from* other datasets as well. This leads to another important feature—*discoverability*: linked data applications can discover new data items at run-time by following RDF links (Bizer, 2009). A definition can also concern real-world objects if they are given a URI that can be dereferenced to a description of the data item (Berners-Lee, 2006). Auer et al. (2011) provided further features interpreted as benefits, e.g., *uniformity*—using a uniform data model, *coherence*—by use of typed RDF links, diverse datasets are related, *integrability*—relatively easy syntactic integration, *timeliness*—direct access to evolving data instead of time-consuming ETL processes.

Taking into account the global nature of linked data, it is important to discuss another feature—*size*. Dadzie and Rowe (2011) pointed out that the size and scale of the Web of Data presents a challenge for data processing. Linked data is sometimes treated as part of the big data landscape (Hitzler & Janowicz, 2013; Janssen & Kuk, 2016). It addresses one of the 4V of big data—the variety (De Mauro et al., 2016).

[4] RDF was first recommended in 1999. The first W3C working draft for SPARQL was published on 12 October 2004.

[5] For example, Redland RDF Libraries, http://librdf.org.

The variability of big data can be reduced by applying linked data principles to take advantage of self-describing capabilities (see Sect. 4.5.2).

Conformance of published datasets to linked data practices is important for easier discovery and efficient integration of linked data in applications. Therefore, in recent years, a lot of research has focused on this topic (Hogan et al., 2012; Hyland et al., 2014; Schmachtenberg et al., 2014). The best practices can be roughly grouped into interlinking, vocabulary reuse, and metadata provisioning. Lóscio et al. (2017) mentioned 35 best practices. They concluded that following them would facilitate the interaction between publishers and consumers. Data should be discoverable and understandable by humans and machines.

Extending the notion of self-describing, linked data should be published along with *metadata* in order to increase its usability for consumers (Bizer et al., 2009; Hitzler & Janowicz, 2013; Jain et al., 2010). Metadata is expected to supply at least a description of contents, even just categorization, and the origin of data. Jain et al. (2010) criticized linked data for being "merely more data" and a "weakly linked triple collection." First of all, there was a lack of conceptual description of datasets; hence, knowledge discovery was hampered. Secondly, the absence of schema level links made it difficult to identify relevant datasets by navigation. Finally, there was a lack of expressivity, i.e., the full potential of RDF Schema and OWL was not leveraged, and linked data focused mainly on instances. They postulated the use of an upper-level ontology to mitigate the above deficiencies. Hitzler and Janowicz (2013) claimed that linked data does indeed open the possibility to publish linked descriptions, and Latif et al. (2009) analyzed the 'insufficient description' issue from a provenance point of view. Information on how data was prepared is a special kind of metadata, important in the value chain context. The provenance is deemed necessary both from a scientific (Bechhofer et al., 2013) and practical point of view (Kobilarov et al., 2009). Hogan et al. (2012) made an important distinction between "instance data" and vocabulary, and Schmachtenberg et al. (2014) found that provenance and license metadata is still rarely provided by the data sources.

### *3.2.3 Linked Data Life Cycles*

Several publications concerned a linked data life cycle to describe data sourcing. According to Latif et al. (2009), linked data can generally be produced directly or "RDFized" from raw data. They distinguished four roles in their value chain: (1) a raw data provider—provides data in an arbitrary format for further processing; (2) a linked data provider—offers data through dereferenceable URIs, a SPARQL endpoint, API, or an RDF dump; (3) a linked data application provider—processes linked data and generates human-readable output (e.g., RDFa); (4) an end-user—consumes the linked data.

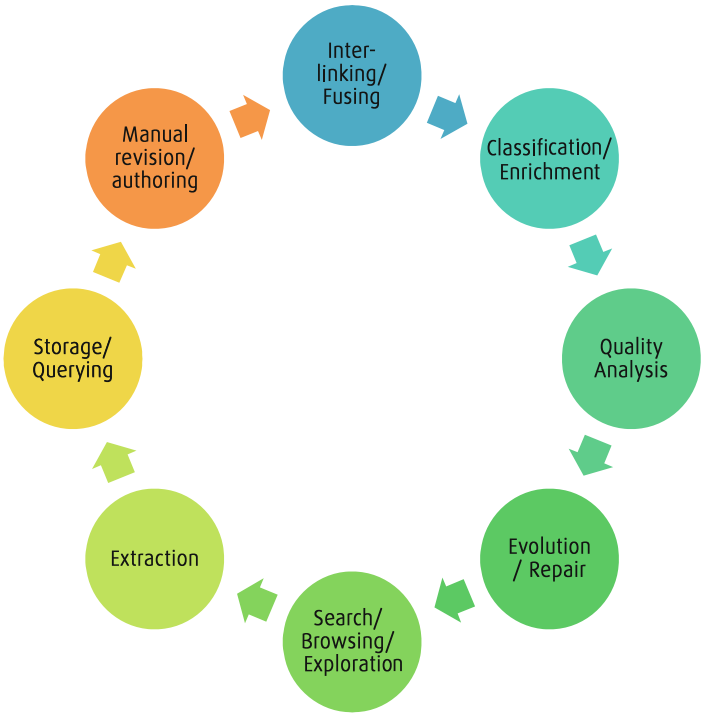



**Fig. 3.1** Linked Data publishing life cycle (LOD2 project). Source: (Auer et al., 2012)

The LOD2 project[6] contributed a more detailed linked data life cycle, which consisted of the following steps (Auer et al., 2012, 2011): (1) extraction, (2) storage and querying, (3) manual revision and authoring, (4) interlinking and fusing, (5) classification and enrichment, (6) quality analysis, (7) evaluation and repair, (8) search, browsing and exploration (Fig. 3.1). The names of the steps are self-explanatory. They were used to integrate the landscape of tools developed by organizations involved in the project—the so-called LOD2 stack. The life cycle focused on the publication process, i.e., which tools should be used to make data available as linked data. The integration of tools was assured by following three pillars of architecture design: (1) software integration and deployment using the Debian packaging system, (2) use of a central SPARQL endpoint and standardized vocabularies, (3) integration of user interfaces based on REST-enabled web applications. The same life cycle was adopted by the GeoKnow[7] project.

Attard et al. (2015) identified the lack of an open data life cycle tailored to the specific needs of open government data and therefore proposed their own life

[6] "Creating Knowledge out of Interlinked Data," http://lod2.eu.

[7] "Geospatial Data and the Semantic Web," http://geoknow.eu/Welcome.html.

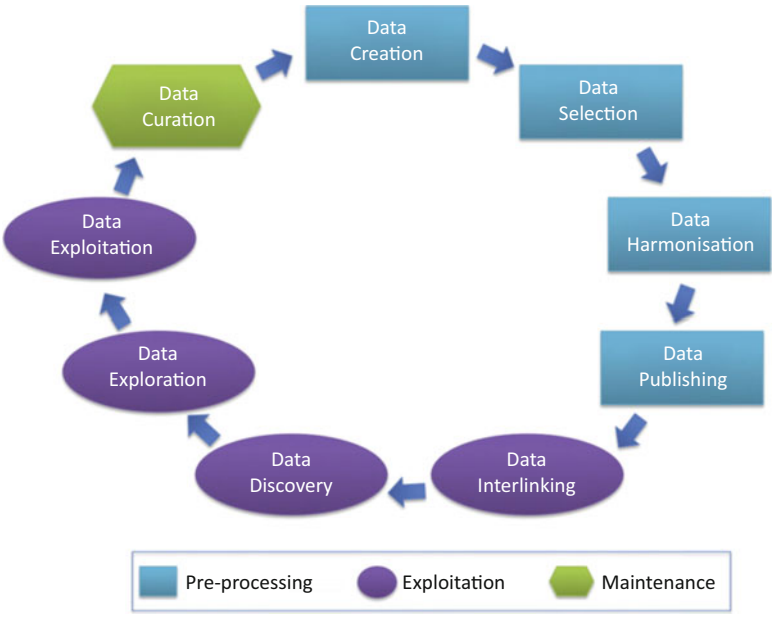



**Fig. 3.2** Open government data life cycle. Source: (Attard et al., 2015)

cycle (see Fig. 3.2). It consisted of three phases: preprocessing—preparing data for publication, exploitation—using the published data, and maintenance—making data sustainable. Altogether, there were nine steps. An open government data life cycle usually started with the *creation of data*. *Data selection* was designed to decide which data will be published and which will not be published, e.g., personal data. *Data harmonization* aimed at preparing data to conform to certain publishing standards. The last step of preprocessing was *data publishing*, where data was made available on the government portal. The exploitation phase started with *data interlinking*, as defined by the Five Star Deployment Scheme for Open Data (Berners-Lee, 2006), and linking of data provided a context for its interpretation. The next step, *data discovery*, was to raise awareness of the existence of the published dataset. *Data exploration* was a trivial step where data was just browsed or visualized. A more advanced way of consuming data was *data exploitation*—the user could distribute open data in the form of analyses or mashups, or expand upon the open data. The only step in the maintenance phase was *data curation*. This step was vital in ensuring that the published data was sustainable: data could be updated, enriched with metadata, completed, etc. It is worth noting that although the life cycle was proposed for open government data, it explicitly included a reference to linked data as a natural way to express open data.

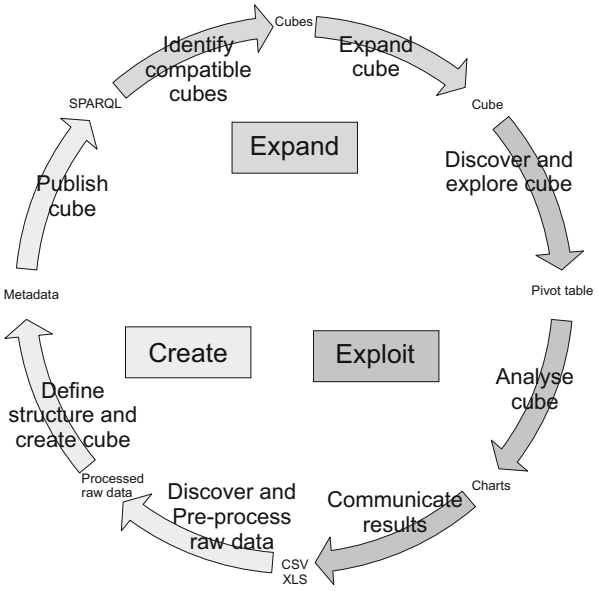



**Fig. 3.3** Linked Statistical Data Lifecycle. Source: (Tambouris et al., 2015) ©Springer

Yet another life cycle was developed within the *OpenCube* project.[8] In the first version, several steps were grouped into two phases: publish and reuse (Kalampokis et al., 2014). It was then refined, and three phases were proposed: create, expand, and exploit (cf. Fig. 3.3). The 'create' phase defined how raw data was transformed into linked data cubes. First, raw data had to be discovered and preprocessed—filtered, sorted, cleansed. In the next step, raw data was transformed to RDF using RDF Data Cube vocabulary, where a cube was defined in terms of dimensions, measures, and attributes. The cube was then published in the last step—it was made available through different interfaces, e.g., SPARQL endpoint, dump, or API. The 'expand' phase aimed at completing existing data with additional data. One needed to identify compatible cubes, which were then used to expand the original cube. The 'exploit' phase specified how the created cubes could be utilized for analytics and visualization. In the first step, a cube needed to be discovered and explored. The next step was analysis, where summaries could be computed or predictive models built. The last step was the communication of results, usually by means of visualization. The life cycle was the foundation for the design of the architecture. For each step of the life cycle, the following layers were defined: (a) user interface, (b) data management, (c) infrastructure, (d) storage, (e) model.

[8] "Publishing and Enriching Linked Open Statistical Data for the Development of Data Analytics and Enhanced Visualization Services," 2-year EU FP7 Project, http://opencube-project.eu/.

OpenCube aimed at the development of a linked statistical data cloud. Within the context of this work, it is relevant to analyze how data expansion was done. The first component, *Compatibility explorer*, searched linked data sources, identified compatible cubes, and established typed links to remote compatible cubes. Discovery of new cubes was based on a compatible structure (data structure definition in semantic terms). *Aggregator* then computed aggregations of cell values across hierarchies and dimensions. The third component, *Expander*, created a new expanded cube based on two compatible cubes. It was possible to expand a set of dimensions, a set of attributes of a dimension, or a set of measures. Finally, the *OLAP Browser* performed OLAP operations on the linked data representation. The toolkit provided was a proof of concept of open data analytics.

In all three presented approaches, the life cycle was introduced to take control over linked data quality. Issues concerning the quality of data are described in Sect. 3.5.

### *3.2.4 Linked Data Contribution*

In order to justify cash expenditure, we need projects that create value based on the datasets made available. Two approaches concern the mechanisms of value creation: either existing mechanisms can be made cheaper, or new ones can be invented. One of the possible cases is replacing existing mechanisms for data exchange by linked data technologies. This approach can reduce the cost of data exchange among public bodies maintaining, for example, core registries. The other possible case is to create new value based on published data.

The linked data principle contributes value to information through combination with other data (Tinholt, 2013). Taking into account the economic perspective, not only does a supply have to be offered, but it should also meet a certain demand on the consumers' side. Going back to linked data principles, they are intended to foster reuse, linkage, and consumption of that data.

A better user experience with linked data can be achieved through visualization techniques. Dadzie and Rowe (2011) claimed that the consumption of linked data was difficult to non-technical users and identified requirements for its visualization. Providing visualization can be perceived as a transformation for increasing the value of data. There was only one paper from the analyzed collection focusing on this domain.

Practical implementation of linked data was presented by only one paper. Kobilarov et al. (2009) analyzed how linked data was used by the BBC. Nevertheless, this commercial approach was appealing to other applications as well.

To sum up, linked data realizes the vision of evolving the Web into a global data space, where various applications operate on an ad hoc defined set of data sources using standard web access mechanisms (Bizer et al., 2009). It recommends best

practices for connecting and exposing pieces of data, information, and knowledge on the Semantic Web. Moreover, it is perceived as a pragmatic approach and is gaining in importance (Auer et al., 2011). The Web of Data is a product, whereas linked data is the prescription (Hogan et al., 2012). In the scope of this paper, linked data is understood as a new data publishing and data integration paradigm. In a more tangible context, we can speak about the linked data technology stack, which allows for the building of sophisticated applications and solutions deployed in enterprise environments (Hogan, 2014).

## 3.3 Linked Data Assets for Reuse

### *3.3.1 People and Organizations*

There is a multitude of vocabularies for the description of an organization. A search for the term 'organization' in Linked Open Vocabulary[9] returns 1444 results, which include 689 classes (October 2021). Table 3.2 presents the top five vocabularies for describing organizations. The ranking was provided by LOV, and we excluded vocabularies that cannot be dereferenced (non-existing websites).

The lesson learned is that vocabularies evolve. Some of them disappear, others change location. It is then advisable to look at vocabularies developed by renowned organizations. W3C is the standardization body for the World Wide Web; therefore, we will analyze the two ontologies accepted as the W3C recommendation.

*Organization Ontology* (ORG) defines the notion of organization and is used for the publishing of organizational information (Reynolds, 2014). The core ontology defines nine classes and 35 properties. It can represent the following artifacts: organizational structure (suborganizations), reporting structure, locations, and organizational history. It allows domain-specific extensions.

The *Registered Organization Vocabulary* (RegORG) is a profile of the Organization Ontology for describing organizations that have gained legal entity status through a formal registration process (Archer et al., 2013). It defines one class and six properties and captures the fundamental characteristics of a legal entity, e.g., the legal name, the legal identifier, and the activity. RegORG was first developed by the European Commission ISA Programme with support from the Directorate General Internal Market and Services as the Core Business Vocabulary. This Directorate General is responsible for the legislative proposal on interconnecting national business registers. Core Business Vocabulary, which specifies Registered Legal Organizations, can be combined with the more general Organization Ontology (ISA, 2012). A Registered Organization (`rov:RegisteredOrganization`) is a subclass of the Organization Ontology's Formal Organization (`org:FormalOrganization`).

[9] http://lov.okfn.org/dataset/lov/terms?q=Organization.

**Table 3.2** Vocabularies for describing organizations

| Name | URL | Characterization |
|---|---|---|
| `npg:Organization` | http://ns.nature.com/terms/Organization | Represents an organized body of people normally having a legal status. |
| `foaf:Organization` | http://xmlns.com/foaf/0.1/Organization | Represents a kind of Agent corresponding to social institutions, such as companies, societies, etc. |
| `schema:Organization` | http://schema.org/Organization | An organization, such as a school, NGO, corporation, club, etc. |
| `dul:Organization` | http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#Organization | An internally structured, conventionally created SocialAgent, needing a specific Role and Agent that plays it in order to act. |
| `ipo:Organization` | http://purl.org/ipo/core#Organization | Represents a group of people organized aiming at a common goal: social, commercial or political. |

Source: own work

ISA—"Interoperability Solutions for European Public Administration"—was a project of the European Commission conducted in the years 2010–2015 with a global budget of 164 million euro. ISA initiated several other standardization efforts besides RegORG: Core Person Vocabulary[10] for describing a natural person, Core Location Vocabulary[11] for describing a location represented as an address, a geographic name, or a geometry, as well as Core Public Service Vocabulary[12] to capture the fundamental characteristics of a service offered by public administration. Within the ISA Programme, Asset Description Metadata Scheme[13] (ADMS) was also defined to describe semantic interoperability solutions, formerly called semantic assets. The work is continued under the ISA$^2$ Programme[14]—"Interoperability solutions for public administrations, businesses, and citizens", which ran from January 1, 2016 until December 31, 2020 and was dubbed as IT solutions for less bureaucracy.

The problem of identifying and linking of organizations has already been noticed by the European Union and is regulated by Directive 2012/17/EU on Interconnecting Business Registers (EU, 2012). The directive set up a system for the interconnection of business registers and emphasized the need for a common seman-

[10] https://joinup.ec.europa.eu/solution/core-person-vocabulary.

[11] https://joinup.ec.europa.eu/solution/core-location-vocabulary.

[12] https://joinup.ec.europa.eu/solution/core-public-service-vocabulary.

[13] https://joinup.ec.europa.eu/solution/asset-description-metadata-schema-adms.

[14] https://ec.europa.eu/isa2/home_en.

tic standard. RegORG is specifically suitable for implementing this directive—each organization is identified by a unique URI, using the ADMS vocabulary, where an identifier is accompanied by the scheme agency. The directive specifies that the registers must make the following information available through the interconnection free of charge: (a) the name and legal form of the company; (b) the registration number of the company; (c) the registered office of the company and the Member State of registration; (d) information on winding-up or insolvency proceedings.

The largest open database of companies is currently OpenCorporates.[15] It specializes in legal entities and is the best-known implementation of Core Business Vocabulary RegORG. OpenCorporates held information concerning more than 137 million companies as of October 2017 and more than 160 million as of October 2021, being the largest open database of companies in the world. The primary goal is to make information about companies more usable and more widely available for the benefit of the public. This public mission is supported by the appropriate business model, which allows data to be offered for free to journalists, NGOs, and academics.

### *3.3.2 Vocabularies for E-Business*

Various classification codes used by national and international statistics are a good starting point for transition to linked data. For example, EUROSTAT publishes standard code lists for various classification purposes in a semantic form.[16] It is widely used for the exchange and interpretation of statistical data throughout the European Union and it is part of a broader initiative, the EU Semantic Interoperability Catalogue.[17] The catalogue is a collection of assets that can be used for the development of e-government services with semantics in mind. The assets meet the requirements of the Semantic Asset Clearing Process and Quality Framework.

There are several vocabularies that can be used to classify organizations and describe their activities. GoodRelations can be used to exchange information about products and services, pricing, payment options, other terms and conditions, store locations and their opening hours and, possibly, other conditions related to e-commerce (Hepp, 2011). For classification of products, it uses a semantified version of eCl@ss called eClassOWL (Hepp & Radinger, 2010). eCl@ss[18] is the most sophisticated standard for describing the nature and features of products. It covers more than 30,000 types of products and services and defines more than 5000 product attributes. An optional proposal for product description is ProductOntology.[19] It

[15] https://opencorporates.com/info/about.

[16] https://joinup.ec.europa.eu/collection/eurostat.

[17] https://joinup.ec.europa.eu/collection/eu-semantic-interoperability-catalogue.

[18] http://www.eclasscontent.com/.

claims to provide "high-precision identifiers for product types" thanks to the reuse of the identifiers of English Wikipedia. For example, if a carrot is described by Wikipedia on the page https://en.wikipedia.org/wiki/Carrot, the respective URI of the product concept is http://www.productontology.org/doc/Carrot. As of October 2021, the service held about 300,000 precise definitions of products. This ontology is linked from another popular vocabulary, used in over 1,000,000 domains, for the description of products: http://schema.org/Product. It defines a product as "anything that is made available for sale" and is used for embedding of product descriptions in web pages using Microdata or RDFa formats.

Another ontology worth mentioning is Financial Industry Business Ontology (FIBO[20]). Its designation is to define business concepts in the financial services industry. FIBO consists of 10 core domains (top level directories), which are then split into subdomains, modules, and ontologies (as of October 2021). The following domains are distinguished: Business Entities, Business Process, Corporate Actions and Events, Derivatives, Financial Business and Commerce, Foundations, Indices and Indicators, Loans, Market Data, and Securities. Each FIBO ontology has one of the three levels of maturity: release (only Business Entities achieved this level), provisional (under development), and informative (considered deprecated but included for informational purposes). FIBO is developed by the Enterprise Data Management Council (EDMC) and standardized through the Object Management Group (OMG). The ontology is serialized in OWL files.

FIBO is reviewed by EDM Council member firms and applied in practice. As specified by ontology design principles, it represents a consensus in the industry. For example, Deutsche Bank used FIBO's basic language, rules, and classifications to standardize content and meaning within its organization (Zaino, 2016). The most important achievement was elimination of the need for copying, mapping, and translating data before it could be used, leading directly to cost savings.

### *3.3.3 Geospatial Data*

Geospatial data is a special kind of data that also attracts a lot of attention regarding linked data development. Various initiatives come both from governments and businesses.

Among the political initiatives, the largest is INSPIRE—Infrastructure for Spatial Information in Europe. INSPIRE was established by Directive 2007/2/EC, which requested the Commission to build a community geoportal to create a European Union spatial data infrastructure (EU, 2007). The infrastructure shall be used for the purposes of EU environmental and related policies. Member States were also requested to provide access to their infrastructures through the central geoportal.[21]

[19] http://www.productontology.org.

[20] https://spec.edmcouncil.org/fibo/.

The geoportal consists of a viewer, a resource browser, a metadata editor, and a validator. Geodata is naturally modeled in OGC standards that make use of XML/GML.

The Open Geospatial Consortium (OGC[22]) is an international not-for-profit organization committed to developing quality open standards for geospatial content and services, GIS data processing, and data sharing. It developed many standards that are widely used by the geospatial community, the most popular being: GeoSPARQL—A Geographic Query Language for RDF Data,[23] KML (formerly Keyhole Markup Language[24]), network Common Data Form (netCDF) standards suite,[25] Web Coverage Service (WCS[26]), Web Feature Service (WFS[27]) for direct fine-grained access to geographic information at the feature level, Web Map Service (WMS[28]) for requesting geo-registered map images, and Well-Known Text representation of coordinate reference systems (WKT[29]).

Some consortia make an effort to represent spatial data as linked data. INSPIRE also attracted the attention of W3C—there are some standardization efforts going on within Spatial Data on the Web Working Group to represent INSPIRE in RDF. The European projects working on this topic include: GeoKnow—"Geospatial Data and the Semantic Web"[30] for "making the web an exploratory place for geospatial data," MELODIES,[31] and SmartOpenData.[32] The related project is LinkedGeoData,[33] which is an effort to add a spatial dimension to the Web of Data. LinkedGeoData offers information collected by the OpenStreetMap project in the RDF model.

## 3.4 Contexts and Disambiguation

A context refers to any information that can tell us something more about a resource. A commonly cited definition of a context was introduced by Dey (2001) "Context is any information that can be used to characterize the situation of an entity," where entities can be people, places, and things. Context can be described by four main categories: identity, time, location, and activity.

[21] http://inspire-geoportal.ec.europa.eu.
[22] http://www.opengeospatial.org/.
[23] http://www.opengeospatial.org/standards/geosparql.
[24] http://www.opengeospatial.org/standards/kml.
[25] http://www.opengeospatial.org/standards/netcdf.
[26] http://www.opengeospatial.org/standards/wcs.
[27] http://www.opengeospatial.org/standards/wfs.
[28] http://www.opengeospatial.org/standards/wms.
[29] http://www.opengeospatial.org/standards/wkt-crs.
[30] http://geoknow.eu/.
[31] https://www.melodiesproject.eu/.
[32] http://www.smartopendata.eu/.
[33] http://linkedgeodata.org/.

Zimmermann et al. (2007) introduced two extensions to available context definitions that provided a natural understanding of this concept to users of context-aware applications and facilitated the engineering of that concept for software developers of such applications. The context defined: individuality, time, location, activity, and relations of the entity. Relations are distinguished as social, functional, and compositional, and depend on time, location, and entity. For context acquisition, we need to define which mechanisms can be used to capture the attributes. Henricksen et al. (2002) classified these mechanisms as follows: (a) static—fixed and does not change over time, e.g, an identifier, (b) sensed—captured by sensors, (c) profiled—entered by a user, and (d) derived—already processed context also known as higher-level context.

### *3.4.1 Background Knowledge*

Giunchiglia et al. (2012) identified lack of background knowledge as one of the main barriers for the use of semantics. Such background knowledge should be large enough to describe various aspects of reality and at the same time context-sensitive to capture the diversity of the world. There are several knowledge bases that can be considered as a source of background knowledge.

Cyc Knowledge Base[34] is a formalized representation of a fundamental human knowledge: facts, rules of thumb, and heuristics. It consists of over 1,500,000 terms, over 40,000 types of relations (predicates), and more than 25 million assertions (facts and rules), which relate those terms (as of October 2021). Cyc can be used for reasoning about the objects and events of everyday life. Cyc Corp claims that its knowledge base is "the world's largest and most complete general knowledge base and commonsense reasoning engine." From 2001 to early 2017, Cyc Corp offered also an open counterpart of Cyc knowledge base—OpenCyc.[35] Specified portions of the knowledge base could be exported to OWL. The last version of OpenCyc (4.0, released in June 2012) included 239,000 concepts and 2,093,000 facts, mainly taxonomic assertions. OpenCyc distribution was discontinued because the reduced version of Cyc was confusing for potential users.

Suggested Upper Merged Ontology (SUMO[36]) is a highly axiomatized formal upper ontology (Niles & Pease, 2001). As an upper ontology (top-level ontology), it consists of very general terms that are common across all domains. SUMO defines a hierarchy of classes and related rules and relationships. These are expressed in SUO-KIF language with a LISP-like syntax. SUMO together with its domain ontologies form the largest free formal ontology available, with about 25,000 terms and about 80,000 axioms (October 2021). It is combined with other ontologies for its advanced

[34] https://cyc.com.

[35] http://opencyc.org/.

[36] http://www.adampease.org/OP/.

reasoning capabilities offered by axiomatic knowledge. SUMO is the only formal ontology that has been mapped from WordNet synsets.

Yet Another Great Ontology (YAGO[37]) is a large semantic knowledge base derived from several other sources, including Wikipedia (categories, redirects, infoboxes), WikiData, WordNet (e.g., synsets, hyponymy), and GeoNames (Rebele et al., 2016; Suchanek et al., 2007). The latest version 4 contains semantic constraints in the form of SHACL (Shapes Constraint Language). YAGO4 contains more than 50 million entities (e.g, people, organizations, cities) and 2 billion facts about these entities (October 2021). The coverage of the ontology can also be characterized by the size of dump files—about 63 GB of compressed files in Turtle format.

YAGO has some specific features. Its accuracy has been manually evaluated. Every relation is annotated with a confidence value. It combines the clean taxonomy of WordNet with the richness of the Wikipedia category system. Entities are assigned to more than 350,000 classes. Some entities and facts may also have attached spatial and temporal dimensions. YAGO is considered a cleaned and simplified version of Wikidata. It has been linked to the DBpedia ontology and to SUMO, so it is integrated with the linked data cloud.

Comparing the above ontologies, OpenCyc does not contain the axiomatic rules of the Cyc knowledge base and is more like a taxonomy. Cyc, SUMO, and YAGO include also rich axiomatic knowledge. For example, it is possible to infer that two people sharing the same parents must be siblings.

DBpedia[38] is a crowd-sourced community effort to extract structured information from Wikipedia (Lehmann et al., 2015). After extraction and transformation into RDF model, it allows to execute sophisticated queries against Wikipedia. DBpedia is used as a source of identifiers for multiple other knowledge bases and is placed in the center of the linked open data cloud. English DBpedia describes about 4.8 million things, including 1.6 million people, 968,000 locations, 552,000 creative works, and 317,000 organizations (November 2021). Localized versions of DBpedia are available in 140 languages, where 20 languages are considered core chapters. DBpedia is connected to other Linked Datasets by around 50 million RDF links. DBpedia Core Release in January 2021, built from English Wikipedia Language Edition, contained approx. 900 million triples.

Comparing YAGO and DBpedia, they appeared at about the same time and have similar goals. YAGO focuses on providing ontological classes for every entity. DBpedia can only provide classes for those entities that have an infobox on their Wikipedia page. Besides that, these two resources are compatible and can be used simultaneously.

[37] https://yago-knowledge.org/downloads/yago-4.

[38] https://www.dbpedia.org/resources/ontology/.

WordNet[39] is a lexical database for English (Fellbaum, 1998; Miller & Fellbaum, 2007). The key element are synsets (cognitive synonyms), which can be likened to concepts. The main relations among words in WordNet are synonymy and hyperonymy, as instances of lexical relations. Approximately 117,000 synsets are interlinked by means of conceptual relations (October 2021). Majority of relations connect words being the same part of speech (POS). Thus, WordNet consists of four mostly separated subnets—for nouns, verbs, adjectives, and adverbs.

An interesting supplement to WordNet is Lexvo, which contributes to linked open data cloud information about languages, words, characters, and other language-related entities (de Melo, 2015). It defines URIs for terms, languages, scripts, and characters, which are then linked to a variety of resources on the Web. For example, http://www.lexvo.org/page/script/Hang is an identifier of Japanese Katakana script, and http://www.lexvo.org/page/iso639-3/deu represents Standard German language.

BabelNet[40] is a very large multilingual resource. It can be used as an encyclopedic dictionary, a semantic network, or a large knowledge base. It integrates several other knowledge bases, including WordNet, Wikipedia, Wiktionary, Wikidata, and GeoNames. It covers 20 million concepts and named entities in 500 languages. The number of semantic relations is estimated at 1.6 billion (last release February 2021).

### *3.4.2 Contextual Ontologies*

McCarthy (1993) understood context as a way to partition knowledge into a limited set of locally true axioms with common assumptions. Thus, context can also be used as part of reasoning. For example, CYC uses the notion of microtheories. A microtheory is a collection of concepts and facts regarding one particular realm of knowledge. It allows contextual determination of a class membership or property value. Although the knowledge base does not have to be consistent, each microtheory has to be free from contradictions. The truth or falsity of a sentence is context-relative, i.e., considered in the context of a given microtheory. Giunchiglia et al. (2012) also used a context as a tool to restrict reasoning to a subset of facts known by an agent.

Studer et al. (1998) came up with a definition of ontology that merged various preceding definitions emphasizing various features: "An ontology is a formal, explicit specification of a shared conceptualization." One of the important changes was the aspect of *sharing*, which was not considered by the early definition by Gruber (1993). The importance of having a shared view between several parties was discussed by Guarino et al. (2009): ontologies should express a consensus rather than an individual view. The whole conceptualization cannot be shared

[39] http://wordnet.princeton.edu/.

[40] http://www.babelnet.org/.

as at least part of it resides in mind. Nevertheless, there should be a mutual agreement between stakeholders on primitive terms. Without at least minimal shared ontological commitment, the benefits of having an ontology are limited.

An ontology provides a shared model of a domain, but adding an individual's subjective view of a domain was found useful. Such local models may be encoded by contexts. Bouquet et al. (2003) introduced the notion of contextual ontology. An ontology is contextualized when its content is stored locally, i.e., not shared with other ontologies, but mapped to the contents of other ontologies via explicit (context) mappings. Bouquet et al. (2003) proposed also Context OWL (C-OWL), a language that allowed for the representation of contextual ontologies. The syntax was obtained by extending the OWL syntax and semantics.

Upper Mapping and Binding Exchange Layer (UMBEL[41]) is a general reference structure, which provides a scaffolding to link and interoperate other datasets and domain vocabularies. It is based on open W3C standards and conforms to linked data principles. It uses URIs to give each concept a persistent identifier. Ontology is encoded in OWL 2 and uses SKOS.

The core UMBEL ontology contains about 34,000 reference concepts organized according to 31 mostly disjoint super types. Each concept is characterized by a semset,[42] i.e., a set of alternative labels and terms to describe a concept or entity, including synonyms, aliases, acronyms, and jargon. Labels can also be multilingual. UMBEL contains a curated subset of OpenCyc and has about 65,000 formal mappings to other knowledge bases, such as DBpedia, PROTON, GeoNames, and schema.org. UMBEL concepts are linked to over 2 million Wikipedia entities, but UMBEL itself does not aim to provide content about entities.

Development of UMBEL was motivated by the promise offered by semantic technologies—to make heterogeneous information interoperable. UMBEL places itself as a knowledge scaffolding. It can be used as a base vocabulary for the construction of other concept-based domain ontologies. Various ontologies are developed with different needs and represent different world views. Information can be made interoperable by aligning the contexts and the semantics of the information. UMBEL is positioned to help content interoperate on the Web.

## 3.5 Quality of Data

The data quality is typically defined as a suitability for use for a certain application or use scenario. Wang and Strong (1996) referred to it as "fitness for use." Large variation in data quality is observed on the Web and a similar situation is in linked data (Zaveri et al., 2016). There are some curated datasets of high quality but unfortunately most datasets are community-created, i.e., crowdsourced and

[41] http://umbel.org/.

[42] Semsets are similar to the synsets in WordNet, but with a broader use understanding.

extracted (Dadzie & Rowe, 2011; Hitzler & Janowicz, 2013). Automation of production is not a sufficient condition for quality. Completeness and accuracy of links is one of the central issues in large-scale linked data production (Dadzie & Rowe, 2011). According to Latif et al. (2009), a semi-automatic way of creation is the reason for incomplete, incorrect, and out-of-date data. Moreover, even though an error is detected, it cannot be easily corrected due to the nature of linked data—it has to be modified at the source by the raw data provider. Thus, the quality has to be controlled at the publisher by checking the compliance rules (Guerrini & Possemato, 2013; Hogan et al., 2012). Hogan et al. (2012) extracted fourteen concrete guidelines from the tutorial (Bizer et al., 2007) by which they evaluated publishers. Their quality issues are roughly grouped as: naming, link, data, and dereference issues.

### *3.5.1 Classification of Quality Issues*

Referring to the 'fitness for use' notion, OECD (2011) underlined the context dependency of data quality, especially regarding user needs. OECD (2013) explained this dependency as follows: data that is of good quality for a certain application can be of poor quality for other applications. Thus, it is not possible to assess the value of data ex ante (before use).[43] OECD (2011) viewed quality as a multifaceted concept and defined seven dimensions of data quality. They overlap with some of the dimensions by Attard et al. (2015); therefore, we present them tallied in Table 3.3.

Frank and Walker (2016) proposed user-centered methods for measuring the quality of open data. Their findings were based on the needs of small voluntary sector organizations in the UK and India. They used small structured workshops to identify key quality problems and understand how open data can address them. The approach allowed to pinpoint several issues concerning the expansion of open data use. The most important result of this project was the five attributes of the datasets that emerged as being most significant for the surveyed users:

- Discoverability—how easy it is to identify a dataset containing relevant data. Data is no longer kept only within organizations; therefore, it has to be discovered outside. Datasets can be discovered in different ways: general-purpose search engines, dedicated data portals, domain experts. Precise metadata can help in the discovery of datasets.
- Granularity—how precise is the data: whether it concerns society, groups or individuals. In many cases, information about individual people or companies is needed. Unfortunately, for many reasons, including privacy, open data is unlikely to provide such a level of detail.
- Immediate intelligibility—how easy it is to interpret a dataset. Even though the name of the attribute is known, its interpretation is not always straightforward.

[43] Cf. Arrow's information paradox in Sect. 6.2.2.

**Table 3.3** Linked data quality dimensions

| OECD (2011) | Attard et al. (2015) | Definition |
|---|---|---|
| Relevance | – | The degree to which the data serves the purpose |
| – | Usability | How easily can the published data be used. It specifies such measures as to what degree the data is accessible, open, interoperable, complete, and discoverable |
| Accuracy | Accuracy | The extent to which a data or metadata record correctly describes the respective information. This dimension directly impacts the discoverability of datasets—proper description increases the probability that dataset will be found |
| – | Completeness | Related to the number of completed fields in a data or metadata record. The more fields are complete the higher the chance of finding a dataset, hence higher discoverability |
| Coherence | Consistency | Whether data follows a consistent syntactical format, without contradiction or discrepancy within the entire catalogue of metadata |
| Timeliness | Timeliness | The extent to which the data or metadata is up to date. It measures time passed between an event and updating data |
| Accessibility | Accessibility | (defined using two measures) The psychological accessibility expresses how easy it is to discover a relevant dataset through a data catalogue or repository |
| Interpretability | Accessibility | The cognitive accessibility explains how easy it is for a consumer to understand the published data |
| – | Openness | If a datasets is available as a complete set in an open, machine readable format. It directly influences the use, reuse, and redistribution of data |
| Credibility | – | The confidence put on a specific data provider. Trust in the objectivity of data is one of the aspects |

Source: own work

- Trusted/authoritative—how much trust can be attributed to a dataset. Data should preferably come from an authoritative source. People should also understand how data was collected (provenance).
- Linkable to other data—how a given dataset is linked to other sources of data. It reflects the need to discover relationships between data items. The other data has to be present, in a correct format. Linkable data is useful for enrichment or cross-checking.

A more recent extensive survey of linked data quality assessment methods by (Zaveri et al., 2016) provided 18 quality dimensions that can be grouped into accessibility, intrinsic, contextual, and representational categories as well as 69 metrics to precisely measure quality regarding these dimensions.

Quality should not only refer to the raw data. Information can be extracted from the data; therefore, it is necessary to consider the capacity to link data and to extract insights. Thus, the factors to determine data quality are (Reimsbach-Kounatze, 2015):

- Data linkage—information inferred from data depends on how the underlying data is organized and structured and how it can be linked. The same datasets can lead to different information.
- Data analytic capacities—the quality of data also depends on the meaning as extracted or interpreted by the receiver. Different receivers have different analytical techniques and preexisting knowledge and skills, which also leads to different information.

Even broader view on quality should be assumed during data fusion. It is a challenging task that needs to tackle various issues inherent to source data. Khaleghi et al. (2013) proposed the following classification of data-related aspects (see Fig. 3.4), which are grouped into four categories:

- Imperfection—the most common data quality challenge; further aspects of imperfection are uncertainty, imprecision, and granularity.
- Correlation—the covariance of data should be known to produce relevant results, variables should be preferably independent.
- Inconsistency—encompasses spurious, as well as disordered and conflicting data.
- Disparateness—input data can be generated by a variety of sources, thus in various forms and modalities.

Fusion also offers opportunities for quality improvement by verification. This approach is, for example, applied to improve linked data extracted from Wikipedia (Lewoniewski & Węcel, 2017; Lewoniewski et al., 2016, 2017; Węcel & Lewoniewski, 2015).

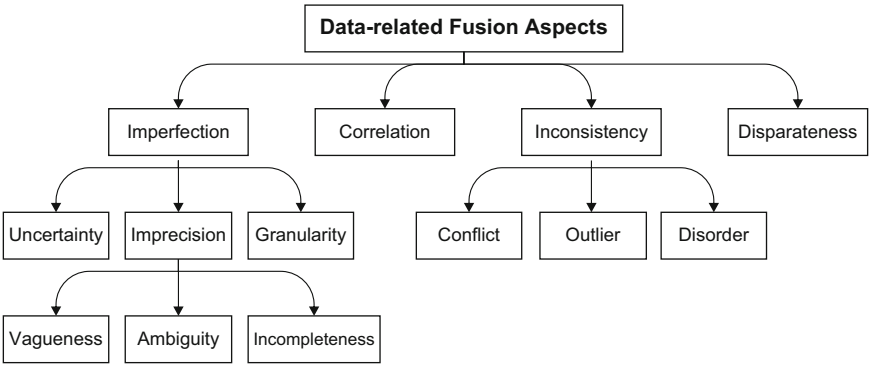



**Fig. 3.4** Taxonomy of data fusion challenges of input data. Source: (Khaleghi et al., 2013)

### 3.5.2 *Data Curation and Repair*

The openness of the Web, where everybody can publish, poses certain challenges for quality. The Web of Data contains data that is "outdated, conflicting, or intentionally wrong" (Bizer et al., 2011). The situation is not improved by governments. Tinholt (2013) found out that countries that share comprehensive and updated data achieved a higher level of user participation and still 96% of the countries analyzed in their research shared data which was not regularly updated. Nevertheless, we need to remember that "disclosing data without proper quality control may jeopardize dataset reuse and negatively affect civic participation" (Vetro et al., 2016). The main challenge of linked data applications is to determine the subset of the Web of Data that should be treated as trustworthy. Kučera et al. (2013) identified two types of strategies for improving data quality: *data-driven* and *process driven*. The first strategy consists in directly modifying the values of data, e.g., normalizing data or correcting invalid data values. The second one focuses on redesigning the data creation processes—quality of data is improved by addressing the root cause, e.g., implementing a data validation step in the data acquisition process.

Quality issues are very often addressed in research projects. Several frameworks were developed including: WIQA (Bizer & Cyganiak, 2009), SWIQA (Fürber & Hepp, 2011), LiQuate (Ruckhaus & Vidal, 2013), RDFUnit (Kontokostas et al., 2014), Open Data Portal Watch.[44] In the linked open data life cycle by LOD2 project, three out of eight steps were related to quality: (3) manual revision and authoring, (6) quality analysis, and (7) evaluation & repair. Quality feature extraction is described in (Ellefi et al., 2017).

In the context of Web, an interesting insight concerning the quality of linked data was provided by Houle (2016). In order to improve the quality of linked data, one usually needs to correct the source data, from which linked data were mapped. As people have own viewpoint on quality, it is usually hard to convince the community for change. That is why various projects that are the source of linked data, e.g., Wikipedia, Wikidata, or OpenStreetMap, are not attracting enough attention and contribution from the community regarding quality. "Reference data users will always have requirements that exceed what they can get from suppliers, so they need the capability to correct errors in the sources when they can, enrich and correct data for themselves when they cannot, and the wisdom to know the difference" (Houle, 2016).

## 3.6 Discoverability of Datasets

Deficit of information is not such a problem as it used to be. "Information has gone from scarce to superabundant" (Gibson, 2010). Discoverability of datasets is an important issue in the context of open data, where the number of datasets is very high

[44] http://data.wu.ac.at/portalwatch.

and constantly growing. Interesting hints are provided by the analysis of reasons for denying access to public information. The main reason for denying freedom of information access in 2013 and 2014 was that the requested data could not be found, whereas only 9% of denials was justified by privacy issues (Deloitte, 2016). After Mehra (2012) we can repeat "the holy grail of search is intent; of browsing, scent; and of classification, meaning." Some portals support only simple search functions that return not only relevant data but also related documents, such as policies or research papers (Alexopoulos et al., 2014). In this section we will show how data discovery can be done exploiting linked data principles.

### *3.6.1 Data Profiling*

Data profiling is the task of examining the data available from a given information source and collecting appropriate metadata. Profiling may serve various purposes. First of all, we may wish to know if known metadata accurately describes the actual values in the source database, or we can discover new metadata, e.g., value patterns, distributions, or foreign keys. By studying the content of the database, we can also assess the data quality (e.g., if data is complete). Moving towards the business domain, such information may be used to assess the risk of integrating new datasets with existing ones. It makes it possible to know the characteristics of data before buying it or downloading a significant portion of it. Moreover, we can discover if the existing data can be used for other purposes. Finally, data profiling can support in understanding the challenges related to any data intensive project. From a managerial point of view, we can refer to master data management and data governance.

The most basic form of data profiling is the analysis of individual columns in a given data source. The discovered metadata usually contains various counts (e.g., number of records, missing values), distribution of values, or identified patterns in data (e.g., in the form of regular expressions). More advanced techniques can detect relations between columns in one or more tables (e.g., how values overlap). Such 'dependency detection' can be used, for example, for discovery of foreign keys (Naumann, 2014). Detailed classification of data profiling tasks is presented in Fig. 3.5.

Naumann (2014) also discussed the profiling of heterogeneous data. Various kinds of heterogeneity can be traditionally distinguished: (a) syntactic heterogeneity, e.g., inconsistent formatting, (b) structural heterogeneity, e.g., unmatched schemata, differently structured information, and (c) semantic heterogeneity, e.g., mismatched meaning of data. Particularly interesting is *data profiling for integration*. With regard to heterogeneity, it is necessary to identify structural and semantic characteristics—structural profiling seeks information about the schema and semantic profiling seeks information about the data. Both levels of profiling help to assess the so-called integrability of data. Schematic similarity is the degree to which the schemata of integrated data sources overlap or complement each other. Data overlap helps to identify objects common to both sources. In the case when

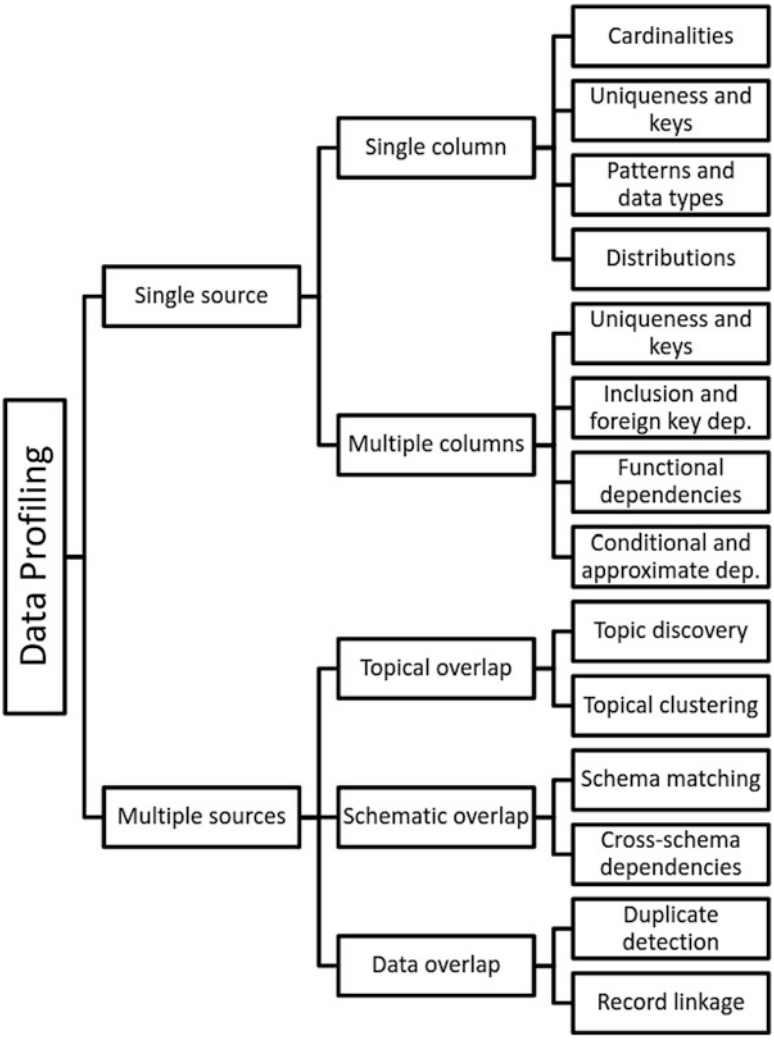



**Fig. 3.5** Data profiling tasks classification. Source: (Naumann, 2014)

external data is not known yet, a topical profiling can be of help. It consists in assigning topics to datasets. For example, Filipowska et al. (2014) presented method to describe CSV files from `open.gov.uk` using DBpedia categories.

Enterprises may increasingly request to profile data that they do not own yet. There are also a lot of non-traditional types of data: non-relational (e.g., linked data), non-structured (e.g., tweets), and fragmented (e.g., open government data). Naumann (2014) called "to develop more efficient and more advanced profiling techniques, in particular for the fast growing areas of *big data* and *linked data*."

Profiling of linked data is particularly challenging—such non-relational models having a loose structure require model-specific profiling methods. Initial methods

focused mostly on various statistics, e.g. ExpLOD (Khatchadourian & Consens, 2010), ProLOD (Böhm et al., 2010), LODStats (Auer et al., 2012). Advanced methods aimed at gaining a deeper understanding of the underlying structure and semantics, e.g., ProLOD++ (Abedjan et al., 2014). An extensive survey of RDF dataset profiling was conducted by Ellefi et al. (2017).

### *3.6.2 Dataset Annotation and Cataloging*

The discoverability of open data is related to the quality of the metadata describing it; therefore, special attention should be paid to its completeness and accuracy (Conradie & Choenni, 2012). Data profiling focuses more on the numbers characterizing datasets. Dataset description is closer to the user—by using an intentionally designed vocabulary for classification, it allows easier identification of datasets.

Several works covered the problem of semantic annotation of research datasets. Shiri (2014) studied potential applications of SKOS-based knowledge organization systems to facilitate analysis, organization, representation, visualization, and access to big data. Singhal and Srivastava (2014) used knowledge from the World Wide Web and organized knowledge bases, such as DBpedia, Yago, Freebase, and WordNet to derive context and annotations for the research datasets. They found that the semantic annotations increased the search accuracy by 18% over the normal search for datasets.

Meusel et al. (2015) investigated to which extent the topical classification of new LOD datasets can be automated using machine learning techniques and the existing annotations as supervision. The best classification technique in their experiments reached an accuracy of 81.62% on the task of assigning one out of the eight classes to a given LOD dataset. The following categories, originally from Linking Open Data cloud diagram, were considered: media, government, publications, life sciences, geographic, social networking, user-generated content, and cross-domain. Interestingly, a deeper inspection of the classification issues revealed problems with the manual classification of datasets in the current LOD cloud. Tygel et al. (2015) pointed out that the tagging process was subject to many problems, such as synonyms, ambiguity, or incoherence. According to their research, these issues were popular in open data portals, making the reuse of data problematic. They proposed an approach for tag reconciliation to improve the quality of tags in a single portal, as well as interlink portals by establishing relations between tags.

Datasets can be discovered by searching metadata but also by similarity to known datasets. Manual search cannot keep pace with the growth of linked data cloud. Röder et al. (2016) studied the use of topic modelling for dataset search, using a novel approach for determining the topical similarity of datasets based on Latent Dirichlet Allocation (LDA), a probabilistic topic modeling technique. They relied solely on the metadata of the datasets. Their method was implemented in an experimental linked dataset search engine—TAPIOCA.[45] The similarity-based

[45] http://aksw.org/Projects/Tapioca.html.

search can be useful both for publishers, e.g., to avoid duplication of datasets, and for consumers, e.g., to supplement datasets they already posses.

Advantages of standardization and better semantic annotations are already observed in the industry. Zaino (2016) described case of application of Financial Business Industry Ontology (FIBO) in Deutsche Bank. The bank had a lot of market data, which was already cataloged, but users needed to know the classification in order to identify the necessary datasets. Thanks to the introduction of FIBO vocabulary, users were able to search for datasets without pre-determined source knowledge.

### *3.6.3 Discovery of Vocabulary*

In the context of the Web and as a consequence of open data, there is no central repository collecting all datasets. Data resources can be published in a distributed manner and should still be discoverable. Certainly, there are repositories that allow to register a dataset and then offer a search interface to locate relevant datasets. Whether it is published in the repository or made available on a local page, a dataset requires metadata describing its content. For this purpose, in the spirit of linked data, dedicated vocabularies can be used.

A broad list of vocabularies available was provided by Ellefi et al. (2017). The most popular vocabularies are Dublin Core (DC), FOAF, SKOS, Data Cube (QB), VoID, SIOC, Creative Commons (CC), and PROV-O.[46] Not all of them are dedicated for a dataset description, e.g., DC or FOAF. Some of them are used just for taxonomy development and utilization for data description is a byproduct of classification, e.g., SKOS. Others, such as CC, describe just licensing. Schmachtenberg et al. (2014) studied usage of well-known vocabularies as used by the dataset beside registries. Table 3.4 presents the most popular vocabularies used in the crawled data in 2014.

We have carried out an own study based on Linked Open Vocabulary[47] in order to identify relevant vocabularies for dataset cataloging and description.

There are several ways to access the description of vocabularies: dump, SPARQL, API, or just a web browser, which also offers visualization. The statistics on dataset usage were provided by LOD2 Stats. Search for datasets can be performed via API,[48] as this is the recommended and convenient way to access. The results are returned in JSON format, easy for further processing. Access to some information requires also the SPARQL endpoint.[49] In our study, we used both API and SPARQL endpoint offered by LOV.

[46] Statistics were collected using http://stats.lod2.eu/.

[47] http://lov.okfn.org (LOV, 760 vocabularies described as of November 2021).

[48] Description of API available here: http://lov.okfn.org/dataset/lov/api.

[49] http://lov.okfn.org/dataset/lov/sparql.

**Table 3.4** Vocabularies used by more than 5% of crawled datasets

| Prefix | Occurrence | Quota | Prefix | Occurrence | Quota |
|---|---|---|---|---|---|
| rdf | 996 | 98.22% | void | 137 | 13.51% |
| rdfs | 736 | 72.58% | bio | 125 | 12.32% |
| foaf | 701 | 69.13% | cube | 114 | 11.24% |
| dcterm | 568 | 56.01% | rss | 99 | 9.76% |
| owl | 370 | 36.49% | odc | 86 | 8.48% |
| wgs84 | 254 | 25.05% | w3con | 77 | 7.60% |
| sioc | 179 | 17.65% | doap | 65 | 6.41% |
| admin | 157 | 15.48% | bibo | 62 | 6.11% |
| skos | 143 | 14.11% | dcat | 59 | 5.82% |

Source: (Schmachtenberg et al., 2014)

To get a list of vocabularies containing the term 'dataset,' we need to use specific methods from API with the appropriate parameters. The search[50] returned only 13 vocabularies. We were really interested in classes that contain 'dataset' as part of the name. The second search[51] resulted in 292 classes. As there were so many results, we decided to visualize the dependencies between them. Having these results, it was possible to obtain detailed information about each vocabulary using the respective API call.For example, for DCAT: http://lov.okfn.org/dataset/lov/api/v2/vocabulary/info?vocab=dcat

We were interested how the vocabulary was interlinked. Using API, it was only possible to get metadata of any vocabulary, but it was not possible to look into the contents. Therefore, SPARQL endpoint was used to get additional information like the number of referenced and referencing datasets. The below SPARQL query displays the number of datasets reusing a given vocabulary along with the number of occurrences:

```
PREFIX vann:<http://purl.org/vocab/vann/>
PREFIX voaf:<http://purl.org/vocommons/voaf#>

SELECT DISTINCT * {
GRAPH <http://lov.okfn.org/dataset/lov>{
  ?vocabURI a voaf:Vocabulary.
  ?vocabURI voaf:occurrencesInDatasets ?occurrencesInDatasets;
  voaf:reusedByDatasets ?reusedByDatasets;
  voaf:reusedByVocabularies ?reusedByVocabularies;
  vann:preferredNamespacePrefix ?prefix .
}}
ORDER BY ?vocabURI
```

[50] http://lov.okfn.org/dataset/lov/api/v2/vocabulary/search?q=dataset.

[51] http://lov.okfn.org/dataset/lov/api/v2/term/search?q=dataset&type=class.

The relations between vocabularies could only be discovered using a visual application. We then extracted all `rdfs:subclassOf` relations in order to show the hierarchical relations between the most popular classes used for the description of datasets. The taxonomy was prepared in GraphViz,[52] with manual adjustments and coloring. Results are presented in Fig. 3.6.

### *3.6.4 Vocabularies for Description of Datasets*

Data Catalog Vocabulary (DCAT) is an RDF vocabulary for describing datasets in data catalogs. The objective is to increase the discoverability of datasets by standardizing metadata. Then catalogs can be published in a decentralized manner, but the federated search is still facilitated – metadata can be consumed from multiple catalogs (Maali et al., 2014). DCAT defines 7 classes and 17 properties. Main classes are: `dcat:Catalog`—represents the catalog, `dcat:Dataset`—represents a dataset in a catalog, `dcat:Distribution`—represents an accessible form of a dataset. The dataset is defined as "a collection of data, published or curated by a single agent, and available for access or download in one or more formats." According to LOV, the vocabulary has 14 incoming links and 14 outgoing links and is used in 77 datasets.

VoID is an RDF Schema vocabulary for expressing metadata about RDF datasets (Alexander et al., 2011). Main applications are data discovery and cataloging. VoID defines 4 classes and 27 properties. According to LOV, the vocabulary has 31 incoming links and 11 outgoing links and is used in 77 datasets. The fundamental concept of VoID is `void:Dataset` defined as "a set of RDF triples that are published, maintained, or aggregated by a single provider." The dataset has to be a meaningful collection of triples and be available through a resolvable HTTP URI or a SPARQL endpoint. VoID defines also a special kind of dataset—`void:Linkset`: a collection of RDF links, i.e., links between instances from different datasets, usually with `owl:sameAs` property.

DCAT and VoID are both RDF vocabularies used for describing datasets. However, they target different forms of data. VoID is for RDF datasets, while DCAT is neutral with regard to format, but usually describes non-RDF datasets. DCAT vocabulary is also more granular about the source; it describes datasets within data catalogs. Using these vocabularies together can be beneficial, especially when RDF is derived from a CSV file. In this case, the source CSV is described using DCAT, and the resulting RDF dataset is described with VoID. Therefore, `dcat:Dataset` term is broader than `void:Dataset`, allowing non-linked data datasets.

```
:void_dataset dct:source :dcat_dataset.                    1
:dcat_dataset dcat:distribution :dcat_dist.                2
:dcat_dist dcat:accessURL <url_to_csv>                     3
```

[52] http://www.graphviz.org/.

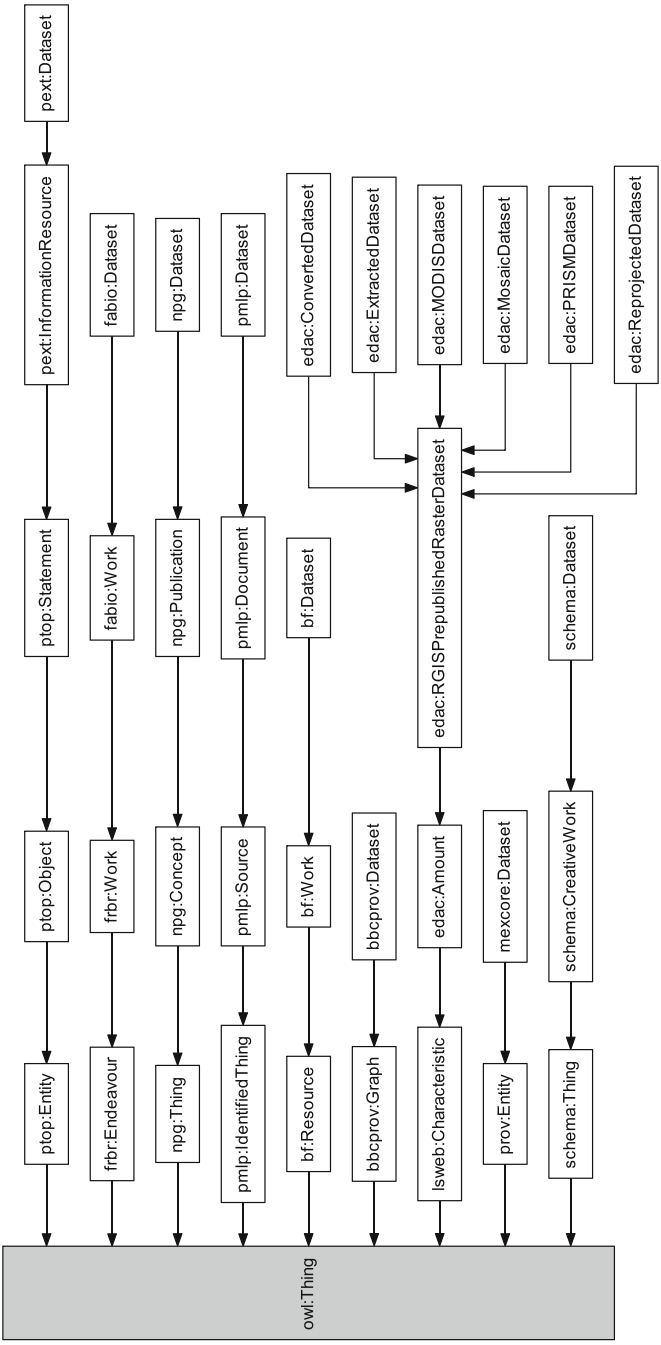



**Fig. 3.6** Hierarchy of classes used for dataset description

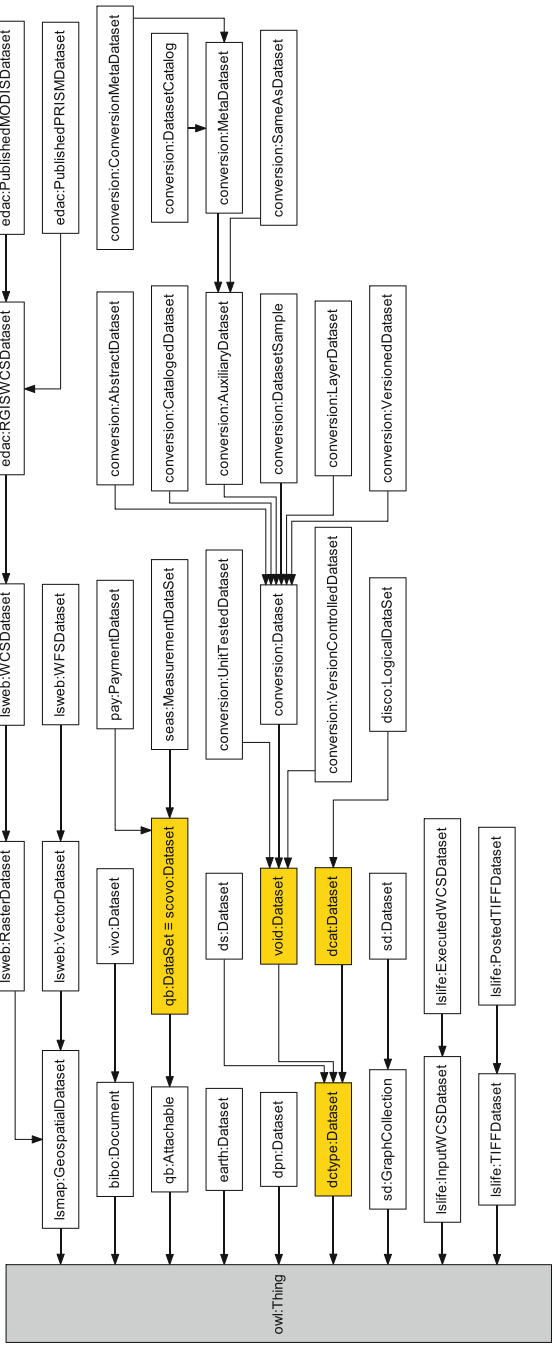



**Fig. 3.6** (continued)

Additional useful vocabularies for the description of datasets include, among others, VANN, ADMS, and DUV. They have not been identified in the previous step because they do not contain 'dataset' as a class but prove to be useful and are very often referenced from the analyzed vocabularies.

VANN is a vocabulary for annotating vocabulary descriptions (Davis, 2010). It is a very simple vocabulary consisting just of six properties, mostly for pointing to an external document, e.g., changes or usage notes. According to LOV, it has 265 incoming links and 3 outgoing links; it is used in 19 datasets. VANN is used by LOV itself. It is very often imported by other vocabularies as indicated by the number of incoming links.

Vocabulary of a Friend (VOAF) is used for the description of RDFS vocabularies or OWL ontologies (Vatant, 2013). It defines 3 classes and 21 properties. Some properties represent metrics of vocabulary elements usage in LOV. According to LOV, it has 18 incoming links and 13 outgoing links. VOAF relies on VoID (one of the outgoing links) and can be used to define networks of vocabularies. The central concept is `voaf:Vocabulary`, a subclass of `void:Dataset`, which defines the vocabulary used in the linked data cloud.

Asset Description Metadata Schema (ADMS) is a profile of DCAT used to describe semantic assets (Dekkers, 2013). It provides four classes and 13 properties. According to LOV, the vocabulary has 26 incoming links and 11 outgoing links. The asset, denoted as `adms:Asset`, is an abstract entity that reflects the intellectual content of the asset and represents those characteristics of the asset that are independent of its physical distribution (`adms:AssetDistribution`). It can be in the form of highly reusable metadata (e.g., XML schema) or reference data (e.g., code lists, taxonomies, dictionaries). In terms of Functional Requirements for Bibliographic Records (IFLA, 1998), an asset is a combination of *work* (a distinct intellectual or artistic creation) and *expression* (the intellectual or artistic realization of a work), while distribution is entity *manifestation*. `adms:Asset` is a subclass of `dcat:Dataset` and `adms:AssetDistribution` is a subclass of `dcat:Distribution`.

DataID[53] ontology is an effort by the DBpedia community to define a core vocabulary for detailed description of datasets and their different manifestations (Freudenberg & Brümmer, 2016). It aims at establishing a uniform way to describe and deliver dataset metadata for arbitrary datasets. The model integrates DCAT, VoID, Prov-O, and FOAF. DataID extends DCAT with capabilities to describe dataset hierarchies, permissions, distributions, fine-grained technical details, and machine-readable licensing information. `dataid:Dataset` is a subclass of `void:Dataset`. The `dataid:DataId` class is the most generic entity in a DataID graph about one or more datasets.

The vocabularies presented so far focused on the publishing process and favored the publisher's view. Dataset Usage Vocabulary (DUV) is used to describe consumer experiences, citations, and feedback about the dataset from the human perspective

[53] http://wiki.dbpedia.org/projects/dbpedia-dataid.

(Lóscio et al., 2016). According to DUV, data should be both discoverable and understandable. Understanding of data is a prerequisite to usage and such usage should also be discoverable. The vocabulary aims at facilitating interaction between publishers and consumers. Data producers can track, share, and persist consumer dataset usage. DUV defines 4 classes and 6 properties. Central concepts are `duv:Usage`, which describes actions that can be performed on a given dataset or distribution and what tools are involved, and `duv:UserFeedback`, which can be any kind of user feedback, e.g., describing, questioning, replying. DUV relies on other vocabularies to describe citations, feedback, and usage of datasets published on the Web. DUV reflects the movement from focusing technology towards data understanding.

In quality aspects, DUV refers to another vocabulary—Data Quality Vocabulary (DQV), which extends DCAT with a number of additional properties and classes suitable for expressing the quality of a dataset (Albertoni & Isaac, 2016). It defines 10 classes and 9 properties. Five different types of quality information can be represented: `dqv:QualityAnnotation`—feedback and quality certificates given about the dataset; `dcterms:Standard`—a standard the dataset conforms to; `dqv:QualityPolicy`—a policy or agreement followed; `dqv:QualityMeasurement`—a metric providing quantitative or qualitative information about the dataset; `prov:Entity`—an entity involved in the provenance of the dataset.

To summarize, all the discussed vocabularies have a notion of a dataset and usually distinguish various physical distributions of it. A dataset or its equivalence is defined in the vocabularies as follows: `dcat:Dataset`, `void:Dataset`, `voaf:Vocabulary`, `adms:Asset`, and `dataid:Dataset`. The above list should be understood as representing the same notion and not defining equivalence in ontology terms.

## 3.7 Summary

The development of open data and linked data is very often strongly correlated and it is reflected in a ‘linked open data’ concept. The link between linked data and e-government is also very often studied, and the first is considered an enabler for the transformation of the latter. In fact, linked data can be an enabler for open data as well as for big data.

Linked data is a concept derived from the Semantic Web and is sometimes referred to as “Semantic Web done right.” By using resource description frameworks and uniform resource identifiers, linked data can make the preparation of data mashups easier for open data users. Connection of information among different datasets and sources and further dissemination of information is then facilitated by linking. Transitioning to linked data frees up data from the systems they are generated in and makes it independent of the source format, accessible to other entities, thus easy to integrate and reuse. Knowing the semantics of data residing in different

datasets makes it possible to combine data directly, without sophisticated translation protocols or import/export procedures.

## References


Abedjan, Z., Grutze, T., Jentzsch, A., & Naumann, F. (2014). Profiling and mining RDF data with ProLOD++. In *30th IEEE International Conference on Data Engineering* (pp. 1198–1201). https://doi.org/10.1109/ICDE.2014.6816740 (page 58)

Albertoni, R., Isaac, A. (Eds.). (2016). *Data on the web best practices: Data quality vocabulary W3C working group note*. https://www.w3.org/TR/vocab-dqv/ (visited on 2017-11-01). (page 65)

Alexander, K., Cyganiak, R., Hausenblas, M., & Zhao, J. (2011). *Describing linked datasets with the VoID Vocabulary. W3C interest group note*. https://www.w3.org/TR/void/ (visited on 2017-11-01). (page 61)

Alexopoulos, C., Zuiderwijk, A., Charapabidis, Y., Loukis, E., & Janssen, M. (2014). Designing a second generation of open data platforms: Integrating open data and social media. In *Electronic Government: 13th IFIP WG 8.5 International Conference EGOV 2014, Dublin, Ireland, September 1–3, 2014. Proceedings*. Berlin Heidelberg: Springer (pp. 230–241). https://doi.org/10.1007/978-3-662-44426-9_19 (page 56)

Archer, P., Meimaris, M., & Papantoniou, A. (Eds.). (2013). *Registered organization vocabulary W3C working group note*. http://www.w3.org/TR/vocab-regorg/ (visited on 2017-10-28). (page 43)

Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. *Government Information Quarterly, 32*(4), 399–418 (pages 39, 40, 52, 53)

Auer, S., Bühmann, L., Dirschl, C., Erling, O., Hausenblas, M., Isele, R., Lehmann, J., Martin, M., Mendes, P. N., Nuffelen, B. V., Stadler, C., Tramp, S., & Williams, H. (2012). Managing the life-cycle of linked data with the LOD2 stack. In *The semantic web—ISWC 2012*. LNCS 7650. Berlin Heidelberg: Springer (pp. 1–16). ISBN: 978-3-642-35173-0. https://doi.org/10.1007/978-3-642-35173-0_1 (page 39)

Auer, S., Demter, J., Martin, M., & Lehmann, J. (2012). LODStats—an extensible framework for high-performance dataset analytics. In A. ten Teije, J. Völker, S. Handschuh, H. Stuckenschmidt, M. D'Acquin, A. Nikolov, N. Aussenac-Gilles, & N. Hernandez (Eds.), *Knowledge Engineering and Knowledge Management: 18th International Conference EKAW 2012, Galway City Ireland, October 8–12, 2012. Proceedings* (pp. 353–362). Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-642-33876-2. https://doi.org/10.1007/978-3-642-33876-2_31 (page 58)

Auer, S., Lehmann, J., & Ngomo, A. C. N. (2011). Introduction to linked data and its lifecycle. In *Reasoning web semantic technologies for the web of data (LNCS 6846)* (pp. 1–75). Berlin Heidelberg: Springer. ISBN: 978-3-319-10587-1. (pages 36, 37, 39, 43)

Bechhofer, S., Buchan, I., De Roure, D., Missier, P., Ainsworth, J., Bhagat, J., Couch, P., Cruickshank, D., Delderfield, M., Dunlop, I., Gamble, M., Michaelides, D., Owen, S., Newman, D., Sufi, S., & Goble, C. (2013). Why linked data is not enough for scientists. *Future Generation Computer Systems, 29*(2), 599–611. ISSN: 0167-739X. https://doi.org/10.1016/j.future.2011.08.004 (page 38)

Berners-Lee, T. (2006). *Linked data—design issues*. http://www.w3.org/DesignIssues/LinkedData.html (visited on 2016-03-25). (pages 36, 37, 40)

Berners-Lee, T. (2009). *Putting government data online*. http://www.w3.org/DesignIssues/GovData.html (visited on 2016-03-25). (page 37)

Bizer, C., Cyganiak, R., & Heath, T. (2007). *How to publish linked data on the web*. (page 52)

Bizer, C. (2009). The emerging web of linked data. *IEEE Intelligent Systems, 24*(5), 87–92. (pages 36, 37)

Bizer, C., Boncz, P., Brodie, M. L., & Erling, O. (2011). The meaningful use of big data: Four perspectives—four challenges. *SIGMOD Record, 40*4, 56–60. (page 55)

Bizer, C., & Cyganiak, R. (2009). Quality-driven information filtering using the WIQA policy framework. *Journal of Web Semantics, 7*(1), 1–10. https://doi.org/10.1016/j.websem.2008.02.005 (page 55)

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data—the story so far. *International Journal on Semantic Web and Information Systems, 5*(33), 1–22. (pages 36, 37, 38, 42)

Böhm, C., Naumann, F., Abedjan, Z., Fenz, D., Grütze, T., Hefenbrock, D., Pohl, M., & Sonnabend, D. (2010). Profiling linked open data with ProLOD. In *Workshops Proceedings of the 26th International Conference on Data Engineering (ICDE)* (pp. 175–178). Long Beach, CA. (page 58)

Bouquet, P., Giunchiglia, F., van Harmelen, F., Serafini, L., & Stuckenschmidt, H. (2003). C-OWL: Contextualizing ontologies. In D. Fensel, K. Sycara, & J. Mylopoulos (Eds.), *The Semantic Web-ISWC 2003: Second International Semantic Web Conference Sanibel Island, FL, USA, October 20–23, 2003. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 164–179. ISBN: 978-3-540-39718-2. https://doi.org/10.1007/978-3-540-39718-2_11 (page 51)

Conradie, P., & Choenni, S. (2012). Exploring process barriers to release public sector information in local government. In *Proceedings of the 6th International Conference on Theory and Practice of Electronic—Governance ICEGOV '12* (p. 5). ACM Press. ISBN: 978-1-450-31200-4. https://doi.org/10.1145/2463728.2463731 (page 58)

Dadzie, A. S., & Rowe, M. (2011). Approaches to visualising linked data: A survey | www.semantic-web-journal.net. *Semantic Web, 2*(2), 89–124. https://doi.org/10.3233/SW-2011-0037 (pages 36, 37, 42, 52)

Davis, I. (2010). *VANN: A vocabulary for annotating vocabulary descriptions*. http://purl.org/vocab/vann/ (page 64)

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Reviews, 65*(3), 122–135. https://doi.org/10.1108/LR-06-2015-0061 (page 37)

de Melo, G. (2015). Lexvo.org: Language-related information for the Linguistic Linked Data cloud. *Semantic Web, 6*(4), 393–400. ISSN: 22104968. https://doi.org/10.3233/SW-150171 (page 50)

Dekkers, M. (2013). *Asset description metadata schema (ADMS). W3C working group note*. https://www.w3.org/TR/vocab-adms/ (visited on 2017-11-01). (page 64)

Deloitte (2016). *The value of DDI (data driven innovation)*. (page 56)

Dey, A. K. (2001). Understanding and using context. *Personal and Ubiquitous Computing, 5*(1), 4–7. (page 47)

Ellefi, M. B., Bellahsene, Z., Breslin, J. G., Demidova, E., Dietze, S., Szymański, J., & Todorov, K. (2017). RDF dataset profiling—a survey of features, methods, vocabularies and applications. *Semantic Web, 1*, 1–37. (pages 55, 58, 59)

EU. (2007). Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). *Official Journal of the European Union, L 108*, 1–14. (page 46)

EU. (2012). Directive 2012/17/EU of the European Parliament and of the Council of 13 June 2012 amending Council Directive 89/666/EEC and Directives 2005/56/EC and 2009/101/EC of the European Parliament and of the Council as regards the interconnection of central, com. *Official Journal of the European Union, L 156*, 244–252. (page 44)

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press. ISBN: 978-0-26-206197-1. (page 50)

Filipowska, A., Węcel, K., & Filipiak, D. (2014). Efficient search and browsing of CSV datasets. In H. Sack, A. Filipowska, J. Lehmann, & S. Hellmann (Eds.), *Proceedings of the Posters and Demos Track of 10th International Conference on Semantic Systems—SEMANTiCS2014*

(Vol. 1224, pp. 6–9). CEUR workshop proceedings. Aachen: Sun SITE, Informatik V RWTH Aachen. (page 57)

Frank, M., & Walker, J. (2016). User centred methods for measuring the quality of open data. *The Journal of Community Informatics, 12*(2), 47–68. (page 52)

Freudenberg, M., & Brümmer, M. (Eds.). (2016). *DataID core Ontology. W3C member submission*. http://vmdbpedia.informatik.uni-leipzig.de/temporary/html/dataid-submission-pre.html (visited on 2017-11-01). (page 64)

Fürber, C., & Hepp, M. (2011). SWIQA—a semantic web information quality assessment framework. In V. K. Tuunainen, M. Rossi, & J. Nandhakumar (Eds.), *19th European Conference on Information Systems, ECIS 2011* (pp. 19–30). Helsinki: IEEE Computer Society. (page 55)

Gibson, W. (2010). Data, data everywhere. *The Economist. Special report: Managing information, 394*(8671), 3–5. (page 55)

Giunchiglia, F., Maltese, V., & Dutta, B. (2012). Domains and context: First steps towards managing diversity in knowledge. *Journal of Web Semantics, 12–13*, 53–63. https://doi.org/10.1016/j.websem.2011.11.007 (pages 48, 50)

Gruber, T. (1993). Towards principles for the design of ontologies used for knowledge sharing. In N. Guarino & R. Poli (Eds.), *Formal ontology in conceptual analysis and knowledge representation*. Kluwer Academic Publishers. (page 50)

Guarino, N., Oberle, D., & Staab, S. (2009). What is an ontology? In S. Staab & R. Studer (Eds.), *Handbook on ontologies* (pp. 1–17). Berlin, Heidelberg: Springer. ISBN: 978-3-540-92673-3. https://doi.org/10.1007/978-3-540-92673-3_0 (page 50)

Guerrini, M., & Possemato, T. (2013). Linked data: A new alphabet for the semantic web. *JLIS.it, 4*(1), 67. (pages 36, 52)

Heath, T., & Bizer, C. (2011). *Linked data: Evolving the Web into a global data space* (1st ed., Vol. 1. 1, pp. 1–136). Morgan & Claypool. (pages 36, 37)

Henricksen, K., Indulska, J., & Rakotonirainy, A. (2002). *Modeling context information in pervasive computing systems* (pp. 167–180). (page 48)

Hepp, M. (2011). *Good relations language reference*. http://purl.org/goodrelations/v1 (visited on 2017-10-01). (page 45)

Hepp, M., & Radinger, A. (2010). *eClassOWL—the web ontology for products and services*. http://www.heppnetz.de/projects/eclassowl/ (visited on 2017-10-24). (page 45)

Hitzler, P., & Janowicz, K. (2013). Linked data, big data, and the 4th paradigm. *Semantic Web Journal, 4*(3), 233–235. (pages 36, 37, 38, 52)

Hogan, A. (2014). Linked data & the semantic web standards. In A. Harth, K. Hose, & R. Schenkel (Eds.), *Linked data management* (Chap. 1, pp. 3–48). CRC Press - Taylor & Francis Group. (page 43)

Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., & Decker, S. (2012). An empirical survey of Linked Data conformance. *Journal of Web Semantics, 14*, 14–44. ISSN: 15708268. https://doi.org/10.1016/j.websem.2012.02.001 (pages 38, 43, 52)

Houle, P. (2016). *Data lakes, data ponds, and data droplets*. http://ontology2.com/the-book/data-lakes-ponds-and-droplets.html (visited on 2017-09-09). (page 55)

Hyland, B., Atemezing, G., & Villazon-Terrazas, B. (2014). *Best practices for publishing linked data*. https://www.w3.org/TR/ld-bp/ (visited on 2017-07-08). (page 38)

IFLA. (1998). *FRBR: Functional requirements for bibliographic records. Final report* (Vol. 19, p. 144). IFLA Study Group on the Functional Requirements for Bibliographic Records. International Federation of Library Associations and Institutions. Section on Cataloguing. Standing Committee. ISBN: 978-3-598-11382-6. (page 64)

ISA. (2012). *Case study: How to describe organizations in RDF using the core business vocabulary and the organization ontology?* Case study. European Commission. ISA Programme. (page 43)

Jain, P., Hitzler, P., Yeh, P. Z., Verma, K., & Sheth, A. P. (2010). Linked data is merely more data. In *Linked data meets artificial intelligence. Technical report SS-10-07* (pp. 82–86). AAAI Press. ISBN: 978-1-577-35461-1. (pages 36, 38)

Janssen, M., & Kuk, G. (2016). Big and open linked data (BOLD) in research, policy, and practice. *Journal of Organizational Computing and Electronic Commerce, 26*(1–2), 3–13. ISSN: 1091–9392. https://doi.org/10.108/10919392.2015.1124005 (page 37)

Kalampokis, E., Karamanou, A., Nikolov, A., Haase, P., Cyganiak, R., Roberts, B., Hermans, P., Tambouris, E., & Tarabanis, K. (2014). Creating and utilizing linked open statistical data for the development of advanced analytics services. In *Proceedings of the 2nd International Workshop on Semantic Statistics co-located with 13th International Semantic Web Conference (ISWC 2014)*. Riva del Garda, Italy: CEUR-WS. (page 41)

Khaleghi, B., Khamis, A., Karray, F. O., & Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Information Fusion, 14*(1), 28–44. ISSN: 1566-2535. https://doi.org/10.1016/j.inffus.2011.08.001 (page 54)

Khatchadourian, S., & Consens, M. P. (2010). ExpLOD: Summary-based exploration of interlinking and RDF usage in the linked open data cloud. In *The Semantic Web: Research and Applications, 7th Extended Semantic Web Conference, ESWC 2010, Heraklion, Crete Proceedings, Part II* (Vol. 6089, pp. 272–287). ISBN: 3642134882. https://doi.org/10.1007/978-3-642-13489-0_19 (page 58)

Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C., & Lee, R. (2009). Media meets semantic web—how the BBC uses DBpedia and linked data to make connections. In *The semantic web: Research and applications (ESWC2009)* (Vol. 5554, pp. 723–737). LNCS. Berlin Heidelberg: Springer. ISBN: 3642021204. https://doi.org/10.1007/978-3-642-02121-3_53 (pages 38, 42)

Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., & Zaveri, A. (2014). Test-driven evaluation of linked data quality. In *Proceedings of the 23rd International Conference on World Wide Web—WWW '14* (pp. 747–758). New York: ACM Press. ISBN: 978-1-450-32744-2. https://doi.org/10.1145/2566486.2568002 (page 55)

Kučera, J., Chlapek, D., & Nečaský, M. (2013). Open government data catalogs: Current approaches and quality perspective. In A. K\Ho, C. Leitner, H. Leitold, & A. Prosser (Eds.), *Technology-Enabled Innovation for Democracy, Government and Governance: Second Joint International Conference on Electronic Government and the Information Systems Perspective, and Electronic Democracy, EGOVIS/EDEM 2013, Prague, Czech Republic, August 26-2* (pp. 152–166). Berlin, Heidelberg: Springer. ISBN: 978-3-642-40160-2. https://doi.org/10.1007/978-3-642-40160-2_13 (page 55)

Latif, A., Saeed, A. U., Hoefler, P., Stocker, A., & Wagner, C. (2009). The linked data value chain: A lightweight model for business engineers. In *Proceedings of ISE-MANTICS09 International Conference on Semantic Systems* (pp. 568–575). Graz. ISBN: 978-3-851-25060-2. (pages 36, 38, 52)

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., Van Kleef, P., Auer, S., & Bizer, C. (2015). DBpedia—a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web, 6*(2), 167–195. ISSN: 22104968. https://doi.org/10.3233/SW-140134 (page 49)

Lewoniewski, W., & Węcel, K. (2017). Relative quality assessment of wikipedia articles in different languages using synthetic measure. In W. Abramowicz (Ed.), *Business Information Systems Workshops: BIS 2017 International Workshops, Poznań, Poland, June 28–30, 2017, Revised Papers* (pp. 282–292). Cham: Springer International Publishing. ISBN: 978-3-319-69023-0. https://doi.org/10.1007/978-3-319-69023-0_24 (page 54)

Lewoniewski, W., Węcel, K., & Abramowicz, W. (2016). Quality and importance of wikipedia articles in different languages. In G. Dregvaite & R. Damaševičius (Eds.), *Information and Software Technologies: Proc. of 22nd International Conference, ICIST 2016* (Vol. 639, pp. 613–624). Communications in Computer and Information Science. Springer International Publishing Switzerland. https://doi.org/10.1007/978-3-319-46254-7_50 (page 54)

Lewoniewski, W., Węcel, K., & Abramowicz, W. (2017). Relative quality and popularity evaluation of multilingual wikipedia articles. *Informatics, 4*(4). ISSN: 2227-9709. https://doi.org/10.3390/informatics4040043 (page 54)

Lóscio, B. F., Burle, C., & Calegari, N. (2017). *Data on the web best practices*. https://www.w3.org/TR/dwbp/ (visited on 2017-07-08). (page 38)

Lóscio, B. F., Stephan, E. G., & Purohit, S. (Eds). (2016). *Data on the web best practices: dataset usage vocabulary. W3C working group note*. https://www.w3.org/TR/vocab-duv/ (page 65)

Maali, F., & Erickson, J. (Eds.). (2014). *Data catalog vocabulary (DCAT). W3C recommendation*. https://www.w3.org/TR/vocab-dcat/ (visited on 2017-11-01). (page 61)

McCarthy, J. (1993). Notes on formalizing context. In R. Bajcsy (Ed.), *Thirteenth International Joint Conference on Artificial Intelligence, IJCAI* (pp. 555–560). (page 50)

Mehra, P. (2012). Context-aware computing: Beyond search and location-based services. *Internet Computing IEEE, 16*(2), 12–16. ISSN: 1089-7801. https://doi.org/10.1109/MIC.2012.31 (page 56)

Meusel, R., Spahiu, B., Bizer, C., & Paulheim, H. (2015). Towards automatic topical classification of LOD datasets. In *CEUR Workshop Proceedings* (Vol. 1409). (page 58)

Miller, G., & Fellbaum, C. (2007). WordNet then and now. *Language Resources And Evaluation, 41*, 209. ISSN: 1574020X. (page 50)

Naumann, F. (2014). Data profiling revisited. *ACM SIGMOD Record, 42*(4), 40–49. ISSN: 01635808. https://doi.org/10.1145/2590989.2590995 (pages 56, 57)

Niles, I., & Pease, A. (2001). Towards a standard upper ontology. In *The 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)* (pp. 2–9). ISBN: 1581133774. https://doi.org/10.1145/505168.505170 (page 48)

OECD. (2011). *Quality framework and guidelines for OECD statistical activities*. Paris: OECD Publishing. (pages 52, 53)

OECD. (2013). Exploring the economics of personal data: A survey of methodologies for measuring monetary value. *OECD Digital Economy Papers, 220*, 40. https://doi.org/10.1787/5k486qtxldmq-en (page 52)

Palmer, S. (2017). *Just how dangerous is alexa?* https://www.linkedin.com/pulse/just-how-dangerous-alexa-shelly-palmer/ (visited on 2017-11-02). (page 35)

Rebele, T., Suchanek, F. M., Hoffart, J., Biega, J., Kuzey, E., & Weikum, G. (2016). YAGO: A multilingual knowledge base from wikipedia, wordnet, and geonames. In *The Semantic Web—ISWC 2016—15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part II* (pp. 177–185). https://doi.org/10.1007/978-3-319-46547-0_19 (page 49)

Reimsbach-Kounatze, C. (2015). The proliferation of big data and implications for official statistics and statistical agencies. *OECD Digital Economy Papers, 245*, 3–39. ISSN: 2071-6826. https://doi.org/10.1787/5js7t9wqzvg8-en (page 54)

Reynolds, D. (Ed.). (2014). *The organization ontology. W3C recommendation*. https://www.w3.org/TR/vocab-org/ (visited on 2017-10-28). (page 43)

Röder, M., Ngomo, A. C. N., Ermilov, I., & Both, A. (2016). Detecting similar linked datasets using topic modelling. In *The Semantic Web. Latest Advances and New Domains—Proc. 13th International Conference, ESWC 2016* (pp. 3–19). ISBN: 978-3-319-18817-1. https://doi.org/10.1007/978-3-319-34129-3_1 (page 58)

Ruckhaus, E., & Vidal, M. E. (2013). LiQuate-estimating the quality of links in the linking open data cloud. In Z. Lacroix, E. Ruckhaus, & M. E. Vidal (Eds.), *5th International Workshop on Resource Discovery, RED 2012* (Vol. 8194, pp. 56–82). Lecture Notes in Computer Science. Heraklion: Springer. ISBN: 978-3-642-45262-8. https://doi.org/10.1007/978-3-642-45263-5_4 (page 55)

Schmachtenberg, M., Bizer, C., & Paulheim, H. (2014). Adoption of the linked data best practices in different topical domains. In *The Semantic Web—ISWC 2014: 13th International Semantic Web Conference* (pp. 245–260). Springer. ISBN: 978-3-319-11964-9. https://doi.org/10.1007/978-3-319-11964-9_16 (pages 38, 59, 60)

Shiri, A. (2014). Linked data meets big data: A knowledge organization systems perspective. In *Advances in classification research online* (Vol. 24, pp. 16–20). American Society for Information Science and Technology. (page 58)

Singhal, A., & Srivastava, J. (2014). Generating semantic annotations for research datasets. In *Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics*

*(WIMS14) - WIMS '14*. New York: ACM Press. ISBN: 978-1-450-32538-7. https://doi.org/10.1145/2611040.2611056 (page 58)

Studer, R., Benjamins, V., & Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data & Knowledge Engineering, 25*(1), 161–197. ISSN: 0169-023X. https://doi.org/10.1016/S0169-023X(97)00056-6 (page 50)

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07)* (pp. 697–706). ACM. ISBN: 978-1-595-93654-7. https://doi.org/10.1145/1242572.1242667. (page 49)

Tambouris, E., Kalampokis, E., & Tarabanis, K. (2015). Processing linked open data cubes. In E. Tambouris, M. Janssen, H. J. Scholl, M. A. Wimmer, K. Tarabanis, M. Gascó, B. Klievink, I. Lindgren, & P. Parycek (Eds.), *Proc. 14th IFIP WG 8.5 International Conference, EGOV 2015, Thessaloniki, Greece* (pp. 130–143). Springer. ISBN: 978-3-319-22479-4. https://doi.org/10.1007/978-3-319-22479-4_10 (page 41)

Tinholt, D. (2013). *The open data economy. Unlocking economic value by opening government and public data*. Capgemini Consulting. (pages 42, 55)

Tygel, A., Auer, S., Debattista, J., Orlandi, F., & Campos, M. L. M. (2015). Towards cleaningup open data portals: A metadata reconciliation approach (p. 8). https://doi.org/10.1109/ICSC.2016.54. arXiv: 1510.04501. (page 58)

Vatant, B. (2013). *Vocabulary of a friend (VOAF)*. OKFN. http://purl.org/vocommons/voaf (visited on 2017-11-01). (page 64)

Vetro, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016). Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly, 33*(2), 325–337. ISSN: 0740-624X. https://doi.org/10.1016/j.giq.2016.02.001 (page 55)

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–33. (page 51)

Węcel, K., & Lewoniewski, W. (2015). Modelling the quality of attributes in wikipedia infoboxes. In W. Abramowicz (Ed.), *Business information systems workshops* (Vol. 228, pp. 308–320). Lecture Notes in Business Information Processing. Springer. ISBN: 978-3-319-26761-6. https://doi.org/10.1007/978-3-319-26762-3_27 (page 54)

Zaino, J. (2016). *Banking on FIBO: Financial institutions turn to semantic standard*. http://www.dataversity.net/banking-fibo-financial-institutions-turn-standard-value-compliance/ (visited on 2017-10-24). (pages 46, 59)

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for linked open data: A survey. *Semantic Web Journal, 7*(1), 63–93. ISSN: 22104968. https://doi.org/10.3233/SW-150175. (pages 51, 53)

Zimmermann, A., Lorenz, A., & Oppermann, R. (2007). An operational definition of context. In *6th International and Interdisciplinary Conference CONTEXT 2007* (Vol. 4635, pp. 558–571). ISSN: 0302-9743. https://doi.org/10.1007/978-3-540-74255-5_42 (page 48)

# Chapter 4
# Big Data Organization Challenge


*Business needs answers even when the answers are not perfect (Helland, 2011)*

## 4.1 Introduction

Big data is a consequence of the progress in data storage and transfer technologies. People store more and more data because they hope that one day they will be able to infer knowledge from that data. Already in 2010 Cisco's Internet Business Solutions Group (IBSG) stated that more things were connected to the Internet than people (12.5 billion devices). They also made a prediction that 25 billion devices will be connected by 2015 and 50 billion by 2020 (IBSG, 2011). According to IBM, every day we create 2.5 quintillion ($2.5 \times 10^{18}$) bytes of data, i.e., 2.5 exabytes (Jacobson, 2013).

Collected data needs to be analyzed, otherwise it contributes to the so-called *dark data*. Companies have gained competitive value from analytics for a long time, starting with financial forecasting and budgeting. Successfully competing with analytics depends on capabilities in three critical areas: data-oriented culture, information management, and analytics experience (Kiron & Shockley, 2011). Technologies proposed in this book can support the second and third area. The enthusiasm surrounding big data stemmed from the conviction that it offers easy access to massive amounts of data (Boyd & Crawford, 2012).

Barnaghi et al. (2013) distinguished two kinds of big data: (a) web-related data and knowledge, e.g., Wikipedia linked open data; (b) sensor device data. They pointed out that the integration of social, cyber, and physical data is interesting for the development of a new breed of applications. They become smart by incorporating context awareness into decision-making process. The need for integration in order to combine big data silos was also observed by OECD (2015). In the survey, 56% of companies stated that 'organizational silos' are the biggest impediment to using big data for effective decision making and a barrier for intra-organizational data sharing. Data silos are perceived as a problem rather by big companies (72% of positive answers) rather than small ones (43%).



K. Węcel, *Big, Open and Linked Data*, Business Information Systems,
https://doi.org/10.1007/978-3-031-07147-8_4

Although there is a hype around big data in research and publications the technology needs to find its way to enterprises. Big data technologies, which allow to integrate data from different sources, are not used very often and the reasons are primarily of an organizational nature (Gronau et al., 2016).

## 4.2 Contemporary Solutions for Data Organization

Data warehousing and master data management (MDM) are among the most important solutions for organization and management of classical data. In the context of the growing volume of data available for enterprises, we observe the emergence of new approaches, methodologies, and tools. "The ability to remove information bottlenecks and create information flow has become a difference between economic success and failure" (Morrison, 2015).

NASCIO (2016) proposed five types of necessary tools for enterprise-wide data management programs: (1) data quality tools—to assess and improve quality of data, (2) entity resolution tools—to match data from different sources, (3) data search tools—to find pieces of information in unstructured data, (4) reporting and analytical tools—to generate various reports based on analyses carried out, and (5) data modeling tools—to structure data and define its semantics. All of these areas are of particular relevance to big data organizations. In the next subsections, we will analyze the types of data available and the particular significance of time.

### *4.2.1 Types of Data in Organizations*

Various types of digital resources in an organization are a challenge to be overcome. There is a consensus to distinguish data, information, and knowledge. According to Hackathorn (1998), data are row facts collected through doing business. Information is aggregated data that has meaning for a particular person. Knowledge is aggregated information that changes behavior of the whole organization.

More detailed granularity of knowledge was defined by Zhang (2013a). He defined data as values drawn from some domain of discourse. Information was defined as the meaning of data values as understood by those who use them. Knowledge was a specialized information about a domain that allows one to make decision. Zhang (2013a) additionally defined meta-knowledge as knowledge about knowledge and expertise as specialized operative task-specific knowledge. Many more definitions are available in other publications. For example, Zins (2007) documented 130 definitions of data, information, and knowledge formulated by 45 scholars.

It is interesting to note that while we speak about 'big data' there is no such concept as 'big information' or 'big knowledge.' In fact, we have to distinguish the whole spectrum of knowledge granularities. Zhang (2013a) calls for precise use of

**Table 4.1** Knowledge content granularities

| | Location-based services | Social networks | Healthcare | Retail |
|---|---|---|---|---|
| Knowledge | Restaurant ratings | Social network structures | Diagnoses | Purchase patterns |
| Information | Restaurants | People who tweet | Patients | Customers |
| Data | Latitude-longitude coordinates | Tweets | X-ray images | Transactions |

Source: (Zhang, 2013a)

concepts and declaration what large datasets contain: primitive data elements, pieces of information, or pieces of knowledge. Table 4.1 presents examples for various knowledge content granularities.

Regarding accessibility, there are two broad categories of data: the *public* and the *personal* (Digital Britain, 2009). Public data is open data as covered in Chap. 2. Example resources are geographical information, census, and meteorological information. Personal data is necessary for identification or in delivering public services. It includes such data as date of birth or home address. Data-Driven Development (2015) distinguished also proprietary data, i.e., data that may be shared but there are no institutional structures available for this purpose; therefore, this data is only available on an ad hoc basis. Usually high value is attributed to this data. Examples include location data, geospatial images from satellites, financial transaction data, logistics, and supply chain details.

Data used by public administration can be further classified into several subcategories (Open Data Institute, 2016). The main type of data they worked with was *administrative data*, which should never be shared. Such data were generated as part of the statutory activities of an organization. Examples include point-of-sale receipts, website access logs, and fixed assets maintenance plans. The second type is *reference data*, which is used to provide information about various things or entities, based on their identifiers. It changes rarely and helps to understand other data. Common reference data is often *shared* between many organizations. Examples include product information, company registrations, and broadcasting licenses. The third type is *aggregate data*, i.e., data summarized from low-level data or being the result of analysis. Analysis can reveal trends and patterns within administrative datasets. There are least barriers to provide aggregate data as open data as it does not provide information about individuals.

Looking at the initial source of data, Llinas (2015) distinguished two broad categories in the data fusion domain: hard and soft. Hard data are generated by physical devices and soft data are provided by people. The first is very often associated with Internet of Things (IoT) and is considered a substantial source of big data. Information created (sensors, computers, mobile phones, and the like) already exceeds the available storage space (Gibson, 2010). Sensors built into various devices digitize lots of information that was previously unavailable. There is

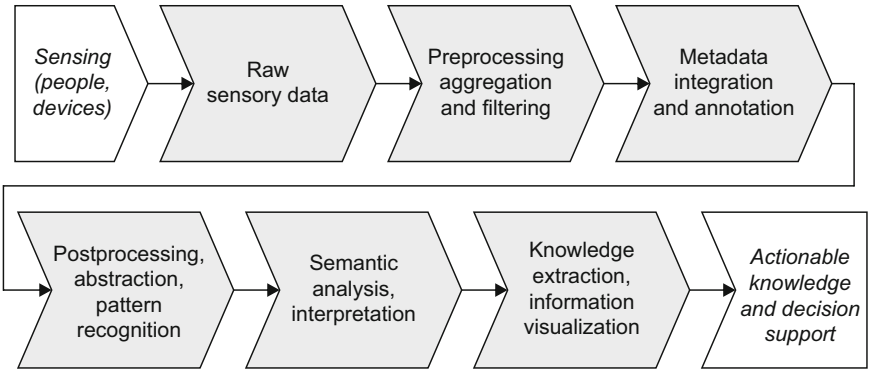



**Fig. 4.1** The process chain for physical-world data on the Web. Source: Barnaghi et al., 2013

an ongoing research on how to combine hard and soft data. For example, Barnaghi et al. (2013) proposed to combine physical-world data with data provided by people on the Web (see Fig. 4.1).

Plachkov (2016) divided soft data into the following categories: (a) observational data—provided by people in written reports, (b) contextual data—characterizes a situation or the surrounding environment, (c) open source and social media—provided by users of social media and other open platforms (crowdsourced). The *observational data* are typically communicated in natural language and as such are semantically rich. If no controlled vocabulary is used, data can be ambiguous. Usually, it is also subjective. Better understanding or disambiguation of observed data can be achieved through *contextual data*, which can provide additional information about the environment of the observed entities. The context can be static (e.g., geographical co-ordinates) or dynamic (phase of the moon). The third category is often referred to as *open data*. They are usually available in big volumes and its relevance to the observed phenomena has to be assessed. Ontology is another tool with the potential to enrich and disambiguate observational data. It can represent entities, relations among them, events, etc.

Special attention should be paid for one of the examples of contextual data—the mobile data. Thanks to the development of technology (mobile phones) and applications, its volume is very fast increasing. This raises data protection and privacy concerns. Naef et al. (2014b) proposed the following principles with regard to this data type:

1. Mobile data should belong to the mobile customer. The operator is only a custodian and should respect rules and preferences regarding the confidentiality of data.
2. Personally identifiable data should not be shared outside the mobile operator without the consent of the customer. Anonymized or aggregated data can be shared with third parties in certain instances.

3. Mobile data can be made available to third party for public interest or scientific projects, e.g., development projects, social programs. The condition is that data should be properly secured and anonymized.

### *4.2.2 Time, Value, and Analytics*

Enterprise data is specific because value of information is changing in time. Data can depreciate in value when it loses its relevance for a particular intended use (Frischmann, 2013).

Hackathorn introduced the concept of the *decision latency* to explain the decrease of the value of information in time. He depicted it on a value-time curve (see Fig. 4.2). This curve represents a simple relation: the longer it takes to make a decision based on new data, the less value can be gained. There is a temporal premium that is motivated by the real-time supply of data (Frischmann, 2013). Hackathorn described response latency as consisting of the sum of *capture latency*, *analysis latency*, and *decision latency* (Taylor, 2012). Capture latency is time between a business event and data capture. Analysis latency represents the time necessary to analyze data. The bigger the volume, variety, or complexity of data, the longer it takes to get results. Finally, decision latency is time necessary to take actions based on delivered information, which usually requires a consensus of the group of people. Hackathorn believed that it is hard to reduce decision latency due to cultural issues. However, it is possible to automate at least some decisions by using business rules approaches. Thus, the decision latency can be reduced almost to zero. We need to focus on capture and analysis latency.

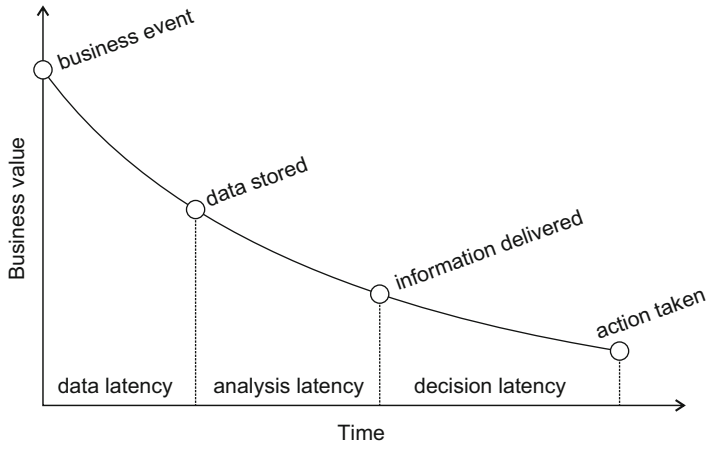



**Fig. 4.2** Value-time curve. Source: based on work by Hackathorn, 1998

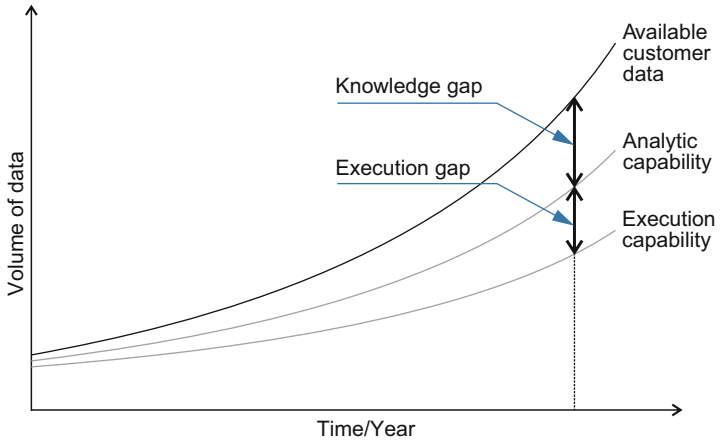



**Fig. 4.3** The knowledge and execution gap. Source: based on presentation by G. Herschel, Gartner Symposium ITXPO 2006

In the above model, what is challenging is pricing. It is due to the fact that data has no intrinsic value, as the *value depends on the context of their use*. There is a number of factors that can affect the value—accuracy and timeliness are the most important (OECD, 2015). Data is more useful and thus valuable if it is accurate and relevant for the particular context in which they are used. As a consequence, "the value of data can perish over time, depreciating as they become less relevant for their intended use" (OECD, 2015). Companies that are able to gather and analyze data in real-time can also get a temporal premium.

There is also another view on the limitations of analytics and it directly refers to exponential growth of data. Such a growth was identified, among others, by Herschel from Gartner (FICO, 2006). There are actually three exponential curves in Fig. 4.3. The 'available customer data' curve represents the growing volume of data resulting from digitizing business. The 'analytic capability' is also growing, as hardware is developing, but it is not capable of analyzing all data. The lowest curve—'execution capability'—reflects how much data is indeed used for decision making. Based on these three curves Herschel defined two gaps: *knowledge gap*—between collection and understanding; and *execution gap*—between understanding and acting.

By the time these gaps were formulated, it was more important to focus on the analysis and then decide what additional data to get. Nowadays, in an era of big data, it seems that first data is captured and then organizations are looking for solutions to gain value from analyzing the possessed data.

Conclusions of Herschel are similar to those of Hackathorn—not all data can be immediately analyzed and even though we have insights, not all data is used to take action. The above two charts can be used to illustrate the 'dark data,' although by that time the concept was not known.

## 4.3 Big Data Definition

Big data can be defined intuitively with regard to its size. In regard to data, size is not everything—this is what George Gallup[1] knew.

Indeed, the first proposed metadefinition by (Jacobs, 2009) focused only on volume: big data should be defined at any point in time as "data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time." Consistent definition was provided by Manyika et al. (2011), who defined it as data for which the "size is beyond the ability of typical database software tools to capture, store, manage, and analyze." Slightly different but still regarding the same feature is definition by Loukides (2010), who defined big data as data for which "the size of the data itself becomes part of the problem." Later big data was a commonly used term to describe data that exceeds the processing capacity of conventional database systems (Wilder-James, 2012). Data can be too big, move too fast, or change the structure so often that traditional databases are useless. To gain value from this data, it is necessary to elaborate methods, techniques, and tools for processing it.

Nowadays, researchers agree that we observe qualitatively and quantitatively new types of data for which new means of its collection, storage, and processing are proposed. The term big data evolved and is typically defined by reference to V's (Chen et al., 2012; Davenport et al., 2012; Hessman, 2013; McAfee & Brynjolfsson, 2012; Sadovskyi et al., 2014; Schroeck et al., 2012):

- *Volume*: refers to collection and processing of vast amounts of all sorts of business-related data. It was initiated by the development and proliferation of digital technologies in organizations. Decreasing the cost of data storage capacity is an incentive to collect even more data.
- *Velocity*: is the speed at which new data is created. It is necessary to use nearly real-time data analytics in order to maximize value of the data. The number of time-bound decision processes is still increasing.
- *Variety*: refers to data structures, content, and formats. It is the result of new sources of data becoming relevant for enterprises. Social media, mobile applications, and sensor networks are among the sources providing companies with valuable but diverse data.

Later on another characteristic of big data appeared as the fourth V—*Veracity* (Hessman, 2013; Schroeck et al., 2012). It refers to the various levels of trust derived from various data sources. It answers the question how reliable is the source and what is the probability that data reflects true values.

Of the V's, volume is the feature that is most discussed. The term is used in conjunction with the following resources: relational data found in the enterprise warehouses of large companies; very large amounts of data that come from scientific experiments; large unstructured collections of data, especially of major web

[1] An American statistician and pioneer of survey sampling techniques for measuring public opinion.

companies. Volume can be quantified in terms of tera, peta, or exabytes. Velocity is harder to quantify. It is mostly connoted to data produced by sensor networks; hence, the characteristic of bit rate of data streams is used. Another interpretation deals with handling the speed of growth in large-scale data repositories. Variety is used to describe heterogeneous nature of data including text, images, audio, and video. They are most challenging to search applications, which need to integrate results across various information types (Hendler, 2013).

The notion of four V's may seem vague at first glance. "The term big data is so amorphous that it hardly has a tangible definition." (Kugler, 2016) This is not the first time when research and technology develop in parallel worlds. For example, the definition of a data warehouse by Inmon (1990) was also vague—it was defined as "a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision making process." It did not prevent enterprises from adopting the concept and developing technology realizing it. Moreover, the developed products like, for example, Microsoft Analytical Services, were not directly referring to the data warehouse definition. 4V's is a conceptual definition, serving as a vision.

As McAfee and Brynjolfsson (2012) noted, big data is similar to analytics and the key differences are observed in the three V's: volume, velocity, variety. The volume and complexity of big data have increased the potential of machine learning and the need for it. Learning from data by developing new algorithms is becoming increasingly important. Organizations need to blend innovative technologies and skills to address these differences and transform their data into useful business information (Sadovskyi et al., 2014). This continuous improvement process is called "computer kaizen" as coined by Google's chief economist Hal Varian (Varian, 2014). The continuous experimentation based on available data is as important for optimizing business processes as continuous improvement in the spirit of *kaizen* has been for production.

## 4.4 Towards Big Data Understanding

Big data as a phenomenon is not only referring to the technology. In order to better understand big data and its role in value creation for enterprises, it is necessary to discussing methodologies, biases, and limitations of big data.

### *4.4.1 Big Data Issues*

Big data issues discussed so far in the literature are mostly related to volume. Bias and representativeness is mentioned among the main issues. Data is representative if it reflects the population. Big data can cause a delusion that we have information about each individual in a population. Usually we do not. Thus, the big volume of

data does not mean that we do not need to worry about representativeness. Boyd and Crawford (2012) warned that methodological issues were still relevant and understanding a notion of 'sample' was even more important. "Quantity of data does not mean that one can ignore the foundational issues of measurement and construct the validity and reliability and dependencies among data" (Kugler, 2016).

As a consequence of non-representativeness, big data usually contains systematic biases (Harford, 2014). Morgan (2015) classified biases that impact the results of big data analysis as: confirmation bias, selection bias, outliers, Simpson's paradox, overfitting and underfitting, confounding variables, and non-normality. Below we describe in detail the biases that are interesting from our research perspective. *Confirmation bias* occurs when there is an intentional or unintentional desire to prove a hypothesis, assumption, or opinion. This means that the analysts usually stop further explorations when they feel that data looks as expected. *Selection bias* occurs when data is not selected randomly or is selected subjectively to suit the purpose of the analysis. Then the population selected does not represent the actual population and the results are skewed. It is usually a matter of study design (Ioannidis, 2005). Surveys are particularly prone to selection bias. First, they are sent to a specific group of people (some of them opt in); second, questions are prepared based on existing assumptions and usually reveal expected results. Increased availability of big data contributes significantly to the selection bias. *Outliers* are values that were measured outside the expected range or do not fit into the normal distribution. They may influence the results of the analysis if not handled properly, e.g., by removal or replacement. It is hard to detect outliers in big data—big datasets make it hard for manual analysis. Moreover, in big volumes rare values seem to be frequent.

Another set of issues is related to inconsistency. Zhang (2013b) distinguished several types of this phenomenon. Temporal inconsistencies refer to temporal attributes, which values can be conflicting. Spatial inconsistencies are observed when datasets contain spatial or geometric dimensions; conflicts can arise from spatial relations between objects. Text inconsistencies occur when the meaning of a message can be in conflict. Functional dependency inconsistencies are related to the modeling of relations in databases (e.g., foreign keys).

Reimsbach-Kounatze (2015) distinguished three types of errors specific for big data: errors caused by poor data quality, errors that come with the inappropriate use of data and analytics, and errors that are caused by the unexpectedly changing environment from which data is collected. The latter issue is especially important for the automation of data analytics. Helland (2011) raised the issue of accuracy but in the context of time necessary to obtain results. By the time we finish calculations the answer might have changed, which is characterized by the phrase "too much to be accurate." His considerations can be referred to granularity, described in Sect. 4.4.2.

Some of the researches point at excessive expectations related to big data (Kugler, 2016). The biggest failures concern forecasting. It is plainly related to ignoring statistical foundations. For example, Google Flu Trends, aiming at the prediction of outbreaks of the disease, failed for two reasons: *too much faith in big data* and *dynamics of algorithms*. First, big data may reflect more phenomena than just the one for which its collected. It should be then correlated with traditional

data collection and analysis, not left alone. Data reliability should be in focus and it is not something that emerges with volume according to law of large numbers. Second, data collection can be wrong. Google have updated the search algorithm many times; therefore, the distribution of typed queries varied. Moreover, their suggestions impacted the behavior of searchers; therefore, data could easily be biased. It resembles the so-called 'observer effect.'[2] There is also another effect—interventions and actions taken based on partial results introduce further bias to data. Finally, regardless of data collection and models used, the phenomenon itself can evolve before it can be caught by the model. Statistical methods for dealing with these failures are becoming increasingly important in big data analytics.

Another manifestation of big data misunderstanding is conviction that the volume of data makes the methodological approach needless. For example, Anderson (2008) argued that technology was so sophisticated that it could discover patterns, trends, and relationships automatically, without providing hypotheses or models. His understanding was that it was not necessary to know why people do something; it was important that people do it and it was possible to track the activities very precisely. "With enough data, the numbers speak for themselves" (Anderson, 2008). The claims were refuted among others by Harford (2014), who stated that statisticians has spent 200 years working on the research methodology and how to avoid many pitfalls when working with data.

### *4.4.2 Big Data Granularity and Self-Similarity*

As already explained, big data is very often associated with data volume. The high volume does not always mean rich content. By the time a disk space was scarce, many compression algorithms were developed. It was driven by observation that data is very often repeated—popular sequences of bits were replaced with a shorter bit representation. The theoretical compression ratio for a data object is determined by the *Kolmogorov complexity*, which is the length of the shortest computer program that produces this object as output (Kolmogorov, 1963).

Here we can also refer to information entropy by Shannon (1948). Entropy refers to disorder or uncertainty. The measure of information entropy is calculated for each possible data value within a given source. It is the negative logarithm of the probability mass function for the value. High-probability events have low information entropy, and low-probability event feature high information entropy, i.e., they carry more 'information.' Thus, less probable values are more interesting (a surprise effect). The information entropy provides an absolute limit on the shortest possible average length of a lossless compression. Both Kolmogorov complexity

[2] The term observer effect refers to changes that the act of observation will make on a phenomenon being observed. In quantum physics, it is related to collapse of the wave function after a measurement has been performed.

and information entropy can be used to characterize *descriptive complexity* of big data.

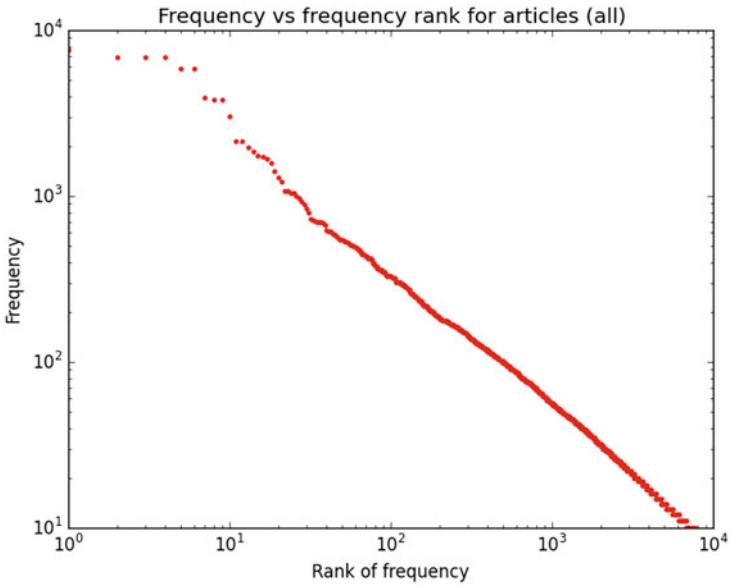
Referring to big data, we need to ask how big it really is. To some extent, big data is an amalgamation of small datasets that are repeated millions of times. Some objects share the same features, some people behave in the same way, some web pages attract similar traffic. The question is what new do we learn if we add additional object, person, or web page.

Another phenomenon characteristic for big data is that it is usually governed by the power law. The power law is a functional relationship between two quantities where one quantity varies as a power of another. In the domain of probability, the quantity is said to follow a power law when the probability of measuring a particular value of some quantity varies inversely as a power of that value (Newman, 2005). It is also known as Zipf's law or the Pareto distribution. It is intriguing that many phenomena in nature follow a power law: distribution of the sizes of cities, the frequency of words in most languages, or the number of friends in a social network. Data that are observed at various scales look similarly—it is like a fractal. It results from the so-called 'scale invariance,' i.e., scaling the argument by a constant factor $c$ causes only a proportionate scaling of the function itself. That is why on the log-log plot it is represented by a straight line—the so-called signature of the power law.

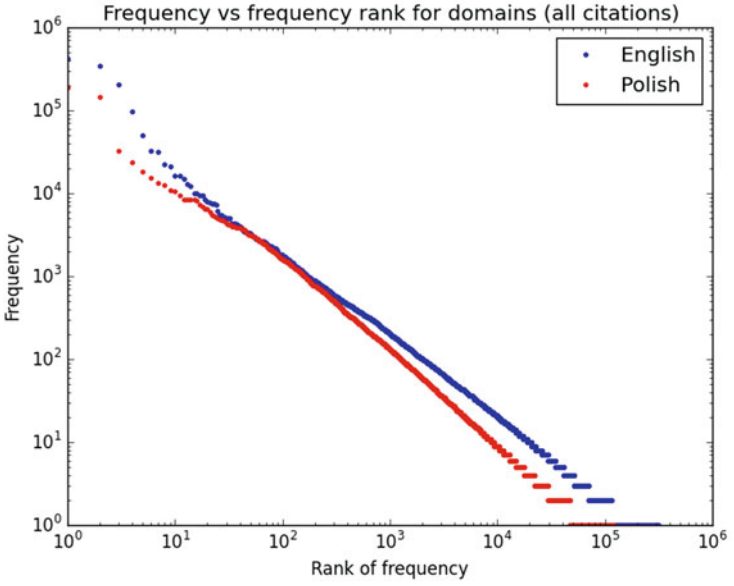
Power law was also discovered in our experiment concerning citations in Wikipedia (Lewoniewski et al., 2017a; Węcel et al., 2016).[3] We have analyzed all references from Polish and English Wikipedia to scientific articles and web pages. Data for the plots were prepared as follows: we have counted how many times each article/domain was cited. The article/domain that was cited the most obtained rank one, and all others were assigned ranks accordingly. The plots show the dependency between frequency of mentions and rank frequency. In order to observe the Zipf's law, we use log-log scale. Figure 4.4a shows results for articles. We can observe that the most popular article (rank $1 = 10^0$) was cited over $7 \cdot 10^3$ times. Articles of rank $7 \cdot 10^3$ were cited at most 10 times (the figure shows only articles cited at least 10 times). Figure 4.4b shows analogous plots for domains. The most popular domain in English Wikipedia was cited over $4 \cdot 10^4$ times whereas in Polish over $2 \cdot 10^5$ times. Domains in English Wikipedia collected more citations than in Polish but the general dependencies in these two languages remain the same.

There are several consequences of the power law. The absolute numbers in the data are not as relevant as the accompanying order of magnitude and only a few things stand out in orders of magnitude. If quantities vary over many orders of magnitude, they cannot be modeled well with a normal distribution and are better modeled by a log-normal or power-law distribution. As noted by Houle (2016), 'exceptional events' are the new normal and as we increase the volume of data such asymmetries in distributions can be a source of bottlenecks. Additionally, the more data is collected, the smaller changes in distributions are observed. Furthermore,

[3] Presented at DBpedia Citation Challenge, during SEMANTiCS 2016 conference in Leipzig, 1st prize award.




(a) articles in Polish Wikipedia




(b) domains in two languages

**Fig. 4.4** Plot of frequency vs. frequency rank (Zipf's law) of Wikipedia citations. Source: Węcel et al., 2016

working with big data might be easier than initially expected because similar conclusions can be drawn from a smaller random sample. Frequent events are not so interesting because they can already be observed in smaller samples. Of course, we can learn something new about entities but the insights for the whole dataset remain the same. What should be interesting for analysts are the 'surprisingly' frequent things that are deviating from theoretical distribution (Kohlschütter, 2011).

The possibility to compress information has also another consequence. Even though there is big data, it may be far less information in the data than necessary to build a stable model. Then stochastic models better account for randomness. The question that needs to be answered is if large amounts of data are really useful. People responsible for analytics can obscure the unreliability of the numbers they use by the beauty of their mathematical models. The association 'garbage in, garbage out' is particularly significant for big data. The analyst "can also miss the big picture in their pursuit of ever more granular data" (Schumpeter, 2011). Concluding, when speaking about big data, we should focus on the analytical potential, and not on the volume itself.

### *4.4.3 Privacy, Ethical, and Social Issues*

Du Gay and Pryke (2002) noted that "accounting tools do not simply aid the measurement of economic activity, they shape the reality they measure." If it is true for such a formalized domain, why not to consider the impact on society. Many years later, social networks, such as Facebook shape a new reality: what we see, what we think, how we interact, even outside the network. The shaping force comes from the content posted by others (Krasnova et al., 2013) or the mechanism itself (Panger, 2016).[4] Big data has to be considered as a socio-technical phenomenon.

When dealing with big data, we need to realize that we indeed deal with data about people. That is why it is important to consider not only the computational perspective but also 'soft' issues. Boyd (2010) considered collecting, aggregating, and analyzing data from social scientists' point of view. He discussed five issues using logic and methodology of social science. First of all, bigger data is not always better data. Here quality matters more than quantity. He stressed the importance of data limits and observing methodologies, e.g., sampling. Second, not all data is created equal. He warned of mistaking the volume of data with accuracy. Third, 'what' and 'why' are different questions. In other words, seeing a pattern does not explain the reasons behind it. The exact question should be known even before data collection. Fourth, we need to be careful of our interpretations. To understand big data, one needs first to understand the methodological processes staying behind analyzing social data. Fifth, just because data is accessible, it does not mean that

[4] In 2014 Facebook manipulated the contents of nearly 700,000 users' news feeds to induce changes in their emotions.

using it is ethical. It was later rephrased as "just because the content is publicly accessible does not mean that it was meant to be consumed by just anyone" (Boyd & Crawford, 2012).

Continuing the discussion about ethics, Boyd (2010) stated that privacy was not about control over data but it was about a collective understanding of the social situation of researched subjects. Sharing of data produced from interaction with a system was considered a violation of privacy. Boyd (2010) was afraid that "our obsession with big data threatens to destabilize social situations" and that is why he proposed five principles: (1) security through obscurity is a reasonable strategy; (2) not all publicly accessible data is meant to be publicized; (3) people who share personally identifiable information are not giving up privacy; (4) aggregating and distributing data out of context is a violation of privacy; (5) privacy is not access control. Methods of securing privacy are described in Sect. 6.4.3.

### *4.4.4 Visualization and Big Data*

Volume of data increasing faster than available analytical techniques causes widening knowledge gap (cf. Sect. 4.2.2). Not only results need to be delivered but also they should be understood. Better understanding of big data is possible through visualization and context. "Big data usually provides a picture so complex that visualization has to be part of the narrative" (Margetts, 2014).

Visualization can also be helpful to avoid misinterpretation of the results. The issues related to understanding of big data can be illustrated with an interesting example provided already in 1973 by Francis Anscombe. He constructed the so-called Anscombe's quartet—four datasets that have nearly identical simple descriptive statistics, like mean, variance, correlation, and regression line, but look very different when presented in the figure. The datasets, each consisting of eleven points, are visualized in Fig. 4.5. The datasets demonstrate the importance of visualization of data on the one hand, and the impact of outliers on statistical properties on the other hand (Anscombe, 1973).

Big data should be presented in an appealing and useful way. It means not only visualization but also providing a broader context. The gap (visualization vs. understanding) can be filled up by data journalism. It involves "finding data sources, interrogating them, matching up across datasets, and visualizing them" (Margetts, 2014). The visualization helps to reveal trends and patterns that would not have been discovered using traditional narrative techniques.

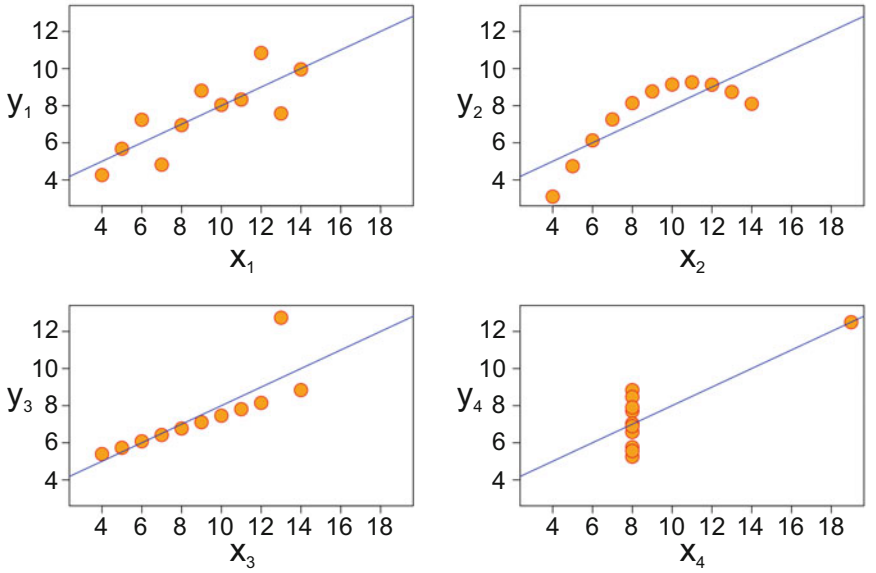



**Fig. 4.5** Visualization of Anscombe's quartet datasets. Source: https://commons.wikimedia.org/wiki/File:Anscombe.svg

## 4.5 Data Resources

### *4.5.1 Big Versus Open*

Open data can be big data, whereas big data is not necessarily open. On the contrary, it is usually proprietary data—its use is restricted to the organization that collected it. This restriction is usually imposed by privacy concerns and trade secret.

In the case such restrictions do not exist, it may be valuable to open up proprietary datasets to others. It does not mean that the whole big data should be made open. It can be opened in an aggregated way or offering computational services in the open algorithm model (see Sect. 6.4.3). OECD (2015) provided a number of reasons why open data can be an optimal strategy even for private companies. As Rufus Pollock stated at the OECD Technology Foresight Forum in October 2012 "The best thing to do with your data will be thought of by someone else," referring to the open data movement.

Big data is not the domain of commercial companies only. Public sector also owns rich and huge data repositories or at least has tools to enforce certain data collection. Big data has the potential to transform a government and can result in new opportunities for a society. Although Janssen and van den Hoven (2015) explicitly referred to combining big, open, and linked data, they did not address the value of this effort. Big and open data can make a government more transparent,

but it also potentially increases control and reduces privacy. A government has to balance the ethical issues (broadening participation, empowering with the inclusion of new voices) versus economic issues (a new source of monetization, crowdsourcing to volunteers as a way to reduce costs) (Sieber & Johnson, 2015).

Margetts (2014) discussed differences and synergies between big and open data. Open data is less homogeneous—it has mixed aims and varying results. It is used mostly in political context as 'transparency' is an attractive label for politicians. Big data is more suitable for rigorous scientific analysis. Although transparency is not exposed in the long term, it may have a more significant impact on transparent government, leading to the 'open-book governance.' Big data has the potential of making government more efficient. Of course, there is also data that is both open and big. It is possible when data is produced to make government transparent and at the same time represents the whole population. Such real-time transactional data can make public organizations both more transparent and more efficient.

Open data offers interesting opportunities for reaching external data. Hendler (2013) used the term 'broad data' to describe the broad spectrum of the vast amount of open data on the Web. The focus should be on open data, not just on the internal data warehouse.

### *4.5.2 Big Data and Semantics*

According to Michael Brodie, there are two challenges related to explosion of data: (1) engineering—efficiently managing data; (2) semantics—doing it meaningfully (Bizer et al., 2011).

Big data and linked data by some people are considered separate worlds. Others believe that combining these worlds can bring certain measurable benefits. In the 'big data' world heterogeneity is an issue that has to be overcome. Semantic technologies can serve as a foundation layer for data integration efforts and can help to make sense of all large data pools in the enterprise and on the Web. It can also link structured and unstructured information. Semantic graphs can become the backbone of any information architecture and can enable entity-centric views on enterprise information and data (Blumauer, 2016).

Linking can take place on two levels: *schema* and *data*. This can also be distinguished as linking of data vs. metadata. Linking metadata allows better dataset discovery. LOD cloud can be an example of linking on the metadata level, but analytics should definitely benefit from linking on the data level.

There are already some successful examples of combining Hadoop with semantic web. For example, Mami et al. (2016) introduced SeBiDa (Semantified Big Data Architecture), a proof-of-concept implementation of the architecture using Big Data components, such as Apache Spark and MongoDB. Their approach was very technical, focusing on architectural issues. They addressed a variety of data by looking at schemata and metadata. There was little about data translation itself.

### *4.5.3 Alternative Data*

The value of data analytics lies in the combination of both internal and external data (Dumbill, 2012; Redman, 2008). Critical information for making a decision very often can be found outside of a company (Biesdorf et al., 2013). The value of data is highly context-dependent, which means that can change significantly when confronted with other dataset (OECD, 2015). Data put in a broader context can reveal additional insights.

Many companies use open data about markets (inflation, unemployment, forecasts), industry (banking, insurance, retail), weather, or traffic. Unfortunately, such data as a commodity does not offer opportunities for gaining competitive advantage.

The real potential is hidden in the so-called *alternative data*. They are a non-trivial data sources that provide information valuable to some entities—those who know how to combine it with their proprietary data. A good example for alternative data is data collected by satellites offered, for example, by Ursa Space Systems Inc. (Ursa[5]). The company is selling reports prepared with space-based data from radar satellites. Their vision is to unlock the potential of data by providing insights derived from satellite imaging. Satellite data can also be combined with other sources. For example, in SIMMO project AIS satellite data was enriched with information from dedicated maritime portals (Abramowicz et al., 2016; Stróÿzyna et al., 2016).

There are also companies that play a role of infomediary and collect sources of alternative data. For example, Quandl built their business model on offering access to alternative data. They have defined the spectrum of diffusion for alternative data (Kamel, 2016). 'Fully diffused' data is data available for everybody, well-known, and popular among enterprises needing it for specific decisions. It is also fully commoditized. Examples include stock prices and economic indicators. Second type is 'diffusing' data—its more and more popular among enterprises and is moderately commoditized. We can count here sentiment data on products or weather data. The third and most sophisticated is 'nascent' data, which encompasses newly discovered datasets, where few have access to. This data is also the most valuable, although the way it can be used is not obvious.

## 4.6 Data Unification Challenge

In this section, we discuss the requirements and foundations necessary to allow the unification of big, open, and linked data. We put this effort in a context of knowledge management. We discuss benefits of bringing semantics in a form of linked data to the enterprise data that is additionally combined with external data. We can look at the process of organizing open data as an analogy. The organization challenge can be

[5] http://ursaspace.com/.

divided into two aspects: (1) providing well-structured and organized metadata for the datasets and (2) structuring and organizing the datasets themselves (Tygel et al., 2015). The first aspect was already discussed in Sect. 3.6.2. The second aspect is the focus of semantic lifting of data (van der Waal et al., 2014; Węcel, 2014). In the context of big data, we need to capture the right elements from available resources.

### *4.6.1 External Data Integration*

Looking for the best ways to integrate structured and unstructured information is primarily a research topic (Hendler, 2014). Our work on this topic dates to 2002 when we considered supplying a data warehouse with external information (Abramowicz et al., 2002). By that time we proposed to use RDF to represent metadata of both textual documents and data warehouse profiles, the long-term needs of business analysts built on the contents of the data warehouse.

Data warehousing solved the issue of accessing a multitude of separate data sources in a homogeneous manner. It made it possible to collect, organize, and access data from a central repository. Prior to data warehousing, users who needed to consolidate multiple sources had to reach each application separately and make an effort to combine the results.

There is a growing requirement of data analysts to reach data outside of the organization (Hendler, 2014). The most important drivers are increasing amount of data, technologies for linking data across datasets, and increasing need to integrate structured and unstructured data. This growing volume of freely available open data on the Web, both structured and unstructured, is referred to as 'broad data.' It consists of millions of datasets made available both by governments and enterprises and supplemented by various datasets and vocabularies from linked open data cloud, including particularly DBpedia, YAGO, and Wikidata (see Sect. 3.4.1). Therefore, it is emphasized that broad data requires linking the metadata, not the data. In the case of 'broad data', variety is more important than scale. *Volume* is already addressed by (distributed) technologies like MapReduce. The issue is actually the *variety* aspect of big data Vs. It seems that there is a better understanding of these issues among enterprises. For example, in the German-speaking economy, big data is defined as the collection, storage, and analysis of heterogeneous data (Gronau et al., 2013). In this definition only the *variety* is exposed.

Within the context of broad data Hendler (2014) proposed an approach to work with data, referred to as DIVE:

- *Discovery.* Data search is complex within enterprise, it is even harder outside—on the Web. Use of various lightweight metadata is proposed to support dataset discovery. The simple approach to discovery is faceted search (e.g. topic, location, data, dataset format). We need also to consider the difficulty of working with multiple search engines where each offers a proprietary protocol. Thus,

more complex searches should involve federated catalogs, similarity of the meaning of keywords, and development of domain-specific metadata.
- *Integration.* Generally, datasets are created independently, without the intention to be used together. Ad hoc integration makes sense when we can manually correlate data from various sources. It is usually not the case in enterprise settings where repeatable processes are necessary. Semantic is required to create data mash-ups. Sometimes an additional dataset is necessary in order to integrate the remaining others. Use of data protocols is another approach to integration. A standardized API or OData protocol should simplify access to data. Various representation issues are tackled with semantics.
- *Validation.* Check if there are any problems resulting from integration. This is actually checking if the previous step was done correctly. Very often it is not straightforward and requires the involvement of domain experts, e.g., the exact meaning of data categories, how data was collected and classified. Problems can be seen only when data is integrated. It is suggested to use correlation between various datasets—this is similar to our approach (Lewoniewski et al., 2016, 2017b; Wçel & Lewoniewski, 2015)
- *Exploration.* It is important to improve human-data interaction. Exploration should take place as early as possible so that hypotheses can be tested before the full analytics pipeline is entered.

Regarding the integration of external data, there is still an untapped potential. Gronau et al. (2013) surveyed about 7000 companies in German speaking countries and confirmed that all surveyed companies had recognized the relevance of data for business success. Majority of data is derived from internal enterprise systems: 82% from ERP and 61% from CRM. Already 52% of enterprises believe that ERP systems do not deliver sufficient data but still only 26% use external market data. Moreover, the use of external data is rather experimental. The success of integrating external data into the company is rated as good or very good by only around 13% of companies in the German-speaking business sector. The propensity to use external data depends on the branch. For example, all manufacturing companies base their decisions on experience rather than data. Over 70% of banks and insurance companies cultivate data-based decision-making. They are more inclined to integrate external data.

### *4.6.2 Wiig Knowledge Management Model*

Knowledge to be useful and valuable must be organized. It can be organized differently for different uses (Wiig, 1993). Knowledge is persisted in a more generic form and an appropriate perspective is defined later, based on the cognitive tasks to be conducted.

Wiig Knowledge Management Model, one of the major theoretical knowledge management models, has four dimensions (Dalkir, 2011): completeness, connected-

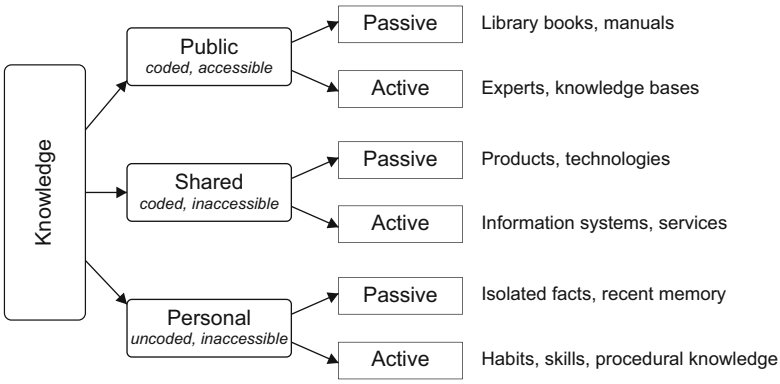



**Fig. 4.6** Wiig's hierarchy of knowledge forms. Source: based on Dalkir, 2011

ness, congruency, and perspective and purpose. *Completeness* explains how much relevant knowledge is available from a given source, whether a human mind or knowledge base. We need to know that knowledge is available, otherwise it is not possible to use it. *Connectedness* refers to various relations defined between knowledge objects. The more connected a knowledge base is, the more coherent it can be (facts may be verified from many sources), thus also the more valuable. *Congruency* means that facts, concepts, values, perspectives, judgments, etc. are consistent, i.e., there are no logical inconsistencies or internal conflicts. It can be achieved through validation. Finally, *perspective and purpose* states that knowledge is usually organized in these two dimensions. There can be different perspectives on the same knowledge content, what can lead to fragmentation of knowledge.

Wiig (1993) distinguished following *forms* of knowledge: public knowledge, shared expertise, and personal knowledge. Detailed break-down is presented in Fig. 4.6. We can also distinguish *types* of knowledge (Wiig, 1993): (1) factual knowledge: data, casual chains, measurements, readings; (2) conceptual knowledge: systems, concepts, perspectives; (3) expectational knowledge: judgments, hypotheses, expectations; (4) methodological knowledge: reasoning, strategies, decision-making methods. Forms of knowledge along with types of knowledge form a matrix. The matrix with examples is presented in Table 4.2.

### *4.6.3 New Theory of Data*

Helland (2011) observed that many principles of classical SQL databases eroded when people tried to handle too much data. Data is no more locked in a database, and unlocking it, i.e., releasing to the world, changes semantics of data. Looking from the other side, there are many sources of data leading to inconsistent schemata. Thus, we need to deal with different or even unknown semantics. Patterns in data are

**Table 4.2** Wiig's knowledge management matrix

| Form of knowledge | Type of knowledge | | | |
|---|---|---|---|---|
| | Factual | Conceptual | Expectational | Methodological |
| Public | Number of COVID cases | Climate change | Discounted prices for black Friday | Check the weather before hitting the trail |
| Shared | List of top 10 customers | Segment X offer higher prices | Big companies are less flexible | Call the customer a week after the purchase |
| Personal | The preferred artist | Company X produces good mobile phones | Increase in salary after successful certification | Take medicine before breakfast |

Source: own examples based on (Wiig, 1993)

inferred from continuously examining data—emerge from use, and not from design. Finally, there is too much data to be accurate—by the time we finish calculations, the data is already updated.

Relational database solutions, which started in the 1970s, capture only part of the problem. Helland (2011) called for a new theory and taxonomy of data. First of all, any new theory and taxonomy of data should subsume the classical database. As an extension towards unlocked data, it should provide identity and optional versions. It should be then possible to identify versions that contributed to the knowledge in an enterprise. Changes to the source should trigger a recalculation in other sources. If a particular data source cannot be used for legal reasons, then an enterprise is not allowed to use the knowledge derived from it. Derivation can introduce also lossiness. It should be possible to differentiate the loss stemming from derivation from the loss resulting from inaccuracies of the source data. The last concern is about attribution by pattern—patterns can be derived from attributes that are derived from patterns, and so on.

Stein and Morrison (2014) explored the practical integration challenges, inspired mostly by the cloud technology. Within the data layer, the following challenges were identified: data silos, data proliferation, rigid schemata, high data warehousing cost, new and heterogeneous data types. Hadoop data lakes, late binding, and metadata provenance tools were identified as emerging solutions for the mentioned problems. Overall, the challenges for new data integration are in the following areas:

1. Data—big volumes, different sources, available as a stream.
2. Processing—transformation, inferences, assumptions.

### *4.6.4 Variety and Discoverability*

There are a lot of open data initiatives (see Sect. 2.3). The more portals decide to publish data, paradoxically, the more difficult it is to find data. Such data is very often published 'as-is', in a less rigorous form, usually as a dump from a structured or unstructured repository, without deep reference to other data (Hendler, 2013). It is very often ambiguous, distorted, temporally undetermined, or debated. The aspect of variety is inherent to open data.

Many organizations do not even have structures to organize internal data. According to Shah et al. (2012), less than 44% of employees knew where to find the information they needed for their work. This phenomenon can only get stronger when big data is concerned. Big data is predominantly described through the first V—volume. This, however, is a technical issue, which can be easily handled. Big data technologies like, for example, Hadoop, were developed with horizontal scaling in mind. There is, however, a much bigger challenge from the management perspective—the variety (Hendler, 2014; Heudecker & Kart, 2014).

Nowadays, even more data is out of the control of individual users. There is a growing need for interaction with heterogeneous data. Beside processing big volumes of data, there is also an important problem how to normalize, integrate, and transform the data from many sources (Knoblock & Szekely, 2013). Enterprises expect an easy way to discover external data, integrate this data with own assets, and to perform analysis leading to valuable insights (Hendler, 2014).

The more heterogeneous the data is, the more difficult it is to integrate it. For example, Heudecker and Kart (2014) showed that companies for their big data projects prefer to analyze their own transactional data rather than social media sources. Percentage of companies analyzing transactions increased from 70% in 2013 to 79% in 2014. During the same period, the number of companies studying data from social media dropped. To get value from less structured sources, it is not sufficient to increase storage, but new methods and tools are necessary.

## 4.7 The Model for Linked-Data-Based Unification of Data

Enterprises increasingly perceive linked data principles as a solution to integration of heterogeneous information systems. Based on the requirements and conducted analyses, we come up with the following steps to build a linked-data-based unification of data.

### 4.7.1 General Benefits of Linked Data

Big data is not self-explanatory. The interpretation can be easily lost in a large mass of raw information (Bollier, 2010). Moreover, taken out of context, big data loses its meaning and value (Boyd & Crawford, 2012). Big data is also perceived as separate islands of data that needs to be connected to produce value.

The idea of a graph-oriented representation of information is very close to understanding of information by people. It allows obtaining a broader context by following links to existing data. Linked data can also provide the structure necessary to process data from various points of view. One of the key applications of linked data is supporting consumers in discovering and integrating data (Schmachtenberg et al., 2014).

Big data requires clear and explicit definitions of information elements. Linked data can not only provide contextual information but also allow to trace back a piece of information to its definition (Voskuil, 2017). Contextualizing can concern topics, places, people, or events mentioned in collected data. By linking to the root meaning, it is then much easier to reuse data. When concepts in raw disparate data are annotated with semantic relations, organizations can achieve data-driven decision making (Yankova, 2016). Some initiatives failed due to a combination of two factors: big data inconsistency and human errors in interpreting data (Kugler, 2016). Both challenges can be addressed by the involvement of linked data.

### 4.7.2 Emerging Structuring

There are some important observations by Helland (2011) concerning the structuring: in very large and loosely coupled systems, the *descriptive* schemata are observed, not *prescriptive*. In a prescriptive schema, the data is forced into a fixed format that is understood by the writing system and is consistently shared by all receivers. Descriptive schema means that when the data is written, the writer also describes what was the intention of the message, which should facilitate understanding. In the large and disconnected system, it is impractical to maintain a prescriptive schema. Maintaining consistency becomes ever more difficult. Thus, extremely large systems naturally evolve towards inconsistency when not governed.

Besides defining every schema in advance, it is also impractical to structure everything. It is easier to store data in raw format and apply feature extraction techniques later when certain data is needed. Brute-force processing is one of the possibilities to extract value from large volumes of raw files. This idea is a foundation of a data lake (see Sect. 4.8.2).

Semantic and engineering limits of data structuring and integration were also observed by Bizer et al. (2011). For relational databases integration, we usually assume that they are semantically homogeneous, but this condition rarely holds.

Meaningful integration of data requires manual intervention. This cannot be done in the big scale.

Schemaless approach does not mean that we should work without the structure. Such structures can be provided ad hoc, also depending on the specific temporary needs of analysts. The schema known from relational databases is just replaced with dataset summarization (Bizer et al., 2011). For this summarization it is possible to apply methods from data profiling (see Sect. 3.6.1).

Grounding data processing in RDF data model has certain benefits. There is no need to define a prescriptive schema before data is loaded. Data structure can be defined descriptively as data is used. SPARQL allows to define the implicit structure of a dataset, and various schemata can be used for various analyses. SPARQL offers also federated query capability so data can be extracted from various sources. "The incremental use of RDFS and SPARQL together gives you an agile alternative to the expensive, time-consuming steps associated with planning a typical data warehousing project" (DuCharme, 2013). Technology slowly follows the conceptual requirements. Schemaless flexibility of data technologies becomes increasingly affordable (Bizer et al., 2011).

### *4.7.3 Specific Contribution of Linked Data*

**Provision of Identity** Many problems in the field of information integration are caused by the lack of global identifiers. Sometimes there are even no agreed standards for identification. For example, a product can be identified using a number of standards: EAN (European Article Number), GTIN (Global Trade Item Number), or UPC (Universal Product Code), not mentioning proprietary or domain-specific solutions, e.g., ASIN (Amazon Standard Identification Number) or ISBN (International Standard Book Number).

Linked data significantly contributes to identity provision. After all, URIs were designed to identify any kind of object or concept (Berners-Lee, 2006). When linked data is concerned an *entity resolution* has to be used as the step. The following features concerning entities can be fulfilled:

- entity identity—URIs can serve as names for things,
- entity access—URIs that are based on HTTP can provide additional information for people after dereferencing,
- entity structure—standards, such as XML and JSON can be used to provide structured information for machine processing,
- entity integration—links in a form of URI to other entities can provide additional context and allow discovery of new things.

**Sensemaking** "Sensemaking is a generic phrase that refers to processes of interpretation and meaning production whereby individuals and groups interpret and reflect on phenomena" (Brown et al., 2008). Sensemaking is closely related to organization

of data. On the one hand, organization emerges through sensemaking (Weick et al., 2005). On the other hand, to make sense is to organize, using the technology of language (labeling, categorizing) to identify, regularize, and routinize memories into plausible explanations (Brown et al., 2008). Ferro and Osella (2012) proposed types of data elaboration, i.e., operations performed on retrieved data in order to make sense of it. The following steps were distinguished: (1) data aggregation, (2) data structuring and classification, (3) data geo-referencing, (4) data validation, (5) data mash-up, and (6) visual analytics. Lycett (2013) defined *datafication* as an information technology driven sensemaking process. This term was coined to describe the consequences of the growing generation, collection, and storage of digital information, which is referred to as the digital revolution. In a datafied world, data is first perceived mathematically, and the context for understanding is established later. To make sense of text, images, and videos in a machine-processable way, we need rich semantics, provided in the form of structured metadata, standardized vocabularies, or ontologies. Sensemaking is conducted within a single dataset, potentially with additional context. It is a prerequisite for the last step—interlinking.

**Context** Data is not generic—its understanding requires a context. The context is connected with identity, because it allows to distinguish certain artifacts. By the introduction of a context, an entity-centric view on contents is possible. Such context is even more important for big data, but it can be easily lost in the scale of data or when data is reduced to fit a certain analytical model. "Managing the context in light of big data will be an ongoing challenge" (Boyd & Crawford, 2012).

Open data and linked data in particular offers a lot of contextual sources (see Sect. 3.4). Blumauer (2016) used a term 'knowledge lens' to refer to a context. He perceived the context as a way to support personalized information processing. A knowledge graph is used not only to retrieve a business object but also all accompanying relations of this object, even if they are not explicitly stated, thus helping users to better understand the underlying meaning of business objects. The same data put into various contexts provides various analysis perspectives for knowledge workers. It may be desirable to show only aspects of the context relevant in a given situation. He noted that such an approach is very demanding for a semantic layer on top of the data layer, particularly when data is distributed and heterogeneous.

**Interlinking** Entity-centric views on enterprise information means that people search for facts, not documents. Therefore, such a view is much more natural when dealing with information in the enterprise. This allows to bundle information chunks together in order to build an overall picture of a problem, decision, etc.

Metadata is particularly important in the interlinking task. Open data ecosystem usually requires different types of descriptions for ensuring data interoperability (Zuiderwijk et al., 2014). In the case of interlinking with big data, the main concern is the variety of data. Metadata can improve the ability to find and interpret open

data. It also makes linking easier, allows avoiding unnecessary duplication, and facilitates collaboration among stakeholders.

The real value lies in the connections we can make within big data. The links between data and value are even more important when reaching outside to externally held data (Fennell, 2012).

### *4.7.4 Validity of the Model*

"The goal of linked data is to replace traditional app-data silos with a universal integration platform to provide globally contextualized information using global identifiers, authentication, authorization, storage, and privacy" (Berners-Lee, 2012). Below we discuss to which extent linked data can address data unification challenges posed by the discussed models and enterprise solutions.

Big data was defined in terms of 4Vs. Regarding these features, the following features can be addressed by linked data:

- Volume—not to be addressed by linked data. In this case, efficiency is a key. Various data architectures are proposed, e.g., columnar store, to optimize access to data and analytical capabilities. Introduction of linked data can only increase the overhead.
- Velocity—not to be addressed by linked data for the above mentioned reasons.
- Variety—this is the area where the use of linked data is highly desirable. Linked data is about giving meaning to the data; therefore, the variety can be addressed by the introduction of standardized vocabularies. Unambiguous understanding is achieved through the linking process.
- Veracity—there is a potential for linked data use. If veracity is understood as a feature of the source, then the possibility to verify information in various sources offers certain benefits.

When speaking of big data, efficiency should be in focus. Introduction of semantics for data integration demonstrates certain benefits with regard to understanding but at the same time increases requirements for the infrastructure. There is a hesitation whether an increased functionality justifies a decrease in efficiency. Bizer et al. (2011) proposed to meaningfully combine database technologies with reasoning capabilities of semantic systems. Sometimes SQL can be sufficient to solve a problem but there will be more and more real-world problems where logic more expressive than SPARQL is necessary.

With regard to Wiig Knowledge Management Model the following dimensions are addressed by linked data:

- Completeness—concerns the availability of relevant knowledge. Linked data contributes to this dimension in two ways. First, it offers a plethora of datasets that can serve as background knowledge (see Sect. 3.4.1). Second, it addresses the issue of *discoverability*, which means that requested data can be much easier

to find. Not only internal data within the enterprise can be better described but also employees can reach external sources, offered among others by open data (see Sect. 2.6).

- Connectedness—refers to relations between knowledge objects. This dimension is explicitly addressed by the nature of linked data, which is *linking* various pieces of data (see Sect. 3.2). Linking can be defined on metadata level, dataset level, or even data level. The last is probably a much deeper implementation of connectedness than expected by Wiig. Various knowledge bases can be interlinked in such a way that they form a consistent knowledge graph, blending the boundaries between datasets (see Sect. 4.8.4).
- Congruency—describes the state where there are no conflicts or inconsistencies in knowledge representation. Such a feature is hard to maintain in a big data environment, where a *variety* is one of the inherent features. Linked data can be used to tackle this issue. First, consistency can be checked by using reasoning in terms of Description Logics if an ontology is represented in OWL. Second, linked data can make various representations homogeneous by the introduction of a standardized vocabulary (see Sect. 3.3).
- Perspective and purpose—indicate the need for different perspectives on the same knowledge content. Linked data allows description of knowledge objects with various attributes, whether asserted or inferred. The attributes can refer to various purposes of using knowledge. The perspectives can be provided by contextual ontologies (see Sect. 3.4.2).

Linked data can also address the requirements of the modern theory of data outlined by (Helland, 2011). Below we describe how certain elements can be implemented using technologies developed around linked data:

- Identity and versions—in RDF model URI serves as an identifier. Versions are not handled directly but some solutions do exist—we can use different knowledge base URIs or named graphs. Changes in versions can be described using provenance ontologies.
- Derivation—the work on provenance is advancing, and some dedicated vocabularies are developed. Derived knowledge can be generated from inferencing using OWL DL reasoning, or from rules, using, for example, SWRL—Semantic Web Rule Language (Horrocks et al., 2004). In an advanced scenarios, various mapping techniques can be used. For example, R2RML—RDB to RDF Mapping Language can be used to map data in relational databases to the RDF data model (Das et al., 2012). A superset of this W3C-recommended mapping language is provided by RML—RDF Mapping Language, which can source data from any structure format (Dimou, 2014). Modern mapping languages are expressed in RDF themselves; therefore, maintenance of consistency of mapping is much easier.
- Lossyness of the derivation—there is no inaccuracy introduced by derived data. On the contrary—we can observe data enhancement in the derivation process. Linked data provides a means to uniquely identify things. The disambiguation process can assure that we will refer to proper things. Modern mapping languages

are expressed in RDF themselves; therefore, the maintenance of consistency of the mapping is much easier.
- Attribution by patterns—patterns are a domain of querying and inferencing. For the purpose of new graph construction based on existing attributes, a SPARQL query language can be used (Prud'hommeaux & Seaborne, 2008).
- Classical locked database data—relational data can be easily mapped to the RDF using, for example, R2RML.

We should not think that the full switch to semantics is possible. It is more probable that old and new systems will coexist. Semantics should serve as an addition, not the replacement. Bizer et al. (2011) discussed success metrics for semantic technologies. They concluded that switching to 'semantic-everything' is unrealistic. Enterprises have heavily invested in traditional infrastructure: middleware, data warehouses, and data mining tools. These assets can generate even more value when enriched with open data. Bonch proposed that the introduction of semantic technologies to a company can be admitted successful, based on relaxed criteria. It is sufficient that employees combine multiple LOD datasets and link them to their own data, even though the RDF is not the primary model. Data integration is one of the most significant costs in IT and every simplification in this domain can bring observable savings.

## 4.8 Modern Enterprise Solutions Leveraging Linked Data

Big data is mostly perceived through technology roots. Indeed, many technologies were developed by companies that were faced with specific big data challenges, like unprecedented volume of data (Turck, 2016). To bring value to enterprises, it is necessary to change the focus away from data and solve technical problems. It is necessary to move towards business and processes and focus on how to *use* big data effectively: how to access, reconcile, query, or enrich. The business justification should be in first place.

The change of focus from technology to business is already observed. Manyika et al. (2011) argued that data is becoming a *factor of production*, like physical or human capital. Companies expect new quality of insights (Buchholtz et al., 2014). To benefit from big data, companies need to overcome five management challenges of big data (McAfee & Brynjolfsson, 2012):

- Leadership—it is not sufficient to possess more data—the management needs to ask correct questions. The vision and human insight is necessary, and business leaders should spot opportunities.
- Talent management—effective working with data requires specific skills. There is a growing demand for data scientists.
- Technology—tools to handle big data have improved recently and are usually available as open source. The prices of hardware also dropped.

- Decision making—in the big data era, information is created in other places than decisions are made. There is a problem of trust and transfer.
- Company culture—it is necessary to move from instinct-driven to data-driven activities. Decisions should be based on numbers.

Below we present areas and specific solutions that we believe may be highly impacted by a combination of big, open, and linked data. We show what benefits are expected or are already realized.

### *4.8.1 Data Governance*

Big data success is not assured by implementing one piece of technology (Turck, 2016). It is necessary to set up technologies, people, and processes. The processes should encompass data life cycle, i.e., capturing, storing, cleaning, querying, analyzing, and visualizing. The organizations realized that they need a process to manage the data deluge (Anderson, 2008).

Data governance is the approach to organize data. Architecture, data management, and processing can be simplified by the uniform representation of requirements, data, and rules. Data governance should be driven by strong business opportunities, both strategic and operational, and should not be limited by data infrastructure (De Leenheer, 2016). Enterprises usually identify the need for data governance when they implement master data management, self-service BI and analytics, or regulatory compliance (see also Sect. 4.8.3). A data steward is a new role within an organization responsible for governing (sourcing, use, and maintenance) all types and forms of data in an organization.

According to OECD (2015), for an effective data governance regime the following should be considered: data access and reuse, data portability and interoperability, data linkage and integration, data quality and curation, data ownership and control, and data value and pricing. One of the key questions for data governance is 'what' to govern. Haruray (2016) enumerated the following resources:

1. Data resources: data items, data models, table structures, reference datasets.
2. Technology resources: systems, databases, repositories.
3. Business resources—business terms, glossaries, synonyms, taxonomies for common understanding within a domain.
4. Governance resources: business rules, policies, standards.

Implementation of enterprise-wide data management programs requires certain technologies. Ziadeh (2016) specified five types of tools: data quality, entity resolution, data search, reporting and analysis, and data modeling. *Data quality* tools are used to improve the quality of data. They support data stewards in assessing data content and resolving quality issues. *Entity resolution* tools support in matching data from different sources, agencies, or jurisdictions. To do so, they establish key master records for key entities like people and organizations. As

a result they strengthen enterprise data sharing. *Data search* tools are used for supporting users in finding and accessing high-quality data and information to answer business questions. *Reporting and analysis* tools facilitate generation of graphs, dashboards, and reports. More advanced tools offer capabilities to extract insights from possessed data (e.g., data mining, text mining). *Data modeling* tools are used for describing data attributes, data relationships, semantics at various levels of abstraction and detail, and business rules governing it.

The even stronger need for data governance across an entire organization is spurred by big data. Yet, data governance can be best implemented with semantic technologies (Harper, 2015). Data governance is not only about data, it also concerns schema. From schema it is close to ontological thinking. Structure of data, which in the case of big data is all but fixed, can be represented with vocabularies and taxonomies. This makes linked data a natural complement of big data. Terminology, vocabularies, and ontological models are an essential element to implement big data governance in a sustainable manner (Harper, 2015). Many organizations just started with light-weight ontologies, implementing business glossaries in SKOS. Governance depends on preserving meaning, whether within an application or between different sources, and semantics is best designated to achieve this goal.

Combining big, open, and linked data is an interesting research field. "Data governance will likely become more important in which redundant data and wrong data is reduced and data can be reused for various purposes" (Charalabidis et al., 2015).

### *4.8.2 Data Lakes*

The data lake concept is motivated by the self-service movement. The term was coined by James Dixon, CTO of Pentaho, to describe a new solution for working with enterprise data.

> If you think of a datamart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples. (Dixon, 2010)

Dixon confronted a well-established technology used in enterprises—data warehousing—with a new approach to work with data. Unlike a data warehouse, the data lake stores data in its raw form. There are no prior transformations, no cleansing, no validation. As such, the lake can serve as a staging area in the form of Operational Data Store (ODS) for the data warehouse. More differences are collected in Table 4.3.

A data lake (DL) is not suitable for every company; many of them will still rely on a data warehouse (DW). DL and DW have various target groups. Users of a data lake should be more aware of limitations, e.g., not cleansed, partial, or biased data.

**Table 4.3** Comparison of a data warehouse and a data lake

| Data warehouse | Data lake |
|---|---|
| Structured data designed for specific purposes | All data, any format, also unstructured |
| Mostly quantitative metrics data | All types of data, also qualitative |
| Only data used for reporting | All data in native format |
| Rigid design | More flexible design |
| Predefined reports | Ad hoc reports |
| Restricted flexibility of reporting (OLAP) | Big data analytic of any type |
| Used by business professionals | Used by data scientists |

Source: own elaboration

From technological point of view the data lake is a distributed file system. The most popular implementation is based on HDFS, Apache Hadoop Distributed File System, which is used to build a single repository of data (Stein & Morrison, 2014). There are also solutions offered by cloud companies: Azure Data Lake by Microsoft and Amazon's S3.

The Hadoop data lake preserves data in its original form and captures changes to data and contextual semantics, which may be especially useful for compliance and internal audit (Stein & Morrison, 2014). Another organizational implication is the difference in deployment and maintenance costs. The lake can be an order of magnitude less expensive than data warehouses on a per-terabyte basis (Stein & Morrison, 2014). The data lake does not require scrutiny characteristics for the data warehouse.

An important observation for data models in the data lake is that they emerge with usage over time; they are not imposed in advance. The consistency of data is achieved only in applications using data. This is similar to late binding in programming languages. Such an approach, i.e., structuring only when data is needed, challenges the data organization. The organizational structure becomes even more important. Data may not be integrated but has to be discoverable. For this purpose, the metadata layer has to be present to describe content of the datasets in the lake. Metadata is usually decoupled from its underlying data and stored independently. Nevertheless, organizations should focus on *semantic consistency* in upstream applications and data stores instead of information consolidation in a data lake (Yankova, 2016).

A step towards organizing a schemaless data was done by Franz,[6] which built a semantic data lake (Woodie, 2015). The semantic data lake leverages a structure that a graph database can provide to unstructured data. Technically, data is stored in a key-value store in HBase and is indexed with Franz's graph database, making it accessible with SPARQL queries. According to Aasman, the Franz CEO, a semantic data lake opened Hadoop analytics to external linked data, when a semantic graph managed by AllegroGraph was introduced (Woodie, 2015).

[6] http://www.franz.com/.

Fowler (2015) discussed the distinction between the data lake and the data shore. The lake stores raw data and the shore handles curated information. Both of them are the evolution of the traditional data warehouse model. While the lake can be compared to ODS, the data shore is resembling the data warehouse with its structural representation of data offered by development of NoSQL databases. Lake shores have much bigger potential for supporting analytics in the company as it represents contextual governed data. This is what Azure is offering with U-SQL language. Lake shore is an extension that allows to add semantics. Organized data in known structures can be used for analytics and knowledge discovery.

Storing data in the data lake does not guarantee reuse of data. The data lake does not offer any support for data governance. Without governance, the data lake will end up as a collection of disconnected data silos (Gartner, 2014) and it will not be possible to manage the insights and reuse findings (White, 2015). Managing the life cycle of data is necessary to avoid the so-called 'data swamps' when data becomes unusable and loses any operational value. The main challenge is not creating the data lake but taking advantage of the new data access (Stein & Morrison, 2014).

One of the predictions for data lakes is its analytic expansion (Martin, 2016). There is a hope that once the organization's assets are contained in RDF graph, it is possible to overcome the 'dark data' phenomenon. The context and meaning of data is then understood before the analysis is carried out, which allows the selection of analytics to be performed and refining its results. If the decision tends toward formalizing the data analysis processes, "it is beneficial to move beyond a data lake concept quite quickly in order to develop a more robust logical data warehouse strategy" (Gartner, 2014).

### *4.8.3 Semantic Compliance*

The main motivation for discussion about compliance is increasing complexity of data ecosystems. Furthermore, the regulatory regime for financial institutions observes increasing complexity. The traditional approach to compliance encounters the translation issue. Regulations are specified in natural language, and as such are not accessible for machines. What is encoded in 'computer' language is not easily understood by business users. Additionally, there are a lot of interconnected systems that require mappings and defined transformations.

The complexity of financial systems is no longer manageable with conventional approaches. The increasing compliance burden forces financial institutions to search for new ways to cope with this complexity. One of the initiatives is Semantic Compliance in Finance, which is understood as an ontological approach to regulatory reporting and oversight.[7] Within this initiatives Financial Regulation Ontology (FRO) was defined. FRO aligns finance and legal issues as it is based on two indus-

[7] http://finregont.com/.

try standards: (1) FIBO—Financial Industry Business Ontology for representing clients, securities, funds, derivatives, etc.; (2) LKIF—Legal Knowledge Interchange Format for representing law, SEC rules, forms, submissions and responses. FRO is encoded in OWL (Ziemer, 2016).

FRO is also populated with the full text of US laws and regulations so that users have all information at hand. Defined classes link to the actual paragraph of the regulation and they are used to express regulatory conditions. Ontology editors usually have reasoners, which allow to evaluate class restrictions and retrieve matching instances.

Compliance checking can be conducted using emerging language SHACL (Shapes Constraint Language). SHACL is a language for describing and constraining the contents of RDF graphs (Knublauch & Kontokostas, 2017). 'Shape' is understood as a kind of constraint that is imposed at a given RDF node. SHACL can be used for validation of data, definition of interfaces, and generating user interfaces. There are plans to publish the translation of FIBO in SHACL.

There are opinions that SHACL is better suited for compliance checking than plain OWL (Kontokostas et al., 2016). There are several reasons for this. The semantics in the open world assumption differs from what is expected in the data model specification. For example, a cardinality restriction in OWL is used for inferencing rather than for constraint checking. SHACL is also more expressive and is extensible. Advanced features of SHACL cover rules and functions. The semantic environment can be completed with higher level rules and constraint languages like SWRL. SHACL has no constructive capability and so can only perform validation, whereas Web Rule Language (SWRL) has well-defined semantics that can be used by reasoners.

### *4.8.4 Enterprise Knowledge Graphs*

Enterprise Knowledge Graph (EKG) is a semantic network of concepts, properties, individuals, and links representing and referencing foundational and domain knowledge relevant for an enterprise (Galkin et al., 2016a). Such a graph is a backbone of Linked Enterprise Data (LED). LED is a concept of incorporating the benefits of semantic technologies into enterprise IT environments. Thus, EKG is the next stage in the development of knowledge management systems.

Galkin et al. (2016b) argued that EKG might be considered as an embodiment of LED so that corporate information management can be lifted to a semantic level. They also positioned EKGs in enterprise data architectures. The paper is based on a survey of existing enterprise information systems that implement certain parts of the EKG functionality.

One of the aspects important within EKG concept is an approach to data fusion. The approaches are: unified, transitionary, and federated (Galkin et al., 2016a). In a *unified approach* data is totally fused. This required sophisticated data fusion mechanisms but offers certain advantages. An enterprise has more

control over data and its quality. Moreover, data querying is significantly faster. A *federated approach* is suitable for enterprises that make heavy use of public LOD sources. The flexibility comes at a cost of higher overhead for query expansion and reasoning. A *transitionary approach* is advisable when data security plays a vital role. Unfortunately, access control and federated querying drastically decreases the performance.

## 4.9 Summary

Big data is typically large, unstructured, real-time, and represents a sample of the whole population. Data collection and proper organization may hold the key to better predictions. However, as data is collected quickly in streams, can it be done correctly? There is a belief that large datasets allow a higher form of intelligence that can generate insights that were previously impossible. Current organizations are rather not prepared to handle this vast amount of information—they suffer from background noise and information overload. Improved data organization techniques can have a better impact on the effectiveness of prediction than improvements in modeling techniques. These are the conclusions both from Hackathorn's value-time curve and efforts to combine big data with linked data.

The ability to leverage information is crucial to drive business success. The growing volume, velocity, and variety of data is still a challenge. Moreover, delivery of data and analytics put increased demand on technology but is also interesting from the economic point of view. Uncertainty of data is also growing and adds to the complexity of data processing. Data from an always increasing number of heterogeneous, internal and external sources becomes more relevant for the business. Among various V's characterizing big data, it is necessary to speak of the data value.

Organizations often underestimate the value of data and information they hold. They are just discovering the potential of linked data technologies to solve the complexity of big data. By complexity we understand rather variety than volume. Advances in big data analytics combined with the expansion of open and linked data are freeing information that was once locked in internal information systems. This change in the approach enables greater accountability and improved decision making. Elaborated big data is sometimes referred to as smart data, which provide actionable information from harnessing the challenges posed by the volume, velocity, variety, and veracity of big data. Proper management of 4 V's of big data allows freeing resources to focus more thoroughly on value of data. Therefore, the answer for the question whether open data and linked data can be beneficial for enterprises is definitely positive.

# References


Abramowicz, W., Kalczyński, P. J., & Węcel, K. (2002). *Filtering the Web to feed data warehouses.* Springer Verlag London. ISBN: 1-85233-579-3. (page 90)

Abramowicz, W., Filipiak, D., Małyszko, J., Stróżyna, M., & Wçel, K. (2016). Maritime domain awareness system supplied with external information: Use-case of the SIMMO system. In Szubrycht, T. (Ed.), *Proc. of the 7th International Scientific and Technical Conference NATCON—Naval Technologies for Defence and Security* (pp. 1–20). Polish Naval Academy. Gdynia. ISBN: 9788393015054. (page 89)

Anderson, C. (2008). *The end of theory: The data deluge makes the scientific method obsolete.* http://www.wired.com/2008/06/pb-theory/ (pages 82, 101)

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician, 27*(1), 17–21. ISSN: 0003-1305. https://doi.org/10.1080/00031305.1973.10478966 (page 86)

Barnaghi, P., Sheth, A., & Henson, C. (2013). From data to actionable knowledge: Big data challenges in the Web of Things. *IEEE Intelligent Systems, 28*(6), 6–11. ISSN: 1541–1672. https://doi.org/10.1109/MIS.2013.142 (pages 73, 76)

Berners-Lee, T. (2006). *Linked data—design issues.* http://www.w3.org/DesignIssues/LinkedData.html (visited on 2016-03-25). (page 96)

Berners-Lee, T. (2012). *Growing big linked data from seed: building a demo.* http://bigdata.csail.mit.edu/node/57 (visited on 2017-11-08) (page 98)

Biesdorf, S., Court, D., & Willmott, P. (2013). Big data: What's your plan? *McKinsey Quarterly* 3. (page 89)

Bizer, C., Boncz, P., Brodie, M. L., & Erling, O. (2011). The meaningful use of big data: Four perspectives—four challenges. *SIGMOD Record, 40*(4), 56–60. (pages 88, 95, 96, 98, 100)

Blumauer, A. (2016). *Introducing a graph-based semantic layer in enterprises.* https://www.linkedin.com/pulse/introducingagraphbasedsemanticlayerin-enterprises-andreas-blumauer (visited on 2017-10-01). (pages 88, 97)

Bollier, D. (2010). *The promise and peril of big data* (p. 61). Washington: The Aspen Institute. ISBN: 0898435161. (page 95)

Boyd, D. M. (2010). *Privacy and publicity in the context of big data.* http://www.danah.org/papers/talks/2010/WWW2010.html (pages 85, 86)

Boyd, D. M., & Crawford, K. (2012). Critical questions for big data. *Communication & Society, 15*(5), 662–679. ISSN: 1369-118X. https://doi.org/10.1080/1369118X.2012.678878 (pages 73, 81, 86, 95, 97)

Brown, A. D., Stacey, P., & Nandhakumar, J. (2008). Making sense of sensemaking narratives. *Human Relations, 61*(8), 1035–1062. ISSN: 0018-7267. https://doi.org/10.1177/0018726708094858 (pages 96, 97)

Buchholtz, S., Bukowski, M., & Śniegocki, A. (2014). *Big and open data in Europe. A growth engine or a missed opportunity?* (p. 114) Warsaw: demosEUROPA. ISBN: 978-83-925542-1-9. (page 100)

Charalabidis, Y., Janssen, M., & Krcmar, H. (2015). Introduction to the big, open, and linked data (BOLD), analytics, and interoperability infrastructures in government minitrack. In *Proceedings of the Annual Hawaii International Conference on System Sciences* (p. 2074). (page 102)

Chen, H., Chiang, R., & Storey, V. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165–1188. (page 79)

Dalkir, K. (2011). *Knowledge management in theory and practice.* Cambridge, Massachusetts: MIT Press. (pages 91, 92)

Das, S., Sundara, S., & Cyganiak, R (Eds.) (2012). R2RML: *RDB to RDF mapping language.* W3C Recommendation. https://www.w3.org/TR/r2rml/ (visited on 2017-11-08). (page 99)

Data-Driven Development (2015). *Data-driven development. pathways for progress*. http://reports.weforum.org/data-driven-development/ (page 75)

Davenport, T., Barth, P., & Bean, R. (2012). How 'Big Data' is different. *MIT Sloan Management Review, 54*(1), 43–46. (page 79)

De Leenheer, P. (2016). *The rise of the chief data officer (CDO).* https://www.collibra.com/blog/the-rise-of-the-chief-data-officer-cdo/ (visited on 2017-11-02). (page 101)

Digital Britain. (2009). *Cm 7650* (p. 238). London: Department for Business Innovation & Skills. ISBN: 978-01-017650-2-2. (page 75)

Dimou, A. (2014). *RDF mapping language specification.* http://rml.io/spec.html (visited on 2017-11-08). (page 99)

Dixon, J. (2010). *Pentaho, hadoop, and data lakes.* http://www.pentaho.com/blog/2010/10/15/pentaho-hadoop-and-data-lakes (visited on 2017-11-03). (page 102)

Du Gay, P., & Pryke, M. (2002). *Cultural economy: Cultural analysis and commercial life.* London: Sage. (page 85)

DuCharme, B. (2013). What do RDF and SPARQL bring to big data projects? *Big Data, 1*(1), 38–41. ISSN: 2167-6461, 2167-647X. https://doi.org/10.1089/big.2012.0004 (page 96)

Dumbill, E. (Ed.). (2012). *Planning for big data. A CIO's handbook to the changing data landscape* (p. 84). O'Reilly Media. (page 89)

Fennell, P. (2012). *Linked data underpins the value of big data.* http://broadcast.oreilly.com/2012/05/linked-data-underpins-the-valu.html (visited on 2016-06-27). (page 98)

Ferro, E., & Osella, M. (2012). Business models for PSI re-use: A multidimensional framework. In *Workshop using open data: policy modeling, citizen participation, data journalism* (pp. 1–5). Brussels. (page 97)

FICO. (2006). *Live from gartner symposium ITxpo—analytics: Action based on integrating processes and applications.* http://www.fico.com/en/blogs/analytics-optimization/live_from_gartn_8/ (visited on 2017-10-04). (page 78)

Fowler, M. (2015). *DataLake.* https://martinfowler.com/bliki/DataLake.html (visited on 2017-10-07). (page 104)

Frischmann, B. M. (2013). *Infrastructure: The social value of shared resources* (p. 436). Oxford University Press, ISBN: 978-0199975501. (page 77)

Galkin, M., Auer, S., & Scerri, S. (2016a). Enterprise knowledge graphs: A backbone of linked enterprise data. (page 105)

Galkin, M., Auer, S., & Scerri, S. (2016b). Enterprise knowledge graphs: A survey. In *37th International Conference on Information Systems.* (page 105)

Gartner. (2014). *Gartner says beware of the Data Lake fallacy.* http://www.gartner.com/newsroom/id/2809117 (visited on 2017-09-07). (page 104)

Gibson, W. (2010). Data, data everywhere. *The Economist. Special report: Managing information, 394*(8671), 3–5. (page 75)

Gronau, N., Fohrholz, C., & Weber, N. (2013). *Wettbewerbsfaktor Analytics - Reifegrad ermitteln, Wirtschaftlichkeitspotenziale entdecken.* Abschlussbericht: Universität Potsdam. (pages 90, 91)

Gronau, N., Thim, C., & Fohrholz, C. (2016). Business Analytics in der deutschen Praxis. *Controlling, 28*(8–9), 472–479. ISSN: 0935-0381. https://doi.org/10.15358/0935-0381-2016-8-9-472 (page 74)

Hackathorn, R. D. (1998). *Web farming for the data warehouse* (p. 384). Morgan Kaufmann. ISBN: 978-1558605039. (pages 74, 77)

Harford, T. (2014). Big data: Are we making a big mistake. *Significance, 11*, 14–19. https://doi.org/10.1111/j.1740-9713.2014.00778.x (pages 81, 82)

Harper, J. (2015). *A semantic approach to big data governance.* https://www.datanami.com/2015/12/10/a-semantic-approach-to-bigdata-governance/ (visited on 2017-09-09). (page 102)

Haruray, A. (2016). *United States of data—what to govern.* https://www.linkedin.com/pulse/united-states-data-what-govern-ashishharuray/ (visited on 2017-11-02). (page 101)

Helland, P. (2011). If you have too much data, then "Good Enough" is good enough. *Queue, 9*(5), 40. (pages 73, 81, 92, 93, 95, 99)

Hendler, J. (2013). Broad data: Exploring the emerging web of data. *Big Data, 1*(1), 18–20. ISSN: 2167-6461, 2167-647X. https://doi.org/10.1089/big.2013.1506 (pages 80, 88, 94)

Hendler, J. (2014). Data integration for heterogenous datasets. *Big Data, 2*(4), 205–215. ISSN: 2167-6461. https://doi.org/10.1089/big.2014.0068 (pages 90, 94)

Hessman, T. M. (2013). Putting big data to work. *Industry Week, 262*(4), 14–18. (page 79)

Heudecker, N., & Kart, L. (2014). *Survey analysis: Big data investment grows but deployments remain scarce in 2014*. Gartner. (page 94)

Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B., & Dean, M. (2004). *SWRL: A semantic web rule language. combining OWL and RuleML. W3C member submission.* https://www.w3.org/Submission/SWRL/ (visited on 2017-11-08). (page 99)

Houle, P. (2016). *Data lakes, data ponds, and data droplets*. http://ontology2.com/the-book/data-lakes-ponds-and-droplets.html (visited on 2017-09-09). (page 83)

IBSG. (2011). *The Internet of Things.* Internet Business Solutions Group. http://share.cisco.com/internet-of-things.html (visited on 2015-03-18). (page 73)

Inmon, W. H. (1990). *Building the data warehouse* (1st ed). Wiley. (page 80)

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), 0696–0701. ISSN: 15491277. https://doi.org/10.1371/journal.pmed.0020124. arXiv: 0208024 [gr-qc] (page 81)

Jacobs, A. (2009). The pathologies of big data. *Queue, 7*(6), 10. https://doi.org/10.1145/1563821.1563874 (page 79)

Jacobson, R. (2013). *2.5 quintillion bytes of data created every day. How does CPG & Retail manage it?* https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytesof-data-created-every-day-how-does-cpg-retail-manage-it/ (visited on 2017-11-02). (page 73)

Janssen, M., & van den Hoven, J. (2015). Big and open linked data (BOLD) in government: A challenge to transparency and privacy? *Government Information Quarterly, 32*(4), 363–368. (page 87)

Kamel, T. (2016). *Alternative data—the developing trend in financial data.* https://blog.quandl.com/alternative-data (visited on 2017-11-03). (page 89)

Kiron, D., & Shockley, R. (2011). Creating business value with analytics the leading question. *MIT Sloan Management Review, 53*(1), 57–63. (page 73)

Knoblock, C., & Szekely, P. (2013). Semantics for big data integration and analysis. *2013 AAAI fall symposium series* (pp. 28–31). ISBN: 9781577356424. https://doi.org/10.1609/aimag.v36i1.2565 (page 94)

Knublauch, H., & Kontokostas, D. (Eds.). (2017). *Shapes constraint language (SHACL). W3C recommendation.* https://www.w3.org/TR/shacl/ (visited on 2017-11-03). (page 105)

Kohlschütter, C. (2011). *Why the current obsession with big data.* https://www.quora.com/Why-the-current-obsession-with-big-data/answer/Christian-Kohlsch%C3%BCtter (visited on 2017-10-05). (page 85)

Kolmogorov, A. (1963). On tables of random numbers. *Sankhyā Ser. A., 25*, 369–375. MR: 178484. (page 82)

Kontokostas, D., Mader, C., Dirschl, C., Eck, K., Leuthold, M., Lehmann, J., & Hellmann, S. (2016). Semantically enhanced quality assurance in the JURION business use case. In H. Sac, E. Blomqvist, M. D'Aquin, C. Ghidini, S. P. Ponzetto & C. Lange (Eds.), *The Semantic Web. Latest Advances and New Domains: 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, May 29–June 2, 2016, Proceedings* (pp. 661–676). Springer International Publishing. https://doi.org/10.1007/978-3-319-34129-3_40 (page 105)

Krasnova, H., Wenninger, H., Widjaja, T., & Buxmann, P. (2013). Envy on Facebook: A hidden threat to users' life satisfaction? In *11th International Conference on Wirtschaftsinformatik* (pp. 1–16). Leipzig. https://doi.org/10.7892/boris.47080 (page 85)

Kugler, L. (2016). What happens when big data blunders? *Communications of the ACM, 59*(6), 15–16. ISSN: 00010782. https://doi.org/10.1145/2911975 (pages 80, 81, 95)

Lewoniewski, W., Węcel, K., & Abramowicz, W. (2016). Quality and importance of wikipedia articles in different languages. In G. Dregvaite & R. Damaševičius (Eds.), *Information and Software Technologies: Proc. of 22nd International Conference, ICIST 2016* (Vol. 639, pp. 613–

624). Communications in computer and information science. Springer International Publishing Switzerland. https://doi.org/10.1007/978-3-319-46254-7_50 (page 91)

Lewoniewski, W., Węcel, K., & Abramowicz, W. (2017a). Analysis of references across wikipedia languages. In R. Damaševičius, V. Mikašytỹe (Eds.), *Information and Software Technologies: Proc. of 23rd International Conference, ICIST 2017, Druskininkai, Lithuania, October 12–14, 2017* (Vol. 756, pp. 561–573). Communications in computer and information science. Cham: Springer International Publishing. ISBN: 978-3-319-67642-5. https://doi.org/10.1007/978-3-319-67642-5_47 (page 83)

Lewoniewski, W., Węcel, K., & Abramowicz, W. (2017b). Relative quality and popularity evaluation of multilingual wikipedia articles. *Informatics, 4*(4). ISSN: 2227-9709. https://doi.org/10.3390/informatics4040043 (page 91)

Llinas, J. (2015). Information fusion process design issues for hard and soft information: Developing an initial prototype. In *Intelligent methods for cyber warfare* (pp. 129–149). (page 75)

Loukides, M. (2010). *What is data science?* O'Reilly Radar. https://www.oreilly.com/ideas/what-is-data-science (visited on 2017-04-02). (page 79)

Lycett, M. (2013). Datafication: Making sense of (Big) data in a complex world. *European Journal of Information Systems, 22*(4), 381–386. ISSN: 0960-085X. https://doi.org/10.1057/Ejis.2013.10 (page 97)

Mami, M. N., Scerri, S., Auer, S., & Vidal, M. E. (2016). Towards semantification of big data technology. In S. Madria, & T. Hara (Eds.), *Big Data Analytics and Knowledge Discovery: 18th International Conference, DaWaK 2016, Porto, Portugal, September 6-8, 2016, Proceedings* (pp. 376–390). Springer International Publishing. https://doi.org/10.1007/978-3-319-43946-4_25 (page 88)

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute. (pages 79, 100)

Margetts, H. (2014). Data, data everywhere: open data versus big data in the quest for transparency. In N. Bowles, J. Hamilton, D. Levy (Eds.), *Transparency in politics and the media: Accountability and open government* (pp. 167–178). London: I.B. Tauris. ISBN: 978-1-78076-675-1. (pages 86, 88)

Martin, S. (2016). *The future of data lakes: Four predictions for 2016*. https://www.scientificcomputing.com/blog/2016/01/futuredata-lakes-four-predictions-2016 (visited on 2017-10-07). (page 104)

McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review, 90*(10), 61–68. (pages 79, 80, 100)

Morgan, L. (2015). *7 common biases that skew big data results.* http://www.informationweek.com/big-data/big-data-analytics/7-common-biases-that-skew-big-data-results/d/d-id/1321211 (visited on 2017-10-05) (page 81)

Morrison, A. (2015). *Enterprises hedge their bets with NoSQL databases.* http://usblogs.pwc.com/emerging-technology/enterprises-hedge-their-bets-with-nosql-databases/ (visited on 2017-11-02) (page 74)

Naef, E., Muelbert, P., Raza, S., Frederick, R., Kendall, J., & Gupta, N. (2014b). *Using mobile data for development*. Cartesian and Bill & Melinda Gates Foundation. (page 76)

NASCIO. (2016). *Better decisions, better government: Effective data management through a coordinated approach.* Lexington: National Association of State Chief Information Officers. (page 74)

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics, 46*(5), 323–351. https://doi.org/10.1080/00107510500052444 (page 83)

OECD. (2015). *Data-driven innovation. Big data for growth and well-being* (pp. 1–456). Paris. ISBN: 9789264229358. https://doi.org/10.1787/9789264229358-en (pages 73, 78, 87, 89, 101)

Open Data Institute. (2016). *How to make a business case for open data*. https://theodi.org/guides/how-make-business-case-open-data (visited on 2016-06-27). (page 75)

Panger, G. (2016). Reassessing the Facebook experiment: Critical thinking about the validity of Big Data research. *Information Communication and Society, 19*(8), 1108–1126. ISSN: 1369118X. https://doi.org/10.1080/1369118X.2015.1093525 (page 85)

Plachkov, A. (2016). Soft data-augmented risk assessment and automated course of action generation for maritime situational awareness by. Master thesis. University of Ottawa. https://doi.org/10.20381/ruor-294 (page 76)

Prud'hommeaux, E., Seaborne, A. (Eds.). (2008). *SPARQL query language for RDF W3C recommendation.* https://www.w3.org/TR/rdf-sparqlquery/ (visited on 2017-11-08). (page 100)

Redman, T. C. (2008). *Data driven: Profiting from your most important business asset* (p. 272). Harvard Business Review Press. ISBN: 978-1422119129. (page 89)

Reimsbach-Kounatze, C. (2015). The proliferation of big data and implications for official statistics and statistical agencies. *OECD Digital Economy Papers, 245*, 3–39. ISSN: 2071-6826. https://doi.org/10.1787/5js7t9wqzvg8-en (page 81)

Sadovskyi, O., Engel, T., Heininger, R., Böhm, M., & Krcmar, H. (2014). Analysis of big data enabled business models using a value chain perspective. In *Multikonferenz Wirtschaftsinformatik (MKWI 2014).* Paderborn, Germany. (pages 79, 80)

Schmachtenberg, M., Bizer, C., & Paulheim, H. (2014). Adoption of the linked data best practices in different topical domains. In *The Semantic Web—ISWC 2014: 13th International Semantic Web Conference.* (pp. 245–260). Springer International Publishing. ISBN: 978-3-319-11964-9. https://doi.org/10.1007/978-3-319-11964-9_16 (page 95)

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., & Tufano, P. (2012). *Analytics: The real-world use of big data. How innovative enterprises extract value from uncertain data.* Executive report. IBM Institute for Business Value. (page 79)

Schumpeter. (2011). *Building with big data.* http://www.economist.com/node/18741392 (visited on 2017-10-05). (page 85)

Shah, S., Horne, A., & Capellá, J. (2012). Good data won't guarantee good decisions. *Harvard Business Review, 90*(4). (page 94)

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x (page 82)

Sieber, R. E., & Johnson, P. A. (2015). Civic open data at a crossroads: Dominant models and current challenges. *Government Information Quarterly, 32*(3), 308–315. ISSN: 0740-624X. https://doi.org/10.1016/j.giq.2015.05.003 (page 88)

Stein, B., & Morrison, A. (2014). *The enterprise data lake: Better integration and deeper analytics.* PwC Technology Forecast. (pages 93, 103, 104)

Stróÿzyna, M., Małyszko J,Wçel, K., Filipiak, D., & Abramowicz, W. (2016). Architecture of maritime awareness system supplied with external information. *Annual of Navigation, 23*, 135–149. https://doi.org/10.1515/aon-2016-0009 (page 89)

Taylor, J. (2012). *Decision latency revisited.* http://jtonedm.com/2012/11/21/decision-latency-revisited/ (visited on 2017-10-01). (page 77)

Turck, M. (2016). *Is big data still a thing? (The 2016 big data landscape).* http://mattturck.com/the-new-gold-rush-wall-street-wantsyour-data/ (visited on 2017-10-04). (pages 100, 101)

Tygel, A., Auer, S., Debattista, J., Orlandi, F., & Campos, M. L. M. (2015). *Towards cleaningup open data portals: A metadata reconciliation approach* (p. 8). https://doi.org/10.1109/ICSC.2016.54. arXiv: 1510.04501 (page 90)

van der Waal, S., Węcel, K., Ermilov, I., Janev, V., Milošević, U., & Wainwright, M. (2014). Lifting open data portals to the data web. In S. Auer, V. Bryl, S. Tramp *Linked open data—creating knowledge out of interlinked data* (Vol. 8661, pp. 175–195). Lecture notes in computer science. Springer International Publishing. ISBN: 978-3-319-09845-6. https://doi.org/10.1007/978-3-319-09846-3_9 (page 90)

Varian, H. R. (2014). Beyond big data. *Business Economics, 49*(1), 27–31. ISSN: 0007-666X. https://doi.org/10.1057/be.2014.1 (page 80)

Voskuil, J. (2017). *The business case for linked data.* https://www.linkedin.com/pulse/business-case-linked-data-jan-voskuil/ (visited on 2017-10-07). (page 95)

Węcel, K. (2014). Public procurement in linked open data paradigm. In J. Gołuchowski, & A. Frączkiewicz-Wronka (Eds.), *Studia Ekonomiczne. [Economic studies]* (Vol. 199, pp. 338–348). ISSN: 2083-8611. (page 90)

Węcel, K., & Lewoniewski, W. (2015). Modelling the quality of attributes in wikipedia infoboxes. In W. Abramowicz (Ed.), *Business information systems workshops* (Vol. 228, pp. 308–320). Lecture notes in business information processing. Springer International Publishing. ISBN: 978-3-319-26761-6. https://doi.org/10.1007/978-3-319-26762-3_27 (page 91)

Węcel, K., Lewoniewski, W., & Sobociński, P. (2016). *DBpedia citation challenge. (not only) polish citations in wikipedia: Analysis, comparison, directions.* https://www.slideshare.net/KrzysztofWecel/dbpedia-citationchallenge-not-only-polish-citations-in-wikipedia-analysis-comparison-directions (visited on 2017-10-05). (pages 83, 84)

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization Science, 16*(4), 409–421. ISSN: 1047-7039. https://doi.org/10.1287/orsc.1050.0133. arXiv: z0022 (page 97)

White, A. (2015). *A Data Lake without any information governance is a data cesspool.* http://blogs.gartner.com/andrew_white/2015/09/09/a-data-lake-without-any-information-governance-is-a-datacesspool/ (visited on 2017-10-07). (page 104)

Wiig, K. (1993). *Knowledge management foundations.* Arlington, TX: Schema Press. (pages 91, 92, 93)

Wilder-James, E. (2012). *What is big data?* https://www.oreilly.com/ideas/what-is-big-data (visited on 2017-10-05). (page 79)

Woodie, A. (2015). *Hadoop, triple stores, and the semantic data lake.* https://www.datanami.com/2015/05/26/hadoop-triple-storesand-the-semantic-data-lake/ (visited on 2017-11-05). (page 103)

Yankova, M. (2016). *Building linked data bridges to fish in data lakes.* https://ontotext.com/linked-data-bridges-data-lakes/ (visited on 2017-10-06). (pages 95, 103)

Zhang, D. (2013a). Granularities and inconsistencies in big data analysis. *International Journal of Software Engineering and Knowledge Engineering, 23*(6), 887–893. ISSN: 0218-1940. https://doi.org/10.1142/S0218194013500241 (pages 74, 75)

Zhang, D. (2013b). Inconsistencies in big data. In *Proceedings of the 12th IEEE International Conference on Cognitive Informatics and Cognitive Computing,* ICCI*CC 2013 (pp. 61–67). https://doi.org/10.1109/ICCI-CC.2013.6622226 (page 81)

Ziadeh, A. (2016). *Practical advice for building a data management program.* https://gcn.com/articles/2016/08/11/nascio-data-management.aspx (visited on 2017-10-01). (page 101)

Ziemer, J. (2016). *Financial regulation ontology.* http://finregont.com/fro/html_widoco/index-en.html (visited on 2017-09-07). (page 105)

Zins, C. (2007). Writing information literacy assessment plans: A guide to best practice. *Journal of the American Society for Information Science and Technology, 58*(4), 479–493. https://doi.org/10.1002/asi.20508 (page 74)

Zuiderwijk, A., Janssen, M., & Davis, C. (2014). Innovation with open data: Essential elements of open data ecosystems. *Information Polity, 19*(1–2), 17–33. ISSN: 18758754. https://doi.org/10.3233/IP-140329 (page 97)

# Chapter 5
# Macroeconomic Aspects of Data Value


## 5.1 Introduction

Data itself does not have value. It is useless unless it is applied in business activities. The value is created by its use.

Data was recognized as "innovation currency in the digital age" (Digital Britain, 2009) and defined as "an economic raw input almost on par with capital and labor" (Gibson, 2010). It is also "the new capital of global economy" (Deloitte, 2016). Data and information are also considered the lifeblood of the knowledge economy. Regardless of the definition, there is an enormous pressure to exploit data. As data is monetizable, there are businesses built around the collection, control, and processing of data. Nevertheless, data can also be a public good and we can observe a growing availability of user-generated, non-profit content on the Web.

In the currently observed so-called 'outcome economy' the focus is shifted from selling things to selling results (Daugherty et al., 2015). In a classical approach, the value of information is the amount a decision maker would be willing to pay for information prior to making a decision. The above statements suggest that getting value out of a data revolution is something obvious, or even trivial. It seems that the introduction of technology should be promptly seen in the bottom line of the company. That was not always the case. As Solow (1987) once used to say "you can see the computer age everywhere but in the productivity statistics." This statement was later reviewed and Solow changed his mind: "you can now see computers in the productivity statistics" (Uchitelle, 2000). The results needed time and the critical mass had to be achieved. Valuation is even more difficult when we consider new phenomena, such as open data. For example, it is difficult to evaluate the value of open data initiatives. From the definition, everybody can access data anonymously, for free, at any time, and from any place (Yang et al., 2015).

One of the challenges for macroeconomics is evaluation of the socioeconomic impact of data across the economy. This leads to the second challenge, which is measuring the contribution of data to gross domestic product (GDP) growth. So far,



K. Węcel, *Big, Open and Linked Data*, Business Information Systems,
https://doi.org/10.1007/978-3-031-07147-8_5

the value of data is poorly captured in economic statistics and financial reports. It is also insufficiently appreciated by organizations and individuals (OECD, 2013a).

The goal of the chapter is to review value-related issues of data. We focus both on direct value and value produced by multiplier effects. Particularly, there is a lack of methods for valuation of external data in the form of open data and linked data. One of the possible solutions are two-sided markets, typical for the Web, which are defined as economic networks having two distinct user groups that provide each other with network benefits. We need to analyze these phenomena when reaching for open and linked data, which are distributed mostly thanks to the development of the Web.

## 5.2 Macroeconomic Impact

Macroeconomic studies create models that estimate the impact of phenomena on an economy as a whole. They determine a financial value for economic effects, such as better consumer decision-making, optimized business operations, and how existing or new infrastructures can be used to maximize the overall social benefit (Tennison & Hardinges, 2015).

The role of data has transitioned from the support of business decisions to becoming a good in itself. Facilitating access to information offers certain benefits. Communities across the world make use of open data to deliver social, economic, and political impacts across a variety of sectors (Enabling the Data Revolution, 2015). Greater social value is created with greater use of common resources. Such an increase is possible because data is a non-rivalrous good. It thus deserves a closer look from the macroeconomic point of view. "Just as the supply of basic physical infrastructure is essential to the 'traditional' economy, so the supply of basic information 'infrastructure' is essential to the 'information' economy" (Pollock, 2009).

### *5.2.1 Statistics Collected*

In scientific articles as well as in market reports and government documents, there are a lot of various numbers showing the predicted macroeconomic impact of big, open, and linked data. The analyses were performed in different years and used various methodologies; therefore, the predictions are far from being consistent. Below we present the most popular sources of evidence for the benefits to be brought by data.

Probably the most cited report by Manyika et al. (2013) estimated that open data, understood as data provided by public administration and private companies, can help create a value of $3 trillion a year in seven areas of the global economy. The mentioned sectors were: education, transportation, consumer products, electricity,

oil and gas, health care, and consumer finance. Similar numbers can be found in (Enabling the Data Revolution, 2015)—open data has an economic potential of Sects. 5.3–5.5 trillion both for data-driven businesses and end-users of the services.

Focusing only on EU economy, Tinholt (2013) projected aggregate economic impact from applications based on open data across the EU27 economy to be €140 billion annually. Another report valued the direct market size of open data in EU 28+ at €55.3 billion for 2016, with a growth potential of 36.9% by 2020 (European Data Portal, 2015).

In terms of GDP growth in Europe, it will be significantly impacted by the implementation of open data policies. Buchholtz et al. (2014) estimated increase in GDP of 1.9% by 2020, i.e., €206 billion. The growth would not be homogeneous across Europe: 2.2% in the Northern European countries as opposed to 1.6% in the Southern European countries. This supports the already mentioned division of countries into Global North and Global South in (Data-Driven Development, 2015).

The situation of UK, one of the forerunners of open data, was researched by Deloitte (2013). Deloitte valued the broad economic value of the data held by the public sector in the UK in 2011 to be around £5 billion per year. The narrow, i.e., without wider societal effects, value to consumers, businesses, and the public sector was approximately £1.8 billion in 2011. Total economic and social value of the market for public sector information was valued between £6.2 billion and £7.2 billion. Of this amount, between £15–58 million was contributed by time savings by changing the behavior based on real-time transport data.

It is interesting also to compare the direct value of data with the overall effect. GDP in the EU28 countries could increase by €10 billion in 2020 if datasets were opened. If open data were combined with big data capabilities for data-driven decisions, the total value of the contribution of open data to GDP would be about €100 billion in 2020 (Buchholtz et al., 2014; Deloitte, 2016).

There are also some analyses concerning specific sectors. Information management sector is growing faster than the software industry. According to (Gibson, 2010), the business of information management was estimated to be worth more than $100 billion and growing at almost 10% a year, whereas the whole software business was growing at only 5%. Regarding the geographical information, the German market for geoinformation in 2007 was estimated at €1.4 billion, 50% higher than in 2000 (Tinholt, 2013). In the Netherlands, the geosector accounted for 15,000 full-time employees in 2008 (ACIL, 2008). In Australia, over 31,400 people were directly employed in the spatial information industry in 2008 (ACIL, 2008).

Moving to the employment market, one of the direct and short-term economic impacts of open data is job creation (Tinholt, 2013). 25,000 jobs directly related to open data will be created in Europe between 2016 and 2020 (European Data Portal, 2015). A larger impact and long-term benefits can be derived from the dissemination of skills around big data (Tinholt, 2013). Open data can thus support governments in efforts at increasing employment.

According to OECD (2015), "Labor in the twenty-first century will increasingly rely on data and analytics." The report analyzed various data-related work tasks.

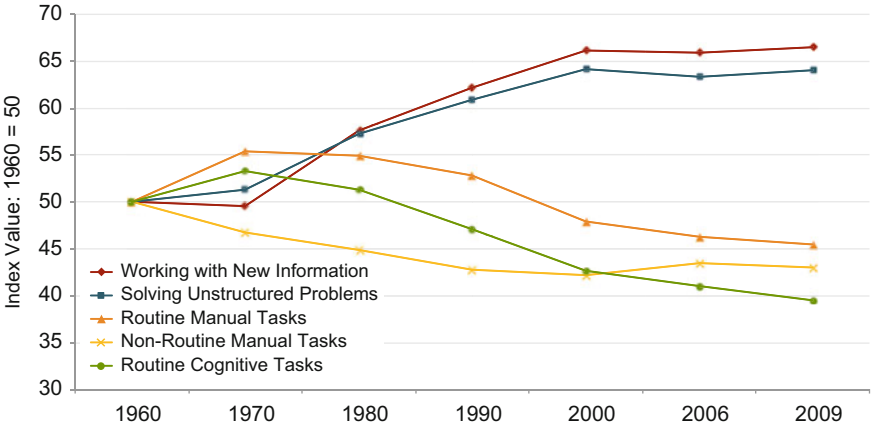



**Fig. 5.1** Index of changing work tasks in the US economy. Source: Levy and Murnane, 2013 ©Third Way (www.thirdway.org)

There are two activities, whose value increased significantly: working with new information and solving unstructured problems (see Fig. 5.1). One of the conclusions was that the lack of data specialists is a missed opportunity for job creation.

More detailed analyses concerning open data were also performed in various countries. For example, in Spain there is a periodical study of the infomediary sector—companies that sell services on top of open data. In 2012, there were around 4000 employees generating €330–550 million annually that can be directly attributed to the reuse of open data (ONTSI, 2012). Those number were increasing in the following years: between €550 and €700 million in 2014 and between €600 and €750 million in 2015 (Vázquez Martínez, 2016). The number of jobs also increased to 5200 in 2015. The majority of employees dedicated to the infomediary sector are highly qualified: 61% hold an advanced degree (master or doctor) and 17% a medium-level degree (bachelor). It is also important that positions are not restricted to computer scientists—people with various backgrounds can contribute: sociology, engineering, mathematics, physics, economics, journalism, and architecture (Vázquez Martínez, 2016).

### *5.2.2 Public Sector Information*

Public sector information (PSI) is an important value contributor from the macroeconomic point of view. According to the literature, it can have also the broader impact. First of all, PSI is increasing democratic participation. It offers the opportunity to perform own analyses, make conscious choices, or even improve the policy-making process. Second, PSI is promoting greater accountability. Public services can be benchmarked and their costs compared. Third, PSI increases social

cohesion by showing information on the provision, distribution, and usage of public services. Fourth, PSI generates environmental benefits. It can contribute to better journey planning by releasing traffic and transport data. Finally, PSI allows identifying previously unknown links between different policy areas.

Public sector information is a broad category and can encompass data from many domains. For research purposes, PSI is usually divided into smaller sectors. For example, Fornefeld et al. (2008) focused on economically relevant data. They distinguished sectors as well as the most important categories of data relevant for re-users:

1. Geographic information, with such categories as: topographic information, cadastral information (including address data), aerial photography.
2. Meteorological information, with such categories as: synoptic observations, radar images, weather predictions.
3. Legal and administrative information, with such categories as: primary and secondary legislation, regulations, official notices, decisions of national and regional courts.

There is no single methodology for estimating the value of public sector information and the value of many parameters is unknown; therefore, estimates can differ significantly. Deloitte (2013) distinguished two kinds of PSI values. The first one—narrow economic value—aggregates the direct value to consumers and PSI holders as well as the value accrued by their supply chain. The second—broader social value—captures all social effects derived from public sector information use. The value of public sector information overall was disaggregated into the following *value components* (Deloitte, 2013) :

- The direct value of PSI for producers and suppliers—the benefits received by producers and suppliers through the sale of information or related value-added services.
- The direct use value of PSI to consumers—the benefits received by consumers (businesses, civil society, or individuals) from directly using and reusing PSI for a variety of purposes.
- The indirect value of PSI from its production and supply—the benefits accumulated up the supply chain to those organizations that interact with suppliers or producers of information.
- The wider societal value from the use and reuse of PSI—benefits to the society of public sector information being exploited, which are not readily captured elsewhere.

### *5.2.3 Benefits by Sectors*

Opening up data provides benefits both for data holders and data users. Data holder usually represents a public sector and data users usually are from the private sector.

Tinholt (2013) divided the key macroeconomic benefits of these sectors in three areas:

- Driving revenue through multiple streams—Public sector benefits from increased tax revenues through increased economic activity. Additional revenue can come from selling high value-added information. Private sector can benefit from new business opportunities.
- Cutting costs and driving efficiency—The government can cut costs through a reduction in transaction costs. Service efficiency is expected to grow through the use of linked data. Private sector can benefit from better decision making based on accurate information, which can also be obtained at a lower cost (no conversion of raw government data necessary).
- Increasing employment and developing future-proof skills—Public sector can benefit from creating jobs and encouraging entrepreneurship. Companies, on the other hand, can gain a skilled workforce.

There can be different economic goals regarding the benefits that can be gained. Consumer surplus is a measure of consumer benefit; it occurs when the consumer is willing to pay more for a given product or service than the current market price or the price paid and is calculated as a difference between these values (see Fig. 5.2). The concept of consumer surplus was developed to measure the social benefits of public goods. Producer surplus is a measure of producer benefit; it is a difference between the market price and the lowest price the producer is willing to accept. A government measures its benefits as revenue divided by expenditure. The above measures are an important tool in the field of welfare economics. The total welfare is calculated as the weighted sum of the above three measures.

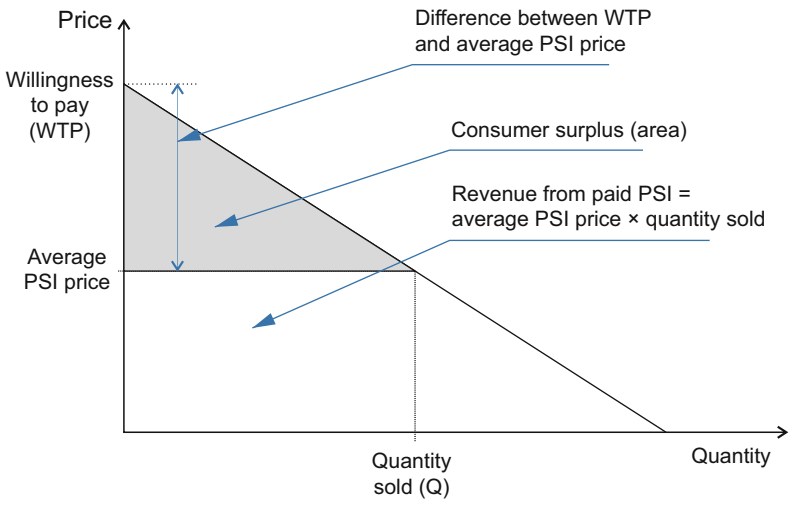



**Fig. 5.2** Consumer surplus for paid PSI. Source: own work based on Deloitte, 2013, p. 189

### 5.2.4 *Data as Infrastructural Resource*

An infrastructure is the basic equipment and structures that are necessary for a country, region, or organization to function properly. Traditionally, the infrastructure is interpreted as large-scale physical facilities provided for public consumption (Frischmann, 2013). The classic examples are transportation systems, communication systems, and basic services and facilities, such as sewage and water systems.

Infrastructure, as an underlying foundation, needs to satisfy three criteria Frischmann (2013):

1. Non-rivalrous criterion—can be consumed in principle an unlimited number of times.
2. Capital good criterion—is used for the production of other goods.
3. General-purpose criterion—can serve different users and usage scenarios.

OECD (2015) agreed that "the economic properties of data suggest that data may be considered as an infrastructure or infrastructural resource." Data has the potential as a driver for value creation and economic growth. The specific properties of data predestining it as an infrastructural resource are: (1) the non-rivalrous nature of data consumption—it can be consumed multiple times by many entities, i.e., availability is not reduced as people consume it; (2) non-excludability—it is impossible to provide data without it being available for many people; (3) the economics of scale and scope in the creation and use of data. Data also depreciates as other capital goods, whose value declines over time, also depending on how it is used (Frischmann, 2013). ICT as a whole seems to be a good candidate for general purpose technology (GPT) in twenty-first century (Buchholtz et al., 2014).

Social welfare can be maximized when a rivalrous good is consumed by the individual who values it the most. Market mechanism is the most efficient for allocating the resources needed to produce such goods. This is not true for non-rivalrous goods. As Frischmann (2013) highlighted, social welfare is maximized when everyone who values the good consumes it, not only the person valuing it the most. By maximizing access to the non-rivalrous, we could in theory maximize social welfare, as every additional unit consumed increases the overall consumer surplus at no additional cost. In fact, such behavior is typical for infrastructures.

Referring to the consumer surplus, users' willingness to pay for the infrastructure, and for data in particular, reflects private demand—the value that they expect to realize. It does not take into account the value that others might capture as a result of their use, i.e., the social value. When a demand manifestation problem occurs, data markets cannot fully serve the social demand for data.

### 5.2.5 *Costs, Investments, and Pricing*

The provision of high-quality data requires significant investment and time for preparation before the data can be shared. OECD (2015) distinguished costs related to: (a) datafication, (b) data collection, (c) data cleaning, and (d) data curation. For effective knowledge sharing, additional resources besides data are necessary, hence additional costs. Data can be accompanied by metadata, complex data models, algorithms for storage and processing, or even secure IT infrastructures.

Looking at significant costs, creators of data are reluctant to share their assets. The overall costs of sharing are perceived as being higher than the expected benefits from sharing. Moreover, data is non-excludable good—once data is published, it can circulate on the Web for free. The free-riding users can then consume the resources without transferring an adequate payment to investors, who in turn are unable to recover their investments.

Another fact discouraging from data opening is that costs and revenues are not 'synchronized.' The effects are delayed in time and should be observed over a longer period. Costs are short-term and easier to predict, e.g., technical development, hosting, legal costs, and administration. Revenues are long-term and harder to predict. They depend on less quantifiable indicators, such as customer behavior, market trends, or product adoption. In a longer term, a broader spectrum of results should be considered (Buchholtz et al., 2014). Benefits include efficient use and reuse of data, innovation, feedback, and transparency; costs include lost revenue and lower competitiveness. Additional concerns, such as national security, privacy, and commercial confidentiality should also be considered in the outcome calculation. Advantages are distributed in time and additionally the impact is indirect.

The temporal discrepancy between costs and revenue was observed earlier during ICT proliferation. Prior to 1990, the macroeconomic impact of ICT was probably negative and a large acceleration in GDP has been observed since 1990 (Jovanovic & Rousseau, 2003). It is not a coincidence that World Wide Web was invented in 1989 and the first web browser was implemented in 1990 by Tim Berners-Lee. The labor productivity growth can be attributed to the proliferation of ICT technologies. According to Buchholtz et al. (2014), the second wave of this revolution can be expected through advancements in big and open data. What is more important, not the direct productivity increase will be the most important for the economy but the indirect effects.

The main driver for the revolution on the macroeconomic level is improving efficiency by cutting the resource waste. Open data can help to save costs through two outcomes. The first is related to the distribution of data, where administrative overhead can be decreased. The second can be achieved by subcontracting or outsourcing application development and service provision (Sieber & Johnson, 2015).

In order to recover at least part of the costs, data suppliers can use various funding and pricing strategies. Pollock (2009) distinguished the following funding schemes for PSI:

- Government funding—opening of data is funded from general government revenues.
- Updater funding—users who make changes to the datasets are charged.
- User funding—users making use of the datasets are charged.

Charging for data is nevertheless a risky business. It is hard to estimate all necessary costs. There are additional and not obvious costs like legal costs related to licensing and software development. Moreover, customers are less willing to pay for digital goods. Another variable here is customer adoption. It is not possible to get stable revenue unless people adopt certain solutions and nobody wants to invest in technology or service that is not stable.

Funding scheme defines who pays for data access. It is still necessary to agree on the price for the access; therefore, Pollock (2009) proposed also charging schemes for PSI:

- Profit-maximizing—prices are set at a level to maximize profit according to a given demand. When there is no competition, it will result in monopoly pricing. PSI holder is funded from its revenues.
- Cost-recovery—prices are set at the average long-run costs, and therefore PSI shall not require direct funding from the government.
- Marginal-cost—prices are set equal to the short-run marginal cost. The cost of supplying data to an additional user tends to zero, so the PSI holder has to be funded by the government.

As an economic conclusion, Pollock (2009) suggested that public sector information in a digital form is best funded out of a combination of 'updater' fees and direct government contributions while users are permitted to use information for free. This is motivated also by the early observation by Shapiro and Varian (1998): competition among sellers of commodity information pushes prices to zero. Such a model, when managed and regulated properly, can offer major societal benefits: broader access to information with a limited funding burden upon government.

Koski (2011) showed that the PSI pricing scheme influences the firms growth, particularly small and medium sized enterprises. Firms in countries that provided geographical information either for free or at maximum marginal costs have grown, on average, 15% more per year compared to firms in countries that priced geographical information according to the cost-recovery principles. It is important to note that after switching to marginal cost pricing, the positive effect was not observed among large companies but only small and medium ones.

This supports the decision by The European Parliament and of the Council, which in the Directive 2003/98/EC stated that "the Member States should encourage public sector bodies to make documents available at charges that do not exceed the marginal costs for reproducing and disseminating the documents."

## 5.3 Direct Value

Data has no intrinsic value, as the value depends on the context of their use (OECD, 2015). From the macroeconomic point of view, it can be understood as the market value, i.e., the total turnover or profits accruing to producers or consumers. Value of data very often is related to the capacity to extract insights from the observed data. Value can be created in a value conversion process, as understood by (Allee, 2008): "value conversion is the act of converting or transforming financial to non-financial value or transforming an intangible input or asset into a financial value or asset." In this section we investigate the macroeconomic value of big, open, and linked data.

### *5.3.1 Value of Information*

Value of information can be defined in different ways depending on the methodology assumed. Classical definition of the value of information refers to the additional revenue that can be expected when using a piece of information $\mathscr{F}$. The value of information associated with $\mathscr{F}$ is defined as "the difference between the maximum expected payoff that a decision-maker can obtain by conditioning on $\mathscr{F}$ and the maximum expected payoff that could be obtained without conditioning on $\mathscr{F}$" (Donaldson-Matasci et al., 2010).

The expected value of perfect information (EVPI) is the maximum amount a decision maker would pay for a perfect information (Hubbard, 2007). EVPI equals the difference between the expected value given perfect information (EV|PI) and the expected value without perfect information (EMV—expected monetary value). In decision theory, we also define the expected value of sample information (EVSI). It is the expected gain that a decision-maker could obtain from having access to a sample of additional observations before making a decision. EVSI is calculated as the difference between EMV with posterior and marginal probabilities and EMV with prior probabilities. EVSI is a relaxation of EVPI—it indicates the value of some limited and incomplete information (Howard, 1966).

Another approach to describe value of information uses a notion of mutual information or a measure of uncertainty reduction, introduced by Shannon (1948). Donaldson-Matasci et al. (2010) showed that fitness value related to EVPI is equal to the reduction in uncertainty about the environment, as described by the mutual information. It can be shown that *information-theoretic* approach is equivalent to *decision-theoretic* approach.

Value of information is highly contextual, i.e., it depends on several additional factors. First of all, the economic value of external data is a function of pre-existing knowledge and depends on the data already possessed, which is caused by redundancies in the external data (Dalessandro et al., 2014; OECD, 2015). Furthermore, the value that can be extracted from data is not only a function of the data, but also a function of the analytical capabilities (OECD, 2015). As Dalessandro

et al. (2014) remarked "data has no intrinsic value, and the estimate of its value is only as good as the abilities of the modeler undertaking this exercise."

Summarizing, there are factors beyond the data themselves that determine their value:

- **Data linkage**—The same datasets can lead to different information depending on their structure, including their links to other datasets. Therefore, the value can also differ.
- **Data analytic capacities**—The value of data depends on the extracted or interpreted meaning. Once more, the same datasets can thus lead to different information depending on the analytical capacity, skills, available techniques, and technologies for data analysis.

Dalessandro et al. (2014) studied economic value of adding incremental data. They supported the use case when external data was bought from third-party data providers. Their model laid the foundations for estimating the fair price of purchased data. Unlike other approaches that as a starting point assumed the seller's perspective, they modeled the situation from the buyer's perspective. In their novel approach, they related the improved model performance to monetary gain. They studied a cost-sensitive classification as an example, i.e., each prediction can have a positive (gain) or negative (cost) value depending on the actual classification. The idea was extended from a fair price for buyers to a 'deserved' price for data providers. The sellers should be rewarded proportionally to the effects their data caused on the buyers side.

Particularly context-sensitive is the valuation of privacy and personal data. OECD (2013b) showed that the monetary valuation of the same data could diverge significantly among market participants. For example, social security numbers in United States can be obtained from specialized data brokers for less than $10. At the same time, various surveys show that individuals are willing to reveal their numbers for $240.

Yet another approach to measure value of information is to recognize its 'tax value.' OECD (2014) discussed issues emerging from the global distribution and interconnectedness of the data ecosystem. The conclusion was that most businesses did not fully took into account the economic value of the data they controlled in their balance sheets. Data was also not appropriately characterized and valued for tax purposes, what also contributed to Base Erosion and Profit Shifting (BEPS) issues.

### *5.3.2 Value of Big Data*

Measuring the value of big data is difficult, because (a) there is huge amount of data and (b) data is used in various situations for various purposes (OECD, 2013b).

Centre for Economics and Business Research (2012) developed a macroeconomic model to calculate the aggregate economy-wide impacts of big data analytics.

Their methodology consisted of three steps that could be used for any similar initiative:

1. Identify and quantify the benefits of big data to business—understanding of the benefits of big data at the enterprise level. There are five key characteristics that collectively signify the extent to which big data has the potential to transform operations of enterprises: data intensity, earnings volatility, product differentiation, supply chain complexity, and IT intensity.
2. Determine current and prospective rates of big data analytics adoption—understanding of the major drivers of the widespread adoption of big data analytics.
3. Calculate the aggregate economy-wide benefits of big data—building the macroeconomic model.

The model was used to measure the impact of big data in the United Kingdom for twelve industry sectors. Three economy-wide benefits were identified. Cost savings and revenue growth were achieved through optimization of operations using one of the six mechanisms: customer intelligence, supply chain management, quality management, risk management, performance management, and fraud detection. The second benefit was product innovation, which meant that enterprises already investing in R&D had bigger chances to receive higher innovation benefits. Third benefit was business creation as the low cost barriers provided incentives for new businesses (start-ups).

Buchholtz et al. (2014) distinguished three *channels* through which big data affects economy:

1. Data-to-Product/Process—this effect can be observed when results of the analysis need to be implemented in the *physical world*. This is a direct way to increase the efficiency of companies. Valuable innovations can be introduced when data analysis reveals potential solutions.
2. Data-to-Information—this impacts the organizations for which *information* is the core source of value. All improvements take place in the *intangible* part of the economy. Changes are transmitted to economy by changes in business models or new business model inventions.
3. Data-to-Management—this provides valuable input to decision making process of the enterprises. Not only insights have to be discovered but also translated into business decisions and then actions. A data-driven decision making is a prerequisite. Data-intensive sectors can observe the highest benefits.

The first key macroeconomic effect associated with big data is increased business process efficiency. It is based on faster data collection (e.g., through sensors) and anomaly detection. The improvements are most visible in better supply chain management. Second key effect is the shift to data-driven decision making. It results in an improved allocation of production factors. The third effect is increased competition and business creation. Proliferation of big data technologies, usually based on open source solutions, lowers barriers to entry. Competition improvement is a result of the opening of public sector data.

**Table 5.1** Additional GDP by sector percent in which data-driven solutions are introduced

| Sector | Share (%) |
|---|---|
| Trade | 23 |
| Manufacturing | 22 |
| Finance and insurance | 13 |
| Public administration | 13 |
| Information and communication | 6 |
| Healthcare and social care | 5 |
| Other | 18 |
| Total | 100 |

Source: based on data in (Buchholtz et al., 2014)

The main business problems addressed by big data are enhancing the customer experience and improving the process efficiency (Heudecker & Kart, 2014). Another area where an increase in the use of big data is observed is the development of information products. Buchholtz et al. (2014) analyzed impact of big data by economy sector. Table 5.1 presents additional GDP by sector. Trade and manufacturing seem to provide the highest value in relation to GDP.

In order to leverage big data for adding value to enterprise Business Intelligence, several components are necessary (Enterprise Management Associates, 2011). First of all, self-service BI is a key to a growing a successful BI community. It is important that tools bridge the gap between IT and the business user. Second component is wide data access, which empowers the business users. It means that data should be federated from multiple sources in an easy to understand way. Third—data blending allows users to augment their analyses with multiple data sources in a data mash-ups manner. It will also be critical to find new business models applicable to the European economy.

### *5.3.3 Value of Open Data*

The EU directive on the reuse of public sector information (EU, 2003) was foreseen to encourage realization of the economic value of public data through its reuse. Governments were empowered to open up the possessed data. It regulated the issues of licensing and charging for information. The motivation for opening was twofold—not only benefits from reuse were targeted but also transparency could bring value to economy. Availability of open data allowed to improve the public services that produced that data. We observe a paradigm shift where hierarchical markets are transformed into an open and networked economy. In this particular setting, "open data is widely presumed to have a positive effect on social, environmental, and economic value" (Jetzek et al., 2014).

Economic impact of open data was widely studied (Farrell, 2012; Jetzek et al., 2013, 2014; Manyika et al., 2013). Similarly to big data, we do not know the value of open data unless it is used in a specific business case. Manyika et al. (2013)

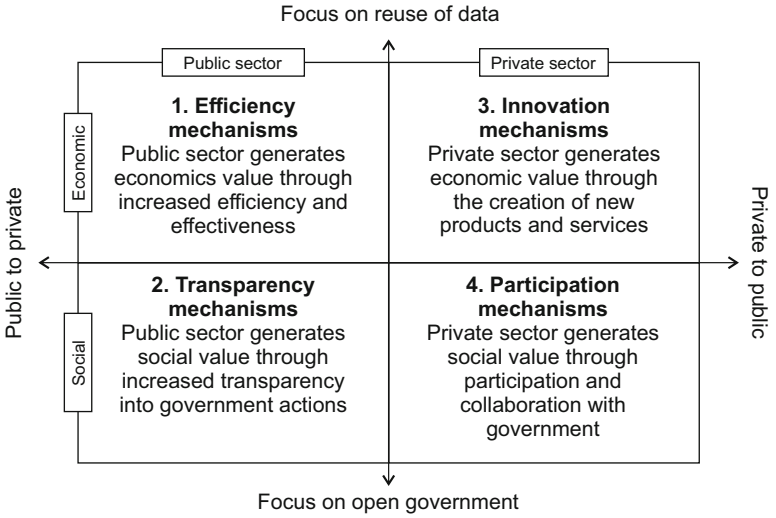



**Fig. 5.3** Four archetypes of OGD value generating mechanisms. Source: based on Jetzek et al., 2013

focused analysis on how the use of open data can create economic value globally in seven domains: education, transportation, consumer products, electric power, oil and gas, health care, and consumer finance. Jetzek et al. (2014) argued that value generation can be empowered by open data sharing and reuse. They proposed a model that described what mechanisms and how contribute to sustainable value. These mechanisms were enabled by a number of contextual factors (see Fig. 5.3).

Often the value from open data is elicited by combining open and proprietary data, in particular big data and its analytics. Open data provides additional *depth* to big data applications and enables entirely new ones. Unfortunately, the majority of open data repositories maintained by governments are too generic and thus of little use for individual citizens. This general data is generated from detailed raw data, which in fact has a much greater value. Usually, there are large volumes of data, but not necessarily open. Some of them are secured by statistical confidentiality, i.e., only authorized state bodies can access detailed data.

Nevertheless, there are some good examples of transforming big data into open data. Farrell (2012) mentioned 'Apps for Democracy,' an open community platform in Washington. The first call for applications attracted almost 50 submissions in 30 days. It generated $2,300,000 for the city at a cost of $50,000.

There is also a belief that value of open data can be generated from diversity. On the Web, we can observe a plethora of smaller sites that provide specialized content and are more respected than big providers. There are also many of them; therefore, various topics can be covered and the intersection between them can be used to compare and verify information. Smaller sites are better connected to their committed user communities. There can also be a number of mobile applications

using the same data and still all of them attract certain people. As Nielsen (2003) stated, diversity is power on the Web.

### 5.3.4 *Value of Linked Data*

Speaking of the value of linked data, we need to refer to Allee (2008), who stated that "the network is the primary economic mechanism for value conversion."

Value of big and open data was highly context-dependent, so is linked data. The value increases when data can be linked to other datasets. Data placed in a larger context can reveal additional insights. Linked data thus creates *super-additive value*, which is greater than the sum of its parts (OECD, 2015). When big data is linked, it offers additional advantages, e.g., provision of a context.

Latif et al. (2009) is one of the first publications to focus on the value of linked data. By introducing roles, they laid foundations for the value chain, which referred to various stages in the linked data life cycle. They emphasized the importance of human-readable data—by consuming this data, the value for the human end-user is generated, which is crucial to the success of any linked data business case. Heath and Bizer (2011) claimed that value stems from reuse. The same is supported by Bechhofer et al. (2013)—publishing does not ensure reusability, thus is not sufficient to create value.

Additional value of linked data is perceived in inferencing (Bechhofer et al., 2013; Dadzie & Rowe, 2011; Hitzler & Janowicz, 2013). Inferences, especially in science, are made possible using RDF Schema and OWL (Bechhofer et al., 2013). Complex queries can be answered not only based on local data but also reaching to external "semantically rich information" (Dadzie & Rowe, 2011). Hitzler and Janowicz (2013) perceived linked data as a support for the fourth paradigm of research, i.e., for pushing forward the frontiers of knowledge by gathering, manipulating, analyzing, and displaying data. The progress is made not by acquiring new data but by mining the existing collections with new analytical techniques.

In order to assess the business value of linked data, it is necessary to explore various business cases and business model patterns for open data and how companies can create additional value from it (see Sect. 7.6). According to Houle (2016), linked data has not changed the world yet. First, consumers look at the quality, which is not satisfactory. Already Even and Shankaranarayanan (2007) called to the need to develop new metrics for data quality assessment that would take various contexts into account. Second, there is no feedback loop for connecting success of the end user to the success of the dataset publisher.

### 5.3.5 *Value of Alternative Data*

Alternative data is a very specific kind of data; therefore, it requires separate methods of valuation. It can be big data, and the bigger is the volume, the higher is the value potential. It is definitely not open. It can be distributed as linked data, thus lowering the cost of integration. The highest volume potential is in its novelty; therefore, the data has to be new for the buyer. The buyer then decides how to use data and what benefits it can bring. As such, it is the buyer's market. Alternative data can only be valuated on a case by case basis. There is, however, one sector of the economy particularly interested in this kind of data.

Some hedge funds from Wall Street[1] are particularly interested in raw data. Many of them use mathematical models and algorithms[2] to evaluate long-term investments. Recently, they also started embracing the capabilities of artificial intelligence, particularly deep learning. Their interest in alternative data is growing. Hedge funds do not share with data providers what exactly they are going to do with data; therefore, it is hard to attribute value to it (Turck, 2017). Only the funds know the real value of data. There are also some marketplaces emerging that help to trade regular datasets and alternative data. One of such examples is Quandl.

What makes a difference for quant funds is data, particularly non-obvious, hard to get alternative data. They may be interested in data that has several key characteristics: level of detail, history, breadth, and rarity. User-level data, even anonymized, is much more interesting than aggregated values. Longer history may be more interesting, but the value of data decreases over time—there will be more companies offering the same data.

An interesting example of a company offering (big) data for a maritime domain is Windward.[3] They have built Windward Mind, which organizes and analyzes the world's maritime data, bringing visibility to the maritime domain. It provides knowledge concerning commodity trade flow: from global patterns to a single vessel. The data has a high volume, contributing to maritime big data, and often is intentionally manipulated. There is also a lot of open data sources from the maritime domain, which allow to enrich datasets in less demanding scenarios. The selection process can rely on the quality of data sources (Stróżyna et al., 2018). There is a good reason for interest in maritime trade—substantial macroeconomic implications for commodity traders, hedge funds, private equity firms, banks, and shipping companies. They make crucial financial decisions about worldwide shipping and flow of goods. The predictions are only as good as the data they rely on.

[1] They are called "quant funds" where *quant* stays for *quantity*.

[2] It is often referred to as 'quantimental investing,' where quantimental is a combination of 'quantitative' and 'fundamental.'

[3] http://www.windward.ai/.

## 5.4 Multiplier Effects

We have so far studied the direct effect, but it is not the only source of value. In this section we discuss various incarnations of a macroeconomic concept called the 'multiplier effect,' which explains how a change in input leads to a larger change in output.

### *5.4.1 Returns to Scale and Returns to Scope*

Terms 'return to scale' and 'return to scope' describe what happens when the scale of production increases in the long run. They characterize the disproportionate change of output when the level of input is changed.

*Returns to scale* concerns the production function and is realized when, for example, doubling of the amount of all factors of production results in more than double the output. Economies of scale are cost advantages that organizations obtain thanks to the size of their operations, reflected in an increased output level with unit costs.

*Returns to scope* is conceptually similar to returns to scale; however, an increase in output is achieved through the diversity of inputs. Economies of scope are advantages formed by a variety, not volume, and are the foundation of diversification plans of corporations (OECD, 2015).

**Big Data Contributes to Increasing Returns to Scale** The accumulation of data can lead to improvements in data-driven services. Better services attract more users, and as a result even more data can be collected, leading to even better services (positive feedback). For example, the accuracy of traffic information in Google Maps is highly depending on the number of users using the service. This feedback reinforces the market position of the service provider and can lead to higher market concentration or even dominance. When the service attracts more users and additionally the network effect can be observed, then increasing returns to scale can occur.

**Linked Data Contributes to Increasing Returns to Scope** If data linkage is possible, then diversification of services leads to even better insights. Linking data allows to contextualize data and it is therefore a source for insights that are greater than the sum of isolated parts. This super-additive value leads to increasing returns to scope. The super-additive nature of linked data has also its drawbacks—it can undermine confidentiality and privacy protection.

The economic properties of data described above have significant implications for economy. As Shapiro and Varian (1998) noticed, "positive feedback makes the strong get stronger and the weak get weaker, leading to extreme outcomes." Diversification of services accompanied by data linking can lead to reinforcement of the feedback. A multi-sided market characterized by returns to scale and returns

to scope, reinforced with network effects can lead to a 'winner takes all' outcome. Success on such a market can easily lead to a monopoly.

An example that comes to mind here is the case of Google, studied by Newman (2014). Google has expanded its control of user data both by increasing the depth of collected data and by moving into new product sectors to collect additional user data. Data collected from a plethora of services—what users write in Gmail, what they watch on YouTube, where they go in Google Maps, what applications they use in Google Play, etc.—allowed Google to become a dominating player in the search market. According to Net Market Share[4] (last data for October 2020), Google search engine share for desktop was 69.3% and 92.9% for mobile.

Another big dominating company is Facebook, with 2.2 billion monthly active users[5] (November 2021). Although it is a dominating social network, it is not a monopoly. People diversify their social media use. The market share of Facebook in the US dropped from maximum 70% in 2011 to 37% by the end of August 2021.

The monopoly resulting from returns to scale and scope has some features of a natural monopoly. It is not a result of consolidation through merges and acquisitions. Natural monopolies can arise in industries that require unique materials, technology, or other similar factors to operate. The unique material in the cases considered above is data.

### *5.4.2 Network Effects and Two-Sided Markets*

There is something special about linked data—the links themselves. They contribute to so-called network effects, exemplified by Metcalfe's law. Network effects, also referred to as demand-side economies of scale, are present if the value of a product or service depends on the number of other consumers using it (Shapiro & Varian, 1998). Telephony is the classical example. Thus, any user joining a network can bring more value than he or she individually perceives. Two main sources of value of linked data are derived from the network effects: massive collaboration between users and creating bidirectional links between resources (Katz & Shapiro, 1994). This explains why achieving critical mass is crucial for the profitability of linked data applications.

Vafopoulos (2011) distinguished network externalities from effects. Network externalities are understood as indirect effects of consumption or production activity stemming from network effects. They are determined by four factors: expectations of consumers, coordination of consumers, switching costs, and compatibility of the network good (Katz & Shapiro, 1994). *Network goods* are defined as goods that derive their value from the network. Linked data can definitely be classified as a network good.

[4] https://www.netmarketshare.com/search-engine-market-share.aspx.

[5] https://www.dreamgrow.com/top-10-social-networking-sites-market-share-of-visits/.

The major value of the Web lies then in the ability to associate resources. Linked data can be perceived as an enabler catalyzing value of other assets. The more we link, the more value it brings. In other words, each new linked dataset adds value to those already published (Archer et al., 2013). Buchholtz et al. (2014) called the ability to recombine datasets and experiment with them as a 'snowball effect.'

Linked data is also changing the data flow—from a simple data flow model: 'owner to infomediary to end-user' to a complex one with dependencies explained as edges in a network where also the direction of relations can vary significantly. Tapscott et al. (2000) introduced five types of value networks: agora, aggregation, value chain, alliance, and distributive networks. As Pellegrini et al. (2013) argued, the "primary value proposition of linked data is rooted in its modularity and connectivity to generate network effects at the data level."

We can distinguish two kinds of network effects: direct and indirect. *Direct network effects* specify that an increase in usage leads to a direct increase in value for other users. For example, a social network implies direct contact among users. This effect is called a same-side network effect. *Indirect network effects* observe another phenomenon: an increase in the usage of one product causes increase in the value of a complementary product, which in turn can increase the value of the original product. This effect is called a cross-side network effect. Examples of complementary goods include software and operating systems. The value of the operating system depends on the number of available applications (cf. failure of Windows Mobile). Economides and Katsamakas (2006) explained why Windows and Linux might compete for software developers rather than for users. Most two-sided markets (or platform-mediated markets) are characterized by indirect network effects. The following example can be used to illustrate the difference between these effects: a gamer who participates in an online platform can benefit from the participation of other gamers (direct network effect) or game developers (indirect network effect).

As a consequence of the discussion above, we can distinguish two sources of economic value of a product displaying network effects: (a) *inherent value*: value derived by an individual from the own use of the product; (b) *network value*: value derived from other people's use of the product.

Two-sided or multi-sided markets are defined as "markets in which one or several platforms enable interactions between end users and try to get two or multiple sides 'on-board' by appropriately charging each side" (Rochet & Tirole, 2006). Such platforms enable not only interaction between various distinct groups of customers but also the exchange of externalities between them, i.e., decisions of one group can impact the results of another group. As a consequence, the prices charged to the members of each group will reflect these externalities (OECD, 2014). For example, more users of a certain portal can generate more clicks on links sponsored by advertisers. This is a positive externality for another side—advertisers. The bigger the user base, the more advertisers have to pay.

A concept very often analyzed in the context of two-sided platforms is a 'free good.' It is a special case of exchange of externalities. The existence of a free good means that there is a companion good, which considered together with the free good

allows to maximize profit. The companion product may be a complement of the free good, a premium version of the free good, or the product on the other side of a two-sided market (Evans, 2011). The existence of a free good is only economically justified if the price of a companion good is higher than the marginal cost of the free good; and usually it is significantly higher. Free goods and services are increasingly common on web-based multi-sided platforms.

Examples of such platforms include Amazon and eBay, which provide marketplaces for sellers and buyers. Google Play (formerly Android Market), Apple's App Store, and Microsoft store link consumers and application developers. Apple's iTunes and Steam link consumers and digital content providers (e.g., video, music). Externalities not only concern the possibility to reach a large user base but also the feedback information, such as a product review. The review posted on a platform may attract many comments that contribute substantively to the review's value and credibility (Marshall & Shipman, 2017).

Google and Facebook are specific service platforms that have developed data- and analytics-enabled multi-sided markets. Here, distinct user groups generate benefits—externalities or spillovers—for the other sides (OECD, 2015). Data is a non-rivalrous capital used by platform operators. It can be used on one side of the market to provide a personalized service. It can be reused on the other side as input for additional services. One of such services, the most profitable one, is marketing. Both Google and Facebook reuse consumer data to provide marketing services to third-parties.

### *5.4.3 Disruptive Innovation*

A disruptive innovation is defined in opposition to a sustaining innovation. The latter are those innovations that mainstream customers demand. However, companies may be deceived when listening to their *current* customers.

Disruption describes a process which allows a smaller company with fewer resources (entrant) to successfully challenge established businesses (incumbent firms) (Christensen et al., 2015). The offer of incumbent firms' often exceeds the performance requirements of the less-demanding customers, and therefore is also too expensive for them. The disruptive innovation has lower performance compared to the expectations of mainstream customers. It does not satisfy current customers, but there are other performance attributes valuable for less-demanding customers, who slowly adopt the new offer. Disruption occurs when mainstream customers also start adopting the entrants' offer.

There are two types of markets neglected by the incumbent firm, as a potential for disruptive innovation: low-end niche (providing a 'good enough' product for less demanding customers) and a new market niche (turning non-consumers into consumers). Disruptive innovation can be characterized by several features (Hölzle, 2017): (a) exponential growth, (b) optimal exploitation of network effects—"the winner takes it all," (c) zero-marginal costs, (d) minimal transaction costs, and

(e) almost no own resources. Disrupters often build business models that are very different from those of incumbent firms. They create new markets—they find a way to turn potential consumers into consumers (Christensen, 1997).

Disruptive innovation phenomenon can be used to understand big data and open data developments observed currently. Open data is typically perceived as a mean for assuring transparency and accountability of a government. Few examples concern the commercialization of PSI. Lakomaa and Kallberg (2013) claimed that open data should be seen as an enabler of innovation outside these traditional sectors. This would also indicate that the previously calculated societal values of open data might be underestimated. Openness as a strategy can shift the commercial value of data to intelligent services based on the data (Farrell, 2012; van Veenstra & van den Broek, 2013). It is then possible to combine technology with open-data initiatives to reach innovators and entrepreneurs across the government. The practices can include organization of hackathons or datapaloozas by government.

The amount of data involved may differ between sectors—some are more data-intensive than others. Data intensity can be measured as the average amount of data per organization. According to Manyika et al. (2011), data intensity is highest in financial services, communication and media, utilities, government, and discrete manufacturing. In these sectors, a typical organization stored on average at least 1 petabyte of data as of 2009. Data intensity can also be measured with the number of data scientists per 1000 employees; the higher number of analytics professionals, the more data-intensive is the sector (OECD, 2013c).

Tambe et al. (2012) studied impact of external information on innovation. They verified the hypothesis that external focus increases returns from IT, especially when combined with decentralized decision making. External focus was understood as the ability of a company to detect changes in its external operating environment and respond to it accordingly. Furthermore, external focus, IT, and decentralization are associated with improved product innovation capabilities.Their results can explain a counter-intuitive observation: companies that "operate in information-rich environments, such as high-technology clusters or areas with high worker mobility have experienced especially high returns to IT investment" (Tambe et al., 2012). This is a strong argument for *cooperation between enterprises*. Capturing external information through networks of customers, suppliers, partners, and new employees can increase the effectiveness of IT investments. In information-rich environments, companies should focus on making accurate and up-to-date information available to decision-makers. Open and linked data can be among the factors as well.

General conclusion from this section is that open and link data is *disruptive*. They address low-end customers and give access to new markets. The quality is not yet satisfactory, but it is steadily improving. It has chances to be accepted in the future as a mainstream.

## 5.5 Summary

There is not yet macroeconomic evidence whether big and linked data will play an important role in the future economy. These concepts just start to emerge and their uptake is just at the beginning. They are compared with the proliferation of other technologies and ideas in the past (steam engine, electricity). The key to understanding the knowledge economy lies in not only perceiving data as assets, but in understanding how these assets can be applied in various configurations in value conversion networks.

Chui et al. (2014) mentioned three predominant value drivers of open data, which can generate more than $3 trillion annually for the global economy:

- Decision making: open data provides a fact base to make more informed and objective decisions.
- New offerings: open data enables organizations to better understand their customers through analysis of context, which should lead to the design of new products and services.
- Accountability: open data reveals details of decisions made by public bodies as well as their consequences and costs. Citizens can then act on it to adjust own behavior.

Open data already has a strong influence on the economy and this impact will become even stronger. Within open data, it is important to identify how the value can be generated. Linked data is even more valuable through enriched contextual information. There is a growing body of evidence that quantifies the utility of open data and demonstrates its impact in many economies. As introduced already by Shapiro and Varian (1998), network effects lead to demand-side economies of scale and positive feedback. Now economists use terms, such as network effects, allocative efficiency, and information asymmetry to explain why open data creates a return on investment.

Throughout the process of data opening, the main concern shifted from questions on how to organize data to how to facilitate data reuse. Publishing of a dataset does not automatically lead to its reuse. The focus also shifted from internal organizations to external users of data.

The lack of intrinsic value of data has consequences for pricing. The value depends on the context of data use. There are also other features that can influence the price, collectively characterized as data quality. More relevant data is more valuable, but the relevance differs between consumers. The value of data can also depreciate over time, so consumers utilizing specific data quicker than others obtain a temporal premium, as studied by Hackathorn (see Sect. 4.2.2). The problem of pricing is deepened by the fact that "information is costly to produce but cheap to reproduce" (Shapiro & Varian, 1998). That is why consumers expect low prices of information or to access information for free. This also explains why free riding is pervasive in society.

# References


ACIL. (2008). *The value of spatial information*. Cooperative Research Centre for Spatial Information. (page 115)

Allee, V. (2008). Value network analysis and value conversion of tangible and intangible assets. *Journal of Intellectual Capital, 9*(1), 5–24. ISSN: 1469-1930. https://doi.org/10.1108/14691930810845777 (pages 122, 127)

Archer, P., Dekkers, M., Goedertier, S., & Loutas, N. (2013). *Study on business models for linked open government data*. European Commission. (page 131)

Bechhofer, S., Buchan, I., De Roure, D., Missier, P, Ainsworth, J., Bhagat, J., Couch, P., Cruickshank, D., Delderfield, M., Dunlop, I., Gamble, M., Michaelides, D., Owen, S., Newman, D., Sufi, S., & Goble, C. (2013). Why linked data is not enough for scientists. *Future Generation Computer Systems, 29*(2), 599–611. ISSN: 0167-739X. https://doi.org/10.1016/j.future.2011.08.004 (page 127)

Buchholtz, S., Bukowski, M., & Śniegocki, A. (2014). *Big and open data in Europe. A growth engine or a missed opportunity?* (p. 114) Warsaw: demosEUROPA. ISBN: 978-83-925542-1-9. (pages 115, 119, 120, 124, 125, 131)

Centre for Economics and Business Research. (2012). *Data equity: Unlocking the value of big data*. London. (page 123)

Christensen, C. M. (1997). *The innovator's dilemma*. Cambridge: Harvard Business School Press. (page 133)

Christensen, C. M., Raynor, M. E., & McDonald, R. (2015). What is disruptive innovation? *Harvard Business Review*. (page 132)

Chui, M., Farrell, D., & Jackson, K. (2014). How government can promote open data and help unleash over 3 million$ in economic value. In *Innovation in local government. open data and information technology* (pp. 4–23). McKinsey & Company. (page 134)

Dadzie, A. S., & Rowe, M. (2011). Approaches to visualising linked data: a survey—www.semantic-web-journal.net. *Semantic Web, 2*(2), 89–124. https://doi.org/10.3233/SW-2011-0037 (page 127)

Dalessandro, B., Perlich, C., & Raeder, T. (2014). Bigger is better, but at what cost? Estimating the economic value of incremental data assets. *Big Data, 2*(2), 87–96. ISSN: 2167-6461. https://doi.org/10.1089/big.2014.0010 (pages 122, 123)

Data-Driven Development (2015). *Data-driven development. pathways for progress*. http://reports.weforum.org/data-driven-development/ (page 115)

Daugherty, P., Banerjee, P., & Biltz, M. J. (2015). *Digital business era: Stretch your boundaries*. Technology Vision. Accenture. (page 113)

Deloitte. (2013). *Market assessment of public sector information*. BIS/13/743. London: Department for Business Innovation & Skills. (pages 115, 117, 118)

Deloitte. (2016). *The value of DDI (data driven innovation)*. (pages 113, 115)

Digital Britain. (2009). *Cm 7650* (p. 238). London: Department for Business Innovation & Skills. ISBN: 978-01-017650-2-2. (page 113)

Donaldson-Matasci, M. C., Bergstrom, C. T., & Lachmann, M. (2010). The fitness value of information. *Oikos, 119*(2), 219–230. https://doi.org/10.1111/j.1600-0706.2009.17781.x (page 122)

Economides, N., & Katsamakas, E. (2006). Two-sided competition of proprietary vs. open source technology platforms and the implications for the software industry. *Management Science, 52*(7), 1057–1071. issn: 0025-1909. https://doi.org/10.1287/mnsc.1060.0549 (page 131)

Enabling the Data Revolution. (2015). Conference Report. Open Data for Development. (pages 114, 115)

Enterprise Management Associates (2011). *Getting big value from big data... Fast*. (page 125)

EU. (2003). 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, 46*(L 345), 90–96. (page 125)

European Data Portal. (2015). *Creating value through open data: study on the impact of re-use of public data resources* (p. 112). Luxembourg: Publications Office of the European Union. ISBN: 978-92-79-52791-3. https://doi.org/10.2759/328101 (page 115)

Evans, D. S. (2011). The antitrust economics of free. *Competition Policy International, 7*(1), 70–89. ISSN: 15540189. https://doi.org/10.2139/ssrn.1813193 (page 132)

Even, A., & Shankaranarayanan, G. (2007). Utility-driven assessment of data quality. *ACM SIGMIS Database, 38*(2), 75. ISSN: 0095-0033. https://doi.org/10.1145/1240616.1240623 (page 127)

Farrell, D. (Ed.). (2012). *Government designed for new times. A global conversation* (pp. 1–116). McKinsey&Company. (pages 125, 126, 133)

Fornefeld, M., Boele-Keimer, G., Recher, S., & Fanning, M. (2008). *Assessment of the re-use of public sector information (PSI) in the geographical information, meteorological information and legal information sectors final report* MICUS Management Consulting GmbH. (page 117)

Frischmann, B. M. (2013). *Infrastructure: The social value of shared resources* (p. 436). Oxford University Press, ISBN: 978-0199975501. (page 119)

Gibson, W. (2010). Data, data everywhere. *The Economist. Special report: Managing information, 394*(8671), 3–5. (pages 113, 115)

Hölzle, K. (2017). *Transformierte Unternehmen brauchen neue Geschäftsmodelle*. [Transformed enterprises need new business models]. MOOC Mastering Digital Transformation. Hasso Platner Institut. (page 132)

Heath, T., & Bizer, C. (2011). *Linked data: Evolving the Web into a global data space* (1st ed. Vol. 1.1, pp. 1–136). Morgan & Claypool. (page 127)

Heudecker, N., & Kart, L. (2014). *Survey analysis: Big data investment grows but deployments remain scarce in 2014*. Gartner. (page 125)

Hitzler, P., & Janowicz, K. (2013). Linked data, big data, and the 4th paradigm. *Semantic Web Journal, 4*(3), 233–235. (page 127)

Houle, P. (2016). *Data lakes, data ponds, and data droplets*. http://ontology2.com/the-book/data-lakes-ponds-and-droplets.html (visited on 2017-09-09). (page 127)

Howard, R. (1966). Information value theory. *IEEE Transactions on Systems Science and Cybernetics, 2*(1), 22–26. ISSN: 0536-1567. https://doi.org/10.1109/TSSC.1966.300074 (page 122)

Hubbard, D. (2007). *How to measure anything: finding the value of intangibles in business*. John Wiley & Sons. (page 122)

Jetzek, T., Avital, M., & Bjorn-Andersen, N. (2013). The generative mechanisms of open government data. In *Proc. of the 21st European Conference on Information Systems*. Utrecht: AIS. (pages 125, 126)

Jetzek, T., Avital, M., & Bjørn-Andersen, N. (2014). Generating sustainable value from open data in a sharing society. In B. Bergvall-Kåreborn & P. A. Nielsen (Eds.), *Creating Value for All Through IT: IFIP WG 8.6 International Conference on Transfer and Diffusion of IT, TDIT 2014, Aalborg, Denmark, June 2–4, 2014. Proceedings* (pp. 62–82). Springer Berlin Heidelberg. ISBN: 978-3-662-43459-8. https://doi.org/10.1007/978-3-662-43459-8_5 (pages 125, 126)

Jovanovic, B., & Rousseau, P. (2003). Two technological revolutions. *Journal of the European Economic Association, 1*(2–3), 419–428. (page 120)

Katz, M. L., & Shapiro, C. (1994). Systems competition and network effects. *Journal of Economic Perspectives, 8*(2), 93–115. ISSN: 0895-3309. https://doi.org/10.1257/jep8.2.93 (page 130)

Koski, H. (2011). Does marginal cost pricing of public sector information spur firm growth? *Keskusteluaiheita Discussion Papers* 1260. (page 121)

Lakomaa, E., & Kallberg, J. (2013). Open data as a foundation for innovation: The enabling effect of free public sector information for entrepreneurs. *IEEE Access, 1*, 558–563. ISSN: 2169–3536. https://doi.org/10.1109/ACCESS.2013.2279164 (page 133)

Latif, A., Saeed, A. U., Hoefler, P., Stocker, A., & Wagner, C. (2009). The linked data value chain: A lightweight model for business engineers. In *Proceedings of ISE- MANTICS09 International Conference on Semantic Systems*, Graz (pp. 568–575). ISBN: 9783851250602. (page 127)

Levy, F., & Murnane, R. (2013). *Dancing with robots: Human skills for computerized work*. Washington, DC: Third Way, NEXT. (page 116)

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute. (page 133)

Manyika, J., Chui, M., Groves, P., Farrell, D., Van Kuiken, S., & Doshi, E. A. (2013). *Open data: Unlocking innovation and performance with liquid information*. McKinsey Global Institute. (pages 114, 125)

Marshall, C. C., & Shipman, F. M. (2017). Who owns the social web? *Communications of the ACM, 60*(5), 52–61. https://doi.org/10.1145/2996181 (page 132)

Newman, N. (2014). Search, antitrust, and the economics of the control of user data. *Yale Journal on Regulation, 31*(2), 401–454. (page 130)

Nielsen, J. (2003). *Diversity is power for specialized sites*. Nielsen Norman Group. https://www.nngroup.com/articles/diversity-is-powerfor-specialized-sites/ (visited on 2017-03-22). (page 127)

OECD. (2013a). Exploring data-driven innovation as a new source of growth: Mapping the policy issues raised by "Big Data. *OECD Digital Economy Papers, 222*, 1–44. ISSN: 2071–6826. https://doi.org/10.1787/5k47zw3fcp43-en (page 114)

OECD. (2013b). Exploring the economics of personal data: A survey of methodologies for measuring monetary value. *OECD Digital Economy Papers, 220*, 40. https://doi.org/10.1787/5k486qtxldmq-en (page 123)

OECD. (2013c). *Supporting investment in knowledge capital, growth and innovation* (p. 362). OECD Publishing. ISBN: 9789264193093. https://doi.org/10.1787/9789264193307-en (page 133)

OECD. (2014). *Addressing the tax challenges of the digital economy* (pp. 1–202). OECD Publishing. ISBN: 978-92-642187-8-9. https://doi.org/10.1787/9789264218789-en (pages 123, 131)

OECD. (2015). *Data-driven innovation. Big data for growth and well-being* (pp. 1–456). Paris. ISBN: 9789264229358. https://doi.org/10.1787/9789264229358-en (pages 115, 119, 120, 122, 127, 129, 132)

ONTSI. (2012). *Characterization study of the infomediary sector*. National Observatory of Telecommunications and Information Society. (page 116)

Pellegrini, T., Dirschl, C., & Eck, K. (2013). Linked data business cube—modelling semantic web business models. In *Proc. of Share-PSI 2.0 Krems Workshop: A Self Sustaining Business Model for Open Data*. (page 131)

Pollock, R. (2009). The economics of public sector information. (pages 114, 120, 121)

Rochet, J. C., & Tirole, J. (2006). Two sided markets: A progress report. *RAND Journal of Economics, 37*(3), 645–667. (page 131)

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x (page 122)

Shapiro, C., & Varian, H. R. (1998). *The information economy* (p. 368). Harvard Business Review Press. ISBN: 978-0875848631. (pages 121, 129, 130, 134)

Sieber, R. E., & Johnson, P. A. (2015). Civic open data at a crossroads: Dominant models and current challenges. *Government Information Quarterly, 32*(3), 308–315. ISSN: 0740-624X. https://doi.org/10.1016/j.giq.2015.05.003 (page 120)

Solow, R. M. (1987). We'd better watch out. *The New York Times Book Review* (p. 36). (page 113)

Stróżyna, M., Eiden, G., Abramowicz, W., Filipiak, D., Małyszko, J., & Węcel, K. (2018). A framework for the quality-based selection and retrieval of open data—a use case from the maritime domain. *Electronic Markets, 28*(2), 219–233. https://doi.org/10.1007/s12525-017-0277-y (page 128)

Tambe, P., Hitt, L. M., & Brynjolfsson, E. (2012). The extroverted firm: How external information practices affect innovation and productivity. *Management Science, 58*(5), 843–859. https://doi.org/10.1287/mnsc.1110.1446 (page 133)

Tapscott, D., Ticoll, D., & Lowy, A. (2000). Digital capital: Harnessing the power of business webs (hardcover). *Harvard Business School Press Books, 1*. ISSN: 0360-8581. https://doi.org/10.1145/336228.336231 (page 131)

Tennison, J., & Hardinges, J. (2015). *The economic impact of open data: What do we already know?* https://medium.com/@ODIHQ/the-economicimpact-of-open-data-what-do-we-already-know-1a119c1958a0 (visited on 2017-10-08). (page 114)

Tinholt, D. (2013). *The open data economy. Unlocking economic value by opening government and public data*. Capgemini Consulting. (pages 115, 118)

Turck, M. (2017). *The new gold rush? Wall street wants your data*. http://mattturck.com/the-new-gold-rush-wall-street-wants-yourdata/ (visited on 2017-09-09). (page 128)

Uchitelle, L. (2000). Productivity finally shows the impact of computers. *New York Times*. (page 113)

Vázquez Martínez, R. (2016). *Chatacterization of the Spanish infomediary sector*. National Observatory of Telecommunications and the Information Society (ONTSI). (page 116)

Vafopoulos, M. (2011). A framework for linked data business models. In *2011 15th Panhellenic Conference on Informatics* (pp. 95–99). IEEE. ISBN: 978-1-61284-962-1. https://doi.org/10.1109/PCI.2011.74 (page 130)

van Veenstra, A. F., & van den Broek, T. A. (2013). Opening moves-drivers, enablers and barriers of open data in a semi-public organization. In *Electronic Government: 12th IFIP WG 8.5 International Conference, EGOV 2013, Koblenz, Germany, September 16–19, 2013. Proceedings*. (Vol. 8074, pp. 50–61) LNCS. ISBN: 9783642403576. https://doi.org/10.1007/978-3-642-40358-3-5 (page 133)

Yang, T. M., Lo, J., & Shiang, J. (2015). To open or not to open? Determinants of open government data. *Journal of Information Science, 41*(5), 596–612. ISSN: 0165-5515. https://doi.org/10.1177/0165551515586715 (page 113)

# Chapter 6
# Microeconomic Aspects of Data Value


## 6.1 Introduction

This chapter continues the considerations initiated in the previous chapter but focuses on a smaller scale. It should help to answer the question what value brings data to individual businesses. The fact that the value of data is context-dependent necessitates studying of specific cases. Here we consider the value of personal data and innovation as the value.

We have a closer look at stakeholders within the open data ecosystem and study it in terms of both supply and demand. The supply side is covered mostly by governments, while the demand side is the domain of enterprises, including infomediaries and actors from civil society. After studying the mutual benefits of these two sides, we move on to discuss data ownership.

The goal of the chapter is to apply induction for the analysis of value of data based on a microeconomic point of view. It will also define data ecosystem, stakeholders, and mutual benefits. The chapter also encompasses the issues of data ownership and personal data. Finally, we explore the critical factors that impact the way organizations create value from data sharing.

## 6.2 Stakeholders

Open data is primarily driven by governments, not by users (Zuiderwijk et al., 2015). Government is representing a shared interest of other stakeholders. Its leading role is emphasized in various models. The concepts 'government,' 'open,' and 'data' can be combined in various configurations. Key perspectives are bureaucratic (associated with the idea of government data), political (associated with the idea of open government), technological (associated with the idea of open data), and economic (Gonzalez-Zapata & Heeks, 2015). The economic perspective emerged from the



K. Węcel, *Big, Open and Linked Data*, Business Information Systems,
https://doi.org/10.1007/978-3-031-07147-8_6

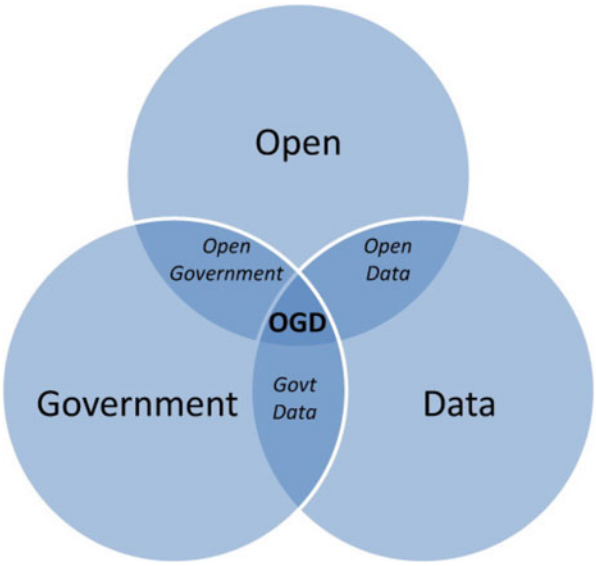



**Fig. 6.1** Foundations of open government data. Source: (Gonzalez-Zapata & Heeks, 2015)

idea of open government data, being the intersection of the above three areas (see Fig. 6.1).

The four derived perspectives can be further characterized by drivers and benefits as well as by the main stakeholders. Details are presented in Table 6.1. It is interesting to note that operational roles dominate in bureaucratic and technological perspectives, while user roles are intrinsic to political and economic perspectives.

The ecosystem delineated by the above perspectives can only be built in collaboration (Zuiderwijk et al., 2014). There is a special role foreseen for policy-makers, who need to use social strategies to encourage open data use and to integrate the use into daily activities. They should show the benefits of open data use and create awareness of users by using open data themselves. Most importantly, they should decrease the effort necessary to use open data technologies (Zuiderwijk et al., 2015). In the following sections, we describe the roles of governments and users.

### *6.2.1 Open Data Ecosystem*

Poikola et al. (2010) defined an open data ecosystem as "a multilevel and multi-dimensional entity where raw material, as far as distribution and developing are concerned, is the target of cooperation." It has to be distinguished from open data infrastructure, which includes all organizations and systems operating with open data, i.e., the whole operational environment. The phrase ecosystem means that we

**Table 6.1** Derived perspectives on open government data

| Perspective | Nature of OGD | Drivers and benefits of OGD | Main OGD actors |
|---|---|---|---|
| Bureaucratic | A policy of data regulations, strategies and processes within government | Improvements in public services through greater efficiency and effectiveness of data management | Public servants; ICT staff [operational role] |
| Technological | A technological innovation within government data systems | Improved government data infrastructure | ICT staff [design and operational roles] |
| Political | A right of free access to public sector data | Better governance through increased transparency, accountability, participation, and empowerment | Citizens [user/beneficiary role] |
| Economic | A mechanism to generate data-based economic value | Economic value through new products, services, revenue, profits, and jobs | Private sector firms, entrepreneurs [user/beneficiary role] |

Source: (Gonzalez-Zapata & Heeks, 2015)

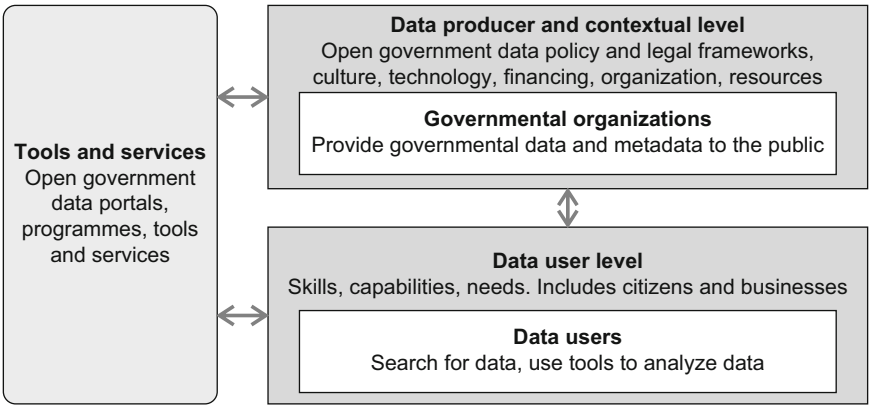



**Fig. 6.2** Elements of an open government data ecosystem. Source: own work based on (Zuiderwijk et al., 2014)

refer not only to technology but also to the living and dynamically changing network of interactions.

Essential challenges that need to be addressed in open data ecosystems relate to policy, legal frameworks, culture as well as technology, financing, and organization (Ubaldi, 2013). One of the important aspects of open data ecosystems is licensing. Creation of an ecosystem without regulations in this area is practically impossible—nobody would know if and how data can be used (Zuiderwijk et al., 2014). Figure 6.2 presents elements of the open government data ecosystem, which encompasses data producers and data users at various levels, along with tools and services.

An open data ecosystem is characterized by interdependencies. For example, open data users depend on the data and metadata provided by the open data publishers, and open data publishers depend on the feedback on their data. Thus, open data publishers and open data users are involved in the network of interactions.

Zuiderwijk et al. (2014) identified several processes that open data ecosystem should support. The first group encompasses: (1) releasing and publishing open data on the Internet, (2) searching, evaluating, and viewing data and their related licenses, (3) cleansing, enriching, linking, analyzing, and visualizing data, (4) discussing, interpreting, and providing feedback to the data provider or other stakeholders. In order to integrate the elements of an ecosystem, some additional elements should be present, forming a second group: (5) user pathways showing directions for how open data can be used, (6) a quality management system, (7) different types of metadata for connecting the elements and ensuring interoperability.

The main role of a government is perceived in policy-making. The government is a facilitator for other players, such as citizens and consumers, businesses, media, and non-governmental organizations. Chui et al. (2014) characterized the stakeholders as follows:

- Citizens and consumers: they can participate in the development of open data initiatives. The government can create dedicated means to support their engagement.
- Business: they create innovative products and services based on open data. To do so, they are dedicated to understand and shape government rules, standards, and regulations about data use.
- Media: they use and interpret open data. The government can involve them in dissemination efforts and discussions of new initiatives.
- Non-governmental organizations: they can use open data for fulfilling their mission. For example, they can promote education of data. Together with the government, they can develop common standards that improve data availability and reuse.

The government is responsible for representing the shared interests of the stakeholders. In fact, the government can serve several roles in order to create value out of data and mitigate risks (Chui et al., 2014):

- Provider: the government is responsible for capturing information electronically. It also releases data publicly and regularly. There is a large responsibility related to data reuse; therefore, there should be a strong focus on improving data quality.
- Catalyst: the government is responsible for building an open data culture. It should bring stakeholders together to make sure that all obstacles are removed. The point is how to motivate various parties to work towards common goals without using legal solutions.
- User: the government is consuming open data. The purpose of using data by the government is to improve various areas of its interest: decision making, service offerings, and accountability. Data is not used in a raw form, but rather

the sophisticated analytics is applied. Therefore, it is also necessary to invest in people, tools, and systems.
- Policy maker: the government is responsible for establishing rules for using data. It should also define data quality standards as well as formats allowing broader reuse of data.

A different set of roles of a government was proposed by Deloitte (2016). Their considerations were anchored in data-driven innovation; therefore, the emphasis was put on less generic aspects. The roles were as follows:

- Data consumer: the government is the first entity to take advantage of rich data resources and realize their potential. It should also consider the transition to making decisions based on data as well as monitored by data.
- Data supplier: the government can offer access to government datasets that can generate economic and social value. The government should identify datasets on demand and provide access to them.
- Regulator: the government should create a suitable regulatory environment. Particular attention should be paid to handling data of private nature. The government can promote opening private data in order to rationalize markets.
- Educator: the government should raise awareness about data-driven innovation. Organizations should be encouraged to base their decisions on data. New set of skills has to be defined for the information age.
- Funder: the government should offer incentives for adopting data-driven processes and practices. The fostering mechanism could be similar to this of R&D. Technological infrastructure should be provided to less technologically advanced organizations.

Although the names are different, we can establish some equivalences. A data supplier is equivalent to a provider, a data consumer—a user, a regulator—a policy maker. Description of an educator has many elements in common with a catalyst. The only role not taken into consideration by Chui et al. (2014) is a funder.

Open data ecosystems have their specializations. For example, Deloitte (2013) proposed public sector information ecosystem (see Fig. 6.3). The main axis of interaction is between a PSI supplier (representing supply) and a PSI consumer (representing demand). Additional players are an enabler and an infomediary. The enabler facilitates open data initiatives without publishing or consuming data, for example, by providing software or platform. It can also be a market organizer. The infomediary works with open data and enhances it, for example, by increasing completeness, accuracy, or accessibility. There is no distinction which role can be played by the government.

An ecosystem with a significantly bigger number of players was proposed by Gonzalez-Zapata and Heeks (2015). They identified eight main open government data stakeholders: politicians, public officials, public sector practitioners, international organizations, civil society activists, funding donors, ICT providers, and academics. The stakeholders were characterized according to two dimensions: power and interest. Figure 6.4 shows the location of the stakeholders in the two

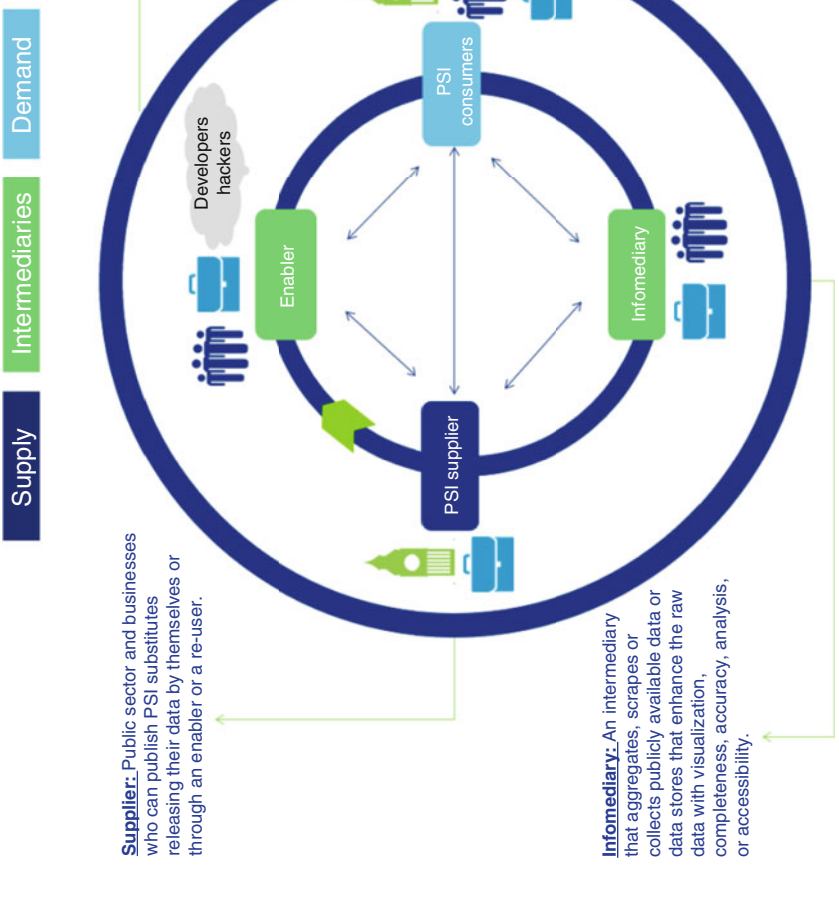



**Fig. 6.3** Public sector information ecosystem. Source: (Deloitte, 2013) licensed under the Open Government Licence v3.0

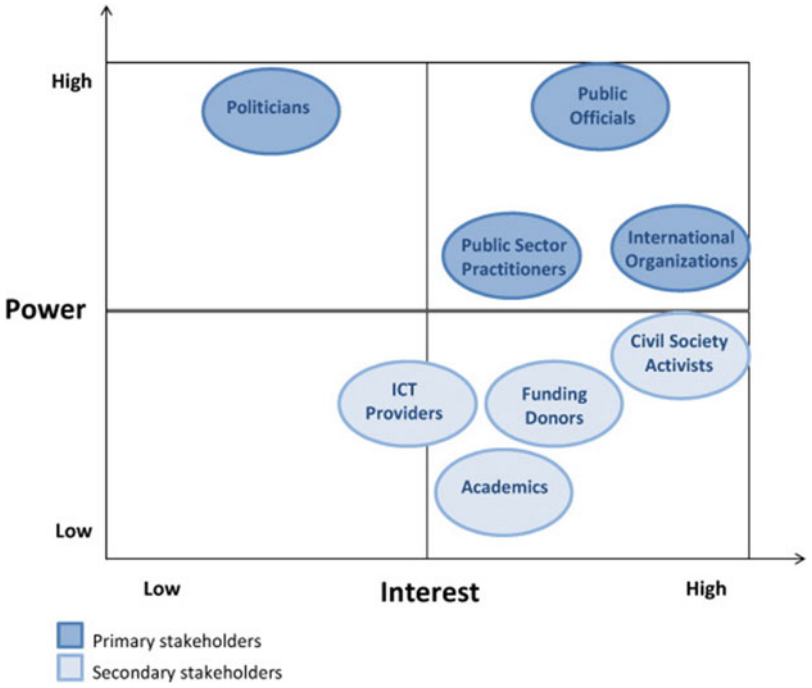



**Fig. 6.4** Stakeholder analysis of open government data in Chile. Source: (Gonzalez-Zapata & Heeks, 2015)

dimensions. What is disappointing is the low interest of politicians. However, the findings of Gonzalez-Zapata and Heeks (2015) were specific as the research was conducted in Chile. Nevertheless, they observed weak participation of entities to whom the data was addressed—private companies that could commercialize data.

Having analyzed the government side, we can move on to data consumers. Deloitte (2013) identified seven public sector information customer archetypes:

1. Large data services company—uses PSI as input to larger data analyses. PSI contributes to profits and new products and services.
2. Independent SME application developer—uses PSI as the key input into products and services in order to generate additional profit.
3. Independent SME focused on efficiency solutions—uses PSI as the key input into products and services. PSI contributes to efficiency savings across public and private sector.
4. Individual user—uses PSI as input to data discovery and own analyses. PSI contributes to improved choice and decision making, leading to savings.
5. Community group—uses PSI as the key input into products and services as well as for own analyses. PSI contributes to improved decision making and efficiency savings across public and private sector.

6. Other non-data companies—can use PSI as input to daily activities, thus improving profits, enabling efficiency, savings, and new products and services.
7. Public sector—can use PSI as input to daily activities, thus enabling efficiency, savings, and new products and services.

Information asymmetries are one of the characteristic features of open data ecosystem. Nielsen (2006) observed a '90-9-1 rule,' describing user participation. Roughly 90% of users are lurkers, i.e., they read or observe, but do not contribute. Next 9% of users contribute from time to time, but it is not their primary activity. Finally, 1% of users are deeply involved and contribute a lot. They account for most contributions. One can have an impression that its their only activity. The distribution of contributors can be estimated with the power law—each next group is one-tenth of the previous group.[1] As (Nielsen, 2006) also noticed, out of about 1.1 billion Internet users, there were only 55 million users (5%), who had weblogs, according to Technorati. Out of those 55 million, only 1.6 million posted daily (3%); however, as many people post several times a day, in fact only 0.1% of users posted daily.

Similar asymmetry was also observed in our research concerning Wikipedia. A very simple measure of the activity of Wikipedia users is the number of edits. Figure 6.5 presents the number of edits by the most active 1000 people. We have used log-log scale here to show that this activity also follows the power law. Their total number of edits was 136,256,516, which made 14.9% of all edits (915,008,845).[2] One thousand users is only a fraction of active users, which Wikipedia defines as "registered users who have performed an action in the last 30 days." Number of active users was 128,267, whereas total number of users was 31,999,162. This means that only 0.4% were active. Once more, the top 1000 Wikipedians (0.003% of all users) were responsible for more than one-seventh of the edits. Thus, more than 99% of Wikipedia users were lurkers.

Lurking can also be valuable in a democratic society where an informed citizen can take effective decisions (Edelmann et al., 2012). The objective of open government data initiatives is to involve as many users as possible. We need to remember that "collaboration is not done for the sake of doing it, but to enable all stakeholders to participate in efficient and effective decisions" (Attard et al., 2015). Nobody is excluded and lurkers can contribute as well if they wish to do it. It can be understood as a kind of externalities where everybody wins. Skewness in the activity of stakeholders is usually reproduced, for example, in value sharing. Lack of representativeness is then connected with big data. Not everybody will adopt open data, and those adopting will be highly varying. The long tail is a fact.

The above situation is not comfortable for sites that depend on user participation. Such sites are usually not representative of average Web users. The most active users certainly differ from the less active group (Nielsen, 2006). For example, a company

[1] It can be formally described as $9 \cdot 10^{-k}$, where $k$ is the rank of group in a range (1–3).

[2] https://en.wikipedia.org/wiki/Wikipedia:Statistics, accessed 2017.10.17.

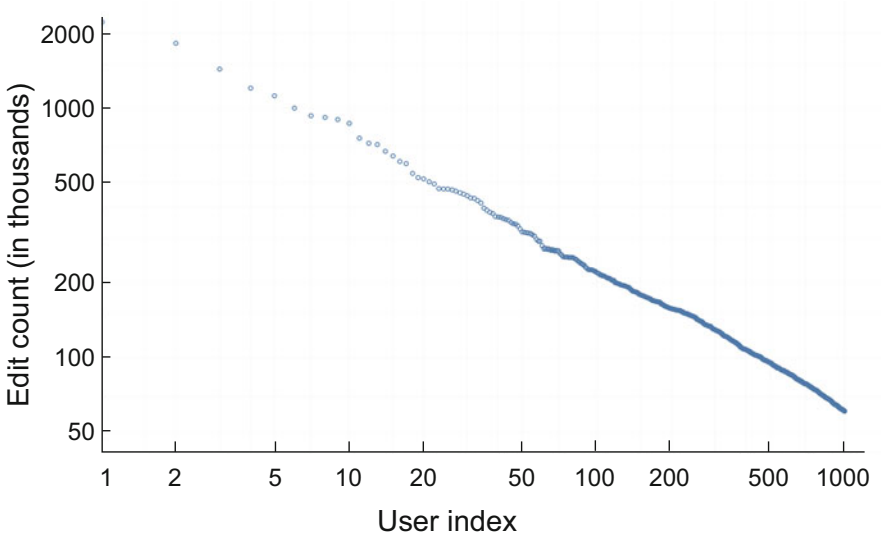



**Fig. 6.5** Number of edits by top 1000 English Wikipedia users on a log-log scale. Source: based on https://en.wikipedia.org/wiki/Wikipedia:List_of_Wikipedians_by_number_of_edits, data downloaded on 2017.10.17

looking for customer feedback gets an unrepresentative sample of opinions about a product. Consumers have similar problems with regard to reviews, which represent only a fraction of users who actually used a product or service. Another example concerns the search for information. Search engines usually sort results based on the number of incoming links. If the majority of linking is done by a small group of people, then the search results can be skewed as well. Small number of users can also cause degradation of information quality. For example, discussion groups can be dominated by low-quality postings.

Fortunately, the awareness of the above phenomena is growing, and so also the responsibility of users using new media. There are good examples where the society was able to impose certain behavior rules, e.g., Wikipedia and Stack Overflow. Edelmann et al. (2012) described such users as *Civil Servant 2.0*. They act as information brokers, are fluent in using the Internet, understand network effects triggered by social media, and as knowledge workers recognize the importance of competition between organizations offering similar information services. We should count them as important citizens of the open data ecosystem.

### *6.2.2 Demand and Supply*

In Sect. 2.4.2 we have described the barriers related to opening datasets. In this section, we characterize data supply quantitatively.

Determination of demand for information is paradoxically difficult, what has already been described by Arrow (1962). The value of information is not known to a buyer unless she has the information. However, in order to verify the information, the buyer effectively purchases the information without cost. The problem is that the seller cannot retain property rights. The seller could reveal part of information, but then the potential purchaser would decide based on less than optimal criteria. The buyer could, for example, act based on value of past information from this source or on the average value of a sample of information. Such a procedure can "lead both to a nonoptimal purchase of information at any given price and also to a nonoptimal allocation of the information purchased" (Arrow, 1962).

Despite the Arrow's paradox, companies strive to determine their demand. According to the research among Swedish IT-entrepreneurs, access to public open data is considered very important for many IT companies. Over 40% find open data essential for the realization of their business plan and over 80% claim that access would support and strengthen their business plan (Lakomaa & Kallberg, 2013).

In order to quantify the demand and supply for open data, we have analyzed data available at `data.gov.uk` for September 2017. There were 42,754 datasets, of which 3540 were unpublished. Only 9920 datasets were viewed at least once, and they attracted 215,933 views. There were 100,344 visits, which means that during one visit there were 2.15 views on average. The detailed breakdown between the main themes (the higher level grouping) is presented in Table 6.2.

With regard to the supply measured by a number of datasets, the most popular category was 'Environment.' It contained 14,174 datasets, making one-third of all datasets. Second place belonged to 'Towns & Cities' with the share of 17.1%.

**Table 6.2** Statistics concerning datasets on data.gov.uk in September 2017

| Theme | Datasets | Views | *Visits* | Datasets (%) | Views (%) | Visits (%) | Ratio |
|---|---|---|---|---|---|---|---|
| Transport | 1213 | 32, 176 | 17, 763 | 2.8 | 14.9 | 17.7 | 5.3:1 |
| Business and economy | 1136 | 13, 741 | 6630 | 2.7 | 6.4 | 6.6 | 2.4:1 |
| Mapping | 2326 | 27, 436 | 9827 | 5.4 | 12.7 | 9.8 | 2.3:1 |
| Defense | 188 | 1750 | 1105 | 0.4 | 0.8 | 1.1 | 1.8:1 |
| Crime and justice | 705 | 6288 | 3097 | 1.6 | 2.9 | 3.1 | 1.8:1 |
| Health | 2131 | 18, 104 | 8602 | 5.0 | 8.4 | 8.6 | 1.7:1 |
| Education | 1320 | 8775 | 4606 | 3.1 | 4.1 | 4.6 | 1.3:1 |
| Society | 3070 | 16, 650 | 8903 | 7.2 | 7.7 | 8.9 | 1.1:1 |
| Government | 3198 | 14, 150 | 7330 | 7.5 | 6.6 | 7.3 | 1:1.1 |
| Government spending | 1847 | 6243 | 2027 | 4.3 | 2.9 | 2.0 | 1:1.5 |
| Environment | 14, 174 | 43, 751 | 17, 286 | 33.2 | 20.3 | 17.2 | 1:1.6 |
| Towns and cities | 7316 | 21, 573 | 10, 763 | 17.1 | 10.0 | 10.7 | 1:1.7 |
| Uncategorized | 4130 | 5296 | 2405 | 9.7 | 2.5 | 2.4 | 1:3.9 |
| Total | 42, 754 | 215, 933 | 100, 344 | 100 | 100 | 100 | |

Source: own calculations based on data from http://data.gov.uk for September 2017. As of 2021, it is not possible to retrieve the updated statistics

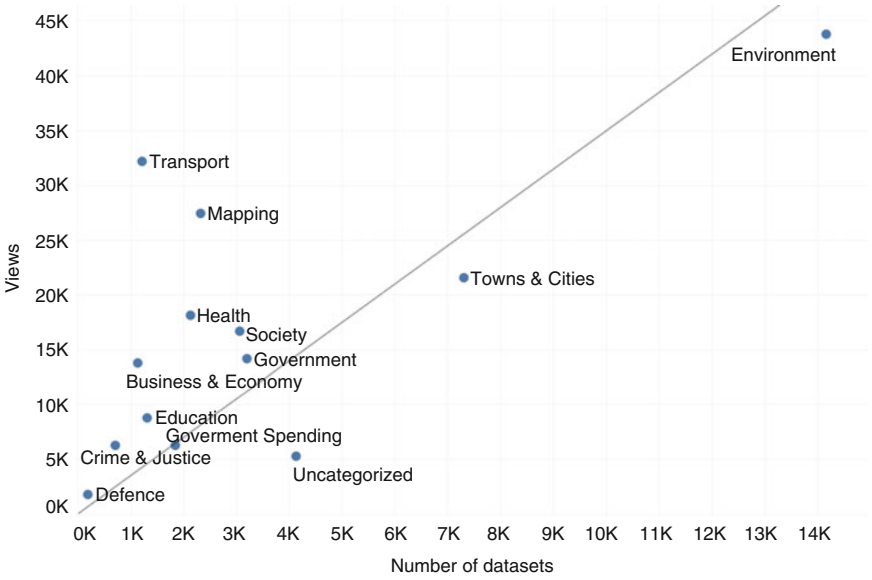



**Fig. 6.6** Supply and demand for datasets in United Kingdom (as of September 2017). Source: own calculations based on data from http://data.gov.uk

About one-tenth of the datasets did not have any theme assigned. The demand side was measured by the actual use of datasets. The leading category was again 'Environment,' with 43,751 views. However, it was only 20.3% of all visits. In order to determine the most relatively demanded theme, we have divided the views share by the dataset share—results are put in the last column 'ratio.' The theme 'Transport' attracted on average 5.3 times more visits than it could be inferred from the dataset share. There were 26.5 views per dataset in this theme. The second place was taken by 'Business & Economy' with 12.1 views per dataset, and the third 'Mapping'—11.8. In theme 'Environment' there were on average only 3.1 views per dataset. Details are presented in Fig. 6.6.

The demand for datasets is very asymmetric. If we consider just 1% of the most popular datasets (99 datasets), they generated 84,951 views, i.e., 39.3% of all the views. This can also be observed for each theme separately, where there are datasets that attract a high demand. In theme 'Transport' there were 1213 datasets that generated 32,176 views in September 2017. When we take just 1% of viewed datasets in this theme, i.e., 6, we get 16,264 views—50.5% of all views. For theme 'Environment' the analogous number is 31.3%. If we count all datasets from two given themes, not only those visited in the analyzed period, the numbers jump to 65.4% and 60.5% respectively. Explanation is straightforward—the distribution of the views follows the power law, which should not be surprising (see Sect. 4.4.2). It is illustrated in Fig. 6.7.

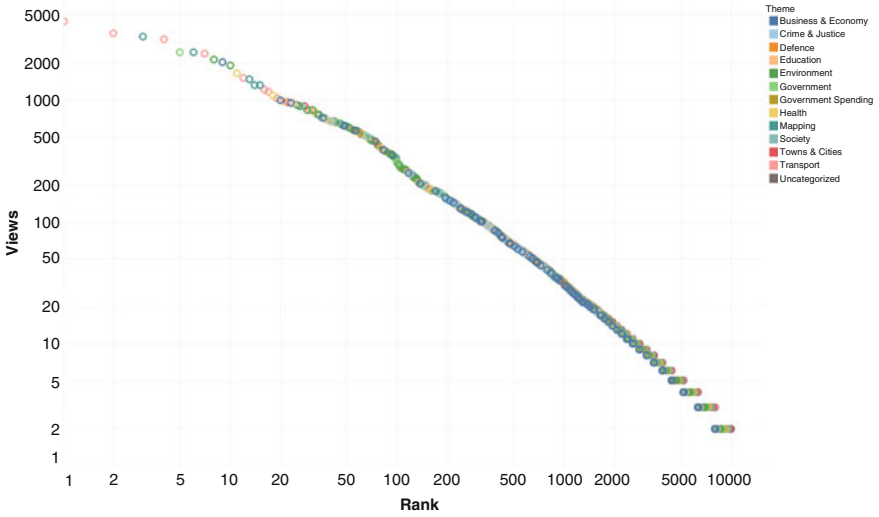



**Fig. 6.7** Number of views per rank for datasets in data.gov.uk (September 2017). Source: own calculations based on data from http://data.gov.uk

Although there are many places where data can be published and then consumed, data management portals like, for example, CKAN focus on metadata management. There is no functionality that allows organizations or individuals to sell or exchange data directly with each other. There is a lack of general marketplaces despite the growing number of data intermediaries. Platforms that offer marketplace functionality are usually tailored to specific scenarios, for example, in logistics or agriculture (Esmeijer et al., 2013).

Organized data sources are more often connected to a specific kind of data or are focused on certain verticals. Besides data brokers, data can also be exchanged through online data markets. There is no strict demarcation line between data brokers and data market providers (OECD, 2015). According to Dumbill (2012), the both provide value-added services, which may be useful for market participants. First of all, they provide a single point of discoverability and comparison for data. Indicators for scope and quality are usually provided. Second, they prepare data for use. Some of the activities can include cleaning data and saving in appropriate formats. Third, they have a better overview of the market; therefore, can offer an economic model for accessing the data. Such models are usually difficult to prepare for those who only publish or consume.

With regard to supply, it is important to think about incentives for data sharing. Many companies accumulate a lot of data, and part of it is a by-product of their core activity, the so-called 'data exhaust' (Turck, 2017). For example, a company offering payment solutions (transfers, credit cards) may also reach for data on what and from people buy. A mobile application may collect geo-location data about user behavior, e.g., movement patterns, buying habits.

Incentives should not only focus on sharing but also on the maintenance of data. Thus, the supply should be stable over time. It is important for a long-term adoption of open data. Many applications would never be created if the creators were not convinced about the stability of data sources. Here, data is the main means of production. One of the negative examples is Twitter, which decided to revoke access to API and reject applications that competed directly with their platform. The developers learned the new risk; therefore, "uncertainty of future access to the API will stifle investment and innovation" (OECD, 2015).

Governments publish data on the Web so that it can be used, reused, and distributed. The typical assumption is the separation of producer and consumer. The roles are definitely distinct but can be played by the same actors. A publisher in one context can be a consumer in another context. Alexopoulos et al. (2014) used the term *data prosumer* to describe this phenomenon. Data consumption can be just data exploration, where a user studies or visualizes open data. It can also have a richer form of data exploitation, where the user adds value to the open data by creating mashups, enriching with other datasets, or making predictive analytics.

### *6.2.3 Joint Production*

Cooperation between companies is relatively common in business. Legal forms of cooperation include, for example, joint ventures and patent pools. Some of the forms are forbidden by antitrust law, for example, a cartel. An international joint venture (IJV) is a partnership between two entities located in two or more countries (Yan & Luo, 2016). Such a strategic partnership allows to avoid some risks related to cross-border transactions or acquisition. It also allows for gaining competitive advantage through access to the partner's resources: people, technologies, or markets. It is perceived as a vehicle for knowledge transfer from common multinational expertise to local businesses with a goal of improving the performance. "A patent pool is a consortium of at least two companies agreeing to cross-license patents relating to a particular technology" (Gibson, 2009). The pool is formed by competitors either to avoid or resolve a conflict. It allows to lower the coordination and negotiation costs, also related to infringement and litigation. Common resources are thus created for the *collective benefit*.

As can be seen from the short explanation above, agreements concerning sharing resources (usually intangible) are very often an important part of these collaboration efforts. Companies decide to share resources under non-discriminatory access regimes "because independent research efforts are inhibited by complexity, expense, strategic concerns, transaction costs, or other impediments" (Frischmann, 2013). The case for collaboration can be very often modeled as the prisoner's dilemma. This dilemma describes the independent decisions of two players—prisoners. It is a classical game in game theory which explains why two rational entities might not cooperate, even if the cooperation is in their best interests. Regarding the sharing of data, one of the specific examples of prisoner's dilemma is 'free riding.'

The motivation for collaboration is simple—individual companies collaborate because it allows them to create value that no single company can deliver on its own (Adner, 2006). Collaborative production is enabled by the ubiquitous presence of ICT, citizens' digital literacy, and their potential willingness to participate online (Edelmann et al., 2012). Models can be created to involve external stakeholders in creation of public value. Moreover, public administration can benefit from increased innovation, when their most valuable resources are used. Engagement of civil society and businesses is a precondition. Particularly, coordination between businesses is necessary. Because it is a collaboration between competitors, i.e., collaboration and competition at the same time, it is sometimes referred to as *coopetition* or *collabpetition*. Such collaboration is crucial for leveraging the potential of the multidisciplinary field of data-driven innovation (Woo, 2013). The number of partnerships can serve as an indicator of market activity.

## 6.3 Mutual Benefits

In the macroeconomic context, we have shown that the direct value of data can be increased thanks to various multiplier effects (see Sect. 5.4). From a microeconomic point of view, such impact concerns both entities participating in a transaction. Innovative and more efficient use of data is beneficial not only for its users but also for data owners. The owners can extract insights from data usage and improve their offer. Opening data is one of the ways to collect valuable feedback from a broader user base. Such opinion is particularly important for the public sector and research institutions.

Sometimes better information can have a negative impact on the welfare. The parties may refuse to trade when one party is known to be better informed than the other. This problem is referred to as the market for *lemons*,[3] and has already been raised by Akerlof (1970) who studied how the quality of goods traded in a market can degrade in the presence of information asymmetry between buyers and sellers.

The welfare is not reduced when the presence of improved information is not known or information is shared, effectively reducing the asymmetry (Brynjolfsson et al., 2011). Reduction of information asymmetry is thus one of the forms of mutual benefits. Proliferation of information on the Internet can help to reduce the problem of insufficient information for buyers. Various information services, whether in the form of open listings or price comparison engines, can make buyers feel more confident in deciding on a purchase. This is also beneficial for sellers who can ask a premium price for products which are of higher quality than general-audience ones.

[3] A lemon in American slang means a device or machine that does not work well.

### 6.3.1 Community Involvement

In order to get the full value from open data initiatives, it is necessary to involve a community. At the foundation are models of collaboration and innovation. A set of such models was proposed by Pisano and Verganti (2008) who considered two criteria: whether the membership is open or closed and whether the governance is flat or hierarchical. These two criteria set up four basic models of collaboration: elite circle, innovation mall, innovation community, and consortium.

Based on the above Edelmann et al. (2012) proposed a collaborative public value production, as depicted in Fig. 6.8. The whole process is directed by the policy cycle of governmental projects. Government and public administration interact with participative stakeholders by sharing data and information about the project. A transaction is necessary to bring innovation to public sector projects. Both exchanges can be top-down (initiated by government) or bottom-up (initiated by stakeholders) processes. Transparency is necessary in order to follow the project development. Government and stakeholders as well as lurkers finally disseminate the results of the project to the civil society. The collaboration and innovation process results in new or increased public value.

Edelmann et al. (2012) pointed at the intrinsic conflict between maximizing shareholder value and public value. Public administration aims at maximizing this value for the benefits of citizens. Moreover, there is no monetary remuneration of citizens and the only social compensation can be in the form of ratings and reputation. As a result, the protocol of collaboration between public administration

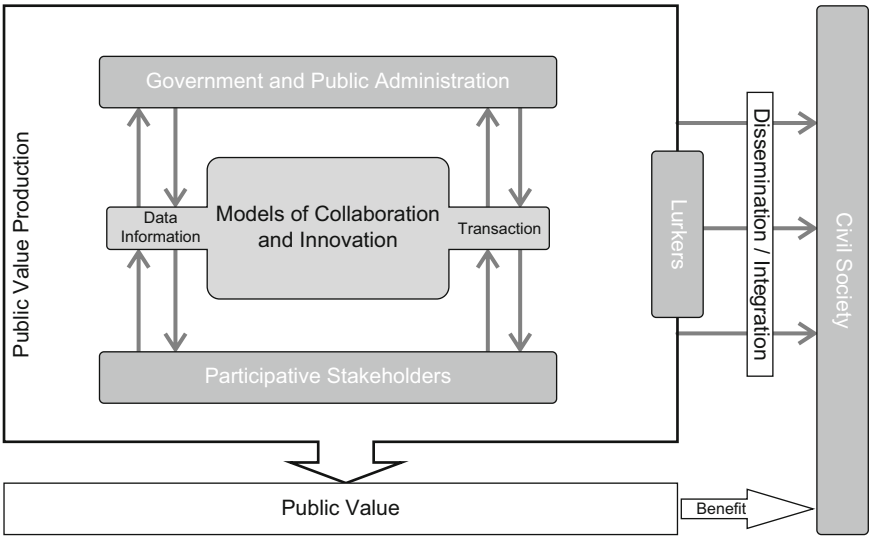



**Fig. 6.8** Collaborative public value production. Source: own work based on (Edelmann et al., 2012)

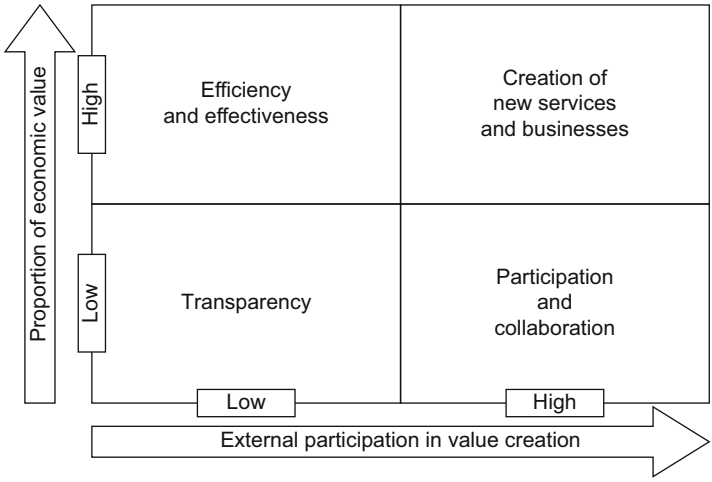



**Fig. 6.9** Value drivers of strategic OGD initiatives. Source: own work based on (Jetzek et al., 2012)

and stakeholders needs to be much more complex than in typical commercial projects.

Jetzek et al. (2013) discussed value drivers of open government data initiatives. One of the dimensions concerned external participation in value creation (see Fig. 6.9). The most interesting initiatives can only be achieved through the involvement of business entities.

The idea of community involvement was extended by Balka et al. (2014) who observed trade-offs between openness to external value creation and closedness for internal value capture. They studied forms of openness and their relation to community involvement. There are three different archetypal forms of openness: transparency, accessibility, and replicability. Community involvement can take one of three stylized forms: observation, co-development, and co-production. Table 6.3 shows relations between these notions: observation implies transparency of information, co-development requires accessibility, and co-production—replicability.

**Table 6.3** Openness vs. community involvement

| | Role of the community | | |
|---|---|---|---|
| Form of openness | Observation | Co-development | Co-production |
| Transparency | Yes | * | * |
| Accessibility | * | Yes | * |
| Replicability | * | * | Yes |

* not required by definition

Source: (Balka et al., 2014)

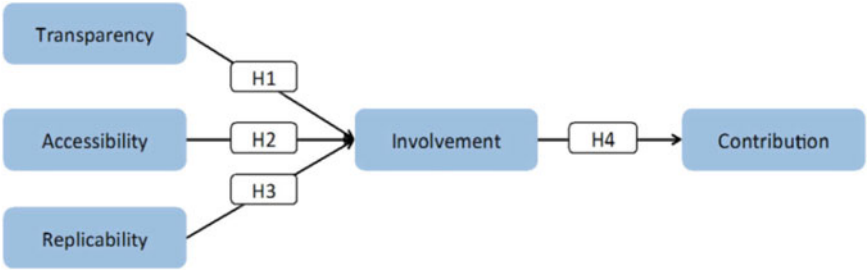



**Fig. 6.10** Balka's research hypotheses. Source: (Balka et al., 2014)

The prerequisite for the involvement of communities is the possibility to observe the development of a new product design; therefore, such information should be transparent (West & O'Mahony, 2008). Transparency encompasses two aspects: the possibility and the right. Communities can co-develop new designs if the design information is accessible and the design can be modified. This requires both the possibility, i.e., the design is accessible in an editable format, and the right, i.e., the design is available under open license. The term co-production means that not only the design but also the production process is distributed; hence; the design needs to be replicated. Replication is understood as the possibility (availability of components) and the right (absence of legal barriers) to produce instances of the design.

Balka et al. (2014) researched the impact of transparency, accessibility, and replicability on contribution and formalized it in a set of four hypotheses (see Fig. 6.10). They first verified if various forms of openness increase the feeling of involvement in an innovation project (hypotheses H1, H2, and H3). Then they checked if the feeling of involvement results in the higher effort and contribution (H4). They concluded that perceived openness affects community members' contribution behavior. Different effects were observed for software and hardware. Transparency and accessibility of the software strongly affected involvement. The effect of software replicability was found to be smaller. No significant effect was found for hardware openness.

### *6.3.2 Value Networks*

Network is very often a key word in value creation. As Allee (2008) stated, "value is an emergent property of the network, so understanding the functioning of the network as a whole is essential to understanding exactly how and why value is created." She dealt with value network analysis that addressed the conversion and utilization of intangible assets. Intangibles are negotiable economic offerings. Knowledge, which is an intangible asset, is one of the most interchangeable

commodities. Knowledge can be exchanged for another form of intangible value (e.g., a favor) or it can be traded for money as a tangible form of negotiable value.

Based on analysis of the literature on open government data, Jetzek et al. (2012) came to the conclusion that value creation initiatives of open government are similar to those met in value networks. Allee (2008) defined value networks as "any set of roles and interactions in which people engage in both tangible and intangible exchanges to achieve economic or social good." In a value network, value is *co-created* or *co-produced*. As the value emerges from exchange there are at least two entities, but usually more entities are involved. Creating value cannot be done one-sidedly, i.e., based on the efforts of a single organization. In a network, the divergent interests of all collaborating partners need to be considered. Fortunately, information is non-rivalry good.

Value networks have also a role in peer production analysis. Morgan et al. (2010) analyzed how firms create and capture value with open source software, and they also contributed to open innovation.

### *6.3.3 Data-Sharing Economy*

According to the calibrated model by Klenow and Rodriguez-Clare (2005) world GDP would be only 6% of its current level if countries did not share ideas. In absolute numbers, it would be $3 trillion instead of $50 trillion. These are the observed scale effects from the non-rivalrous nature of knowledge. Among the channels of international knowledge spillovers, they mentioned trade, foreign direct investments, joint ventures, migration of key personnel, and imitation.

Before we can discuss the phenomenon of sharing, we need to study the Arrow's information paradox. He analyzed what decisions can be made by information owners regarding exploitation. The owner can extract the economic value for own purposes. However, if optimal allocation is to be achieved, the information should rather be sold. The owner is a monopolist to some extent; therefore, he can take advantage of this fact. The problem is that in the absence of legal protection, the information can only be sold once. The information can be easily replicated, so any single buyer can destroy the monopoly. Thus, the only reasonable decision for the owner of information is to stay with the first option. Unfortunately, the monopoly is socially inefficient and not very useful for the owner of the information. There are usually other entities that could use information more effectively. "The very use of the information in any productive way is bound to reveal it, at least in part" (Arrow, 1962). Findings by Arrow are thus a strong argument for sharing of information for a broader social benefit.

Nowadays, the sharing economy revolutionizes the modern consumption (Abramova et al., 2015). People can enjoy the advantages of possession while minimizing the responsibility related to ownership. This is particularly visible on digital marketplaces, which very often refer to sharing, renting, or swapping. Sharing economy has several unique characteristics: the absence of ownership

transfer, focus on services, and tighter interaction between parties. This creates new uncertainties the customers have to deal with and necessitates the development of uncertainty-reducing tools. What also becomes more important than in traditional e-commerce is a community building.

OECD Global Science Forum has identified nine challenges concerning data sharing (OECD, 2015). First, new forms of data are generated but their statistical validity is not yet assessed. This is due to the second issue—there are no micro-data records available for national statistics. There is also a specific challenge concerning individuals' privacy as new forms of personal data are increasingly created and collected. Fourth, there are barriers that impede cross-border collaboration. Therefore, global research agendas require international and interdisciplinary coordination. Moreover, experience should be shared across countries to develop comparable data resources. Next, researchers are often not able to assure the availability of data for reuse. More investments are necessary in data creation and curation to avoid data loss. Finally, there should be incentives for researchers to ensure effective data sharing.

The value of data sharing can also be shown by studying the loss caused by keeping information closed. The first case concerns fraud by automobile insurance, the so-called 'crash for cash.' According to the UK's Insurance Fraud Bureau, the fraudulent collisions cost around £340 million a year (IFB, 2016). When an insurance company analyzes the damage report, it usually is missing information about similar accidents and claims by other companies. A single transaction may look normal, but it may become suspicious when information can be linked to other transactions and related entities. The loss of insurance industry can be reduced if data is integrated and analyzed in a smart way (SAS, 2017).

The reluctance to share can also bring negative consequences. Until 2011, Financial Times (FT) offered an app for iPhone and iPad on Apple's App Store. It was then removed because FT could not agree with Apple on access to its customers. The disagreement concerned not the sharing of revenue, but the sharing of data—FT was not able to interact directly with own subscribers. Therefore, it finally decided to develop own application based on HTML5. This allowed to bypass the restrictions of Apple and know the customers better, resulting in the increase of digital subscribers by 14% (Reuters, 2011).

Another domain of data sharing, bringing significant social benefits, is healthcare. An interesting case of Project Data Sphere (PDS) was described by (Bolen, 2017). It is a not-for-profit initiative established to help cure cancer. The platform allows sharing, integrating, and analyzing historical patient-level data from phase three cancer clinical trials. It promotes innovation and opens new research possibilities. Exploitation of the platform has already resulted in publication in The Lancet Oncology (Wilkerson et al., 2017).

Data sharing and collaboration offers researchers opportunities for exploring new ways to fight cancer. Clinical research generates a lot of data. Each clinical trial is conducted to answer specific questions; therefore, the integration of data is not obvious. Results should be integrated even though the hypotheses were not confirmed. Collecting data from multitude trials conducted by different pharmaceutical

companies allows discovering new patterns. Data sharing can draw attention to links and correlations between phenomena that usually are not measured together.

It is noteworthy that within PDS data sharing occurred in spite of competition. There were many companies joining the initiative, and they were able to discuss in an open and meaningful manner. Moreover, it happened in the healthcare sector, where data is particularly secured. Such consortia allow to define a common understanding of data.

Not every initiative of sharing medical data is so successful. The Liverpool Big Data Collaboration for Health project aimed to link health records between NHS trusts to exploit big data analytics (Baldwin, 2014). It was an innovative combination of IT and healthcare. Based on the initial success, the NHS planned to expand the data collection program in hospitals to include general practice patient care data. This Care.data program was then criticized for such expansion, as the NHS failed to explain the benefits of data to the general public (Glick, 2014).

## 6.4 Data Ownership

Prior to Web 2.0, the online content was created by professionals and technically skilled users. Currently user-contributed content plays an increasingly important role in the development of the Web. It is even overtaking professionally created and curated resources (Marshall & Shipman, 2017). For example, Facebook, the world's most popular media owner, creates no content (Goodwin, 2015). Concerning only English language, Wikipedia, a community-created encyclopedia, is 60 times bigger than the next largest English-language encyclopedia, created by dedicated editors—Encyclopædia Britannica.[4]

Data provision by users has already taken its place in enterprise settings. Similarly to outsourcing where some internal activities of a company are contracted out to another company, we have a notion of crowdsourcing where a company uses contributions from Internet users in order to obtain needed services or ideas. Companies are increasingly able to source knowledge from anywhere and at any time. According to Vivek Bapat (SAP), "work is no longer a place, but an activity" (van Es, 2014). We can thus foresee that companies will be increasingly interested in leveraging this trend, and more and more people will work in crowdsourcing. In the future, not only we do not need to live where we work, but potentially there is no single employer.

The benefits of crowdsourcing have already been observed also in public administration. They should be able to consume crowdsourced data and turn it into actionable information. Such programs are very popular in developing countries when feedback on local issues is the most important (Landry et al., 2016).

[4] https://en.wikipedia.org/wiki/Wikipedia:Size_comparisons.

The important question then arises: who owns data? The answer is important for discussion about value creation. Data typically involves complex assignment of various rights across different data stakeholders (OECD, 2015). When we consider the network structure and interconnectedness of data-driven services, the attribution of value is even more difficult.

### *6.4.1 Access to Data*

Users leaving digital footprints on the Web usually have no idea what will happen with their data and how it will be used. The problem is not that companies want to use data. The real problem is that consumers do not care for data enough, do not have the means to do it, or do not know how to do this (ODINE, 2017). Such a situation is partly caused by companies themselves—most privacy statements are long and complex. This problem was recently highlighted by Australian advocacy group 'Choice,' which asked an actor to read Amazon Kindle's 73,198-word terms and conditions to discover that it requires nine hours (Choice, 2017). People do not understand what is the purpose of data collection and how it is then processed. Automatic analysis of licenses can be helpful (Nejad et al., 2016). Nevertheless, as Berners-Lee (2017) noticed, "we've lost control of our personal data."

Having access to their own personal data, consumers can get insights into their own consumption. They can also discover information that will lead to changes in their behavior. When access is granted to the public, users are empowered to provide feedback that can be used to improve the quality of goods and services (Manyika et al., 2013).

Pentland (2013) postulated to compile a *New Deal on Data* in order to achieve a data-driven society. It should guarantee that the data needed for public goods is readily available and that the interest of citizens is protected. Personal data should be treated as an asset. Pentland (2013) suggested analogy with the English common law tenets of possession, use, and disposal: "1. You have the right to possess data about you. 2. You have the right to full control over the use of your data. 3. You have the right to dispose of or distribute your data."

Marshall and Shipman (2017) looked at the challenge of managing intellectual property as the creation, publication, and sharing of content proliferates. They considered three actions related to content: saving, sharing, and removing. Their analyses were based on different media types and services: tweets, photos, reviews, podcasts, recorded videoconferences and educational videos, recordings from multi-player online games, and Facebook content. They observed an increasing dominance of non-professional digital media, where professionally produced media was copied, excerpted, or remixed.

### 6.4.2 *Ownership Roles*

In Sect. 6.2 we have defined stakeholders on the data landscape. When using data, we need to assign various rights to various stakeholders. These rights are usually much more complex than in the case of other intangible assets (OECD, 2015); therefore, we extend the typology of stakeholders with ownership roles.

The starting point can be activities, i.e., what can be done with data. Some examples include access, create, modify, package, derive benefits, sell, remove, and assign rights to other users. Loshin (2001) identified the following stakeholders that could claim data ownership. These can also be interpreted as various ownership paradigms:

- Creator—the party that creates or generates data owns the data.
- Consumer—the party that consumes data owns that data. Ownership is bound to responsibility.
- Compiler—the party that selects and compiles information from different information sources, owns the data. The compiler is adding value and may expect to leverage the benefits of ownership.
- Enterprise—all data that enters the enterprise is automatically owned by the enterprise.
- Funder—the user that orders the data creation claims ownership.
- Decoder—in environments where data is 'locked' inside particular encoded formats, the party that can unlock the data becomes the owner of that data. The cost of decoding is an investment.
- Packager—the party that collects data and formats for a particular use becomes the owner of that data. It is similar to the compiler paradigm.
- Reader as owner—the value of any data that can be read is subsumed by the reader. The value is derived from learning.
- Subject as owner—the subject of the data claims ownership of that data. Usually this happens in reaction to another party claiming ownership of the same data.
- Purchaser/licensor as owner—the party that buys or licenses data can claim ownership. The purchaser assumes that the investment made in acquiring the data yields ownership (similarly to the funder paradigm).
- Everyone as owner—it expresses the belief that data should be available to all with no restrictions (global data ownership).

As can be seen from the above compilation, usually there will be no single data stakeholder having exclusive rights, and stakeholders will also differ in power, which can be contextual. The situation is particularly complex where personal data is considered. Certain rights of the data subjects cannot be waived (OECD, 2015).

### 6.4.3 *Open Algorithms*

Data ownership is increasingly associated with the ownership of data infrastructures. Data is increasingly collected and managed by private firms. Datasets that were formerly generated by national statistics offices are now better represented in private companies (Enabling the Data Revolution, 2015). A simple example is the price index: many years ago, price lists were collected manually. Currently, big online shops have a much better overview of the market (Eurostat, 2013). Another example concerns census—researchers at the University of Washington demonstrated how Facebook can serve as a more current source of information about migrations (Zagheni et al., 2017). Dataset can also be first established from private data sources, e.g., advanced monitoring systems installed in the road network (Ma, 2016).

As Croll (2011) earlier noticed, the important question is not who owns the data but rather who owns the means of analysis or in other words—who can put that data to work. If data is secured by private companies, it cannot be used for a broader public good. Many firms know the value of their data but usually they do not realize that the data can be used to generate public good value. By opening datasets, companies can also profit themselves, e.g., by helping to grow economies or prevent epidemics. Some companies, mostly telecoms, decided to open samples of data for research purposes, usually organizing various competitions. Their promising results caused a growing demand for more private data to be made available. Data journalist Kenneth Cukier even stated that "not using data is the moral equivalent of burning books."[5] However, he also reminded about responsible handling of data. There is then a dilemma between privacy and utility. Companies willing to open data encounter legal, ethical, and other barriers (see Sect. 2.4.2). Indeed, not all data should be open. Means for secure sharing of personal data are described in Sect. 6.5.2.

Open algorithms are an interesting idea that addresses the problem of data sharing by enterprises without affecting the data itself. The idea is to send the code to the data rather than bring the data to the code; hence, 'open data' became 'open algorithm.' This approach is implemented by the Open Algorithm project (OPAL), which "aims to unlock the potential of private data for public good in a privacy-conscientious, scalable, socially and economically sustainable manner."[6] The founding organizations—Data-Pop Alliance, Imperial College London, Massachusetts Institute of Technology, Orange Group, and World Economic Forum—are developing a platform to unleash the power of big data held by private companies for public good.

OPAL consists of an open platform and algorithms that can be run on the servers of partner companies. OPAL architecture is presented in Fig. 6.11. The left part—'secure datasets'—represents the components on the side of a company owning

[5] "Digitising Europe Initiative," http://www.vodafone-institut.de/event/not-using-data-is-the-moral-equivalent-of-burning-books/.

[6] https://www.opalproject.org/.

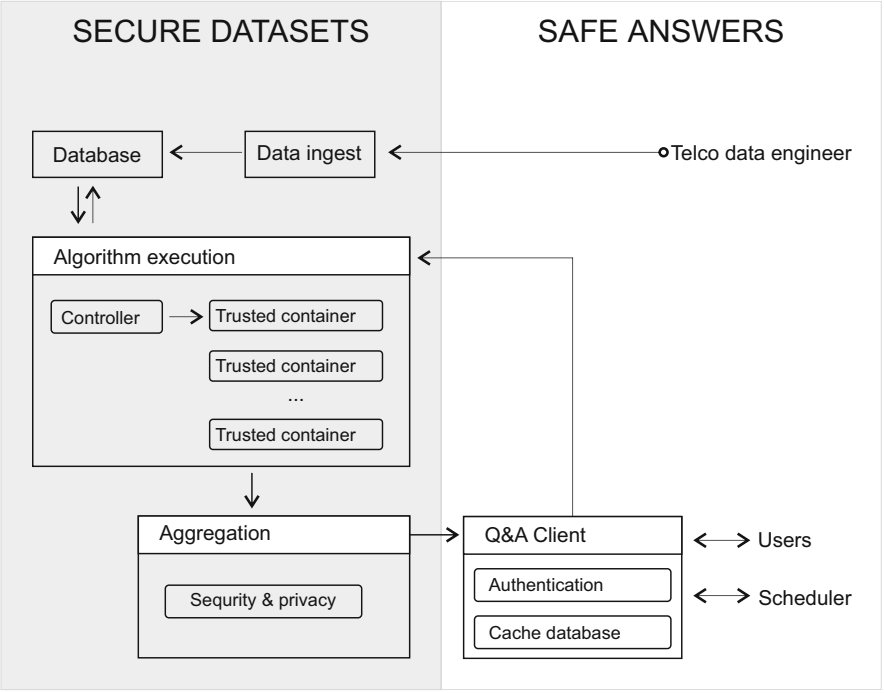



**Fig. 6.11** OPAL architecture. Source: based on https://www.opalproject.org/closer-look/ retrieved in 2019

sensitive data. All processing is executed within the company servers, so data remains private. The right part represents external stakeholders—they can ask a question and receive 'safe answers.'

OPAL is an interesting proposal to solve a confusing and difficult problem of accessing big data sources for research and policy purposes. The findings that may be interesting to policy-makers can be discovered without releasing the data itself. Before OPAL, data held by private companies could only be used and analyzed externally, when data was made available for a data challenge, such as the Orange's D4D Challenges[7] or through bilateral agreements. Direct access to data offers several advantages: greater accuracy, higher frequency, improved granularity, and bias avoidance. Utility of data is increased. There are also additional intangible effects—building trust between all parties involved: private companies, official statistical systems, and citizens (Roca & Letouzé, 2016). The users need not be restricted to public administration. Insights discovered in private datasets can also be offered to other private companies for a price, as long as the privacy of data is assured.

[7] http://www.d4d.orange.com/.

## 6.5 Economics of Personal Data and Privacy

According to EU (1995), personal data is "any information relating to an identified or identifiable natural person ('data subject')." Personal data can be collected in a variety of ways—data can be volunteered by individuals, can be legally observed, or can be inferred. The development of digital technologies and their adoption by organizations and individuals allow the collection of more data—cheaper and faster (Acquisti, 2010). Regardless of the way data is collected, it has to be protected, what implicates legal and economical consequences.

### *6.5.1 Role of Regulations*

The protection of personal data has not always been a legal obligation. The neoclassical economic theory of perfectly competitive markets states that complete information leads to economic efficiency, where 'complete' means that relevant information is available to all market players. The Chicago school of economics points at various inefficiencies related to the protection of privacy. The reason is that potentially relevant information is hidden for some individuals on the market. Thus, economic resources are used inefficiently or rewarded unfairly (Posner, 1981; Stigler, 1980). Chicago School's approach was criticized for assuming the rational behavior of market players. Potential discrimination of some kind is the standard reason behind privacy protection. Taylor (2003) also pointed at the problem of over-investing in collecting personal information. Some economists claimed that data protection could have a positive effect on economic welfare (Hermalin & Katz, 2006).

Establishing a better understanding of the role of personal data in the economy would facilitate policy makers in working on data privacy regulations. Usually transparency and privacy are treated as dichotomous constructs—governments can focus on transparency, neglecting privacy. According to Janssen and van den Hoven (2015), they should be considered as complex and interrelated factors, and only then the impact of big, open, and linked data on privacy can be understood. These values should be balanced with other values, such as security, safety, or openness.

Investigating open data from an economics point of view can help us to find an optimal selection of information to be shared between government, enterprises, and society. A self-regulated, market-driven approach may not be sufficient to achieve an increase in welfare. New technologies like linked data, cheaper storage, etc. are necessary conditions but are not sufficient to achieve a higher level of balance. Sometimes the intervention is necessary to achieve a certain equilibrium. Grant et al. (2014) pointed out that if governments create uniform platforms for information exchange, with appropriate safeguards, they can benefit from ensuring that data is adequately protected. Pentland (2015) postulated to use technology instead relying on traditional bureaucracies to make sure that data stewardship follows the rules

specified by law. Citizens must have effective and direct supervision over data about themselves in order to avoid autocratic control.

### 6.5.2 Secure Sharing of Information

Manovich (2012) distinguished three classes of people in the big data domain, which is also important for personal data: (1) those who create data, both consciously and by leaving digital footprints; (2) those who have the means to collect it; and (3) those who have the expertise to process it. If there are different entities who have the necessary capabilities or skills mentioned above, then data should be transferred between them in some way. The issue of sharing personal data emerges.

Data sharing is intrinsic to open data initiatives. Among the challenges for data sharing, OECD (2015) mentioned the particular situation of personal data. They observed that new forms of personal data are emerging, such as social networking data. What is more troublesome, such data is increasingly created and collected, and the use of these data may pose risks to individuals' privacy.

Depending on the decision taken, information sharing can lead to various economic consequences for data subjects and data holders (Acquisti, 2010). Basically, it has to be decided which data to protect and which to share. On the one hand, individuals want to avoid the misuse of data, and on the other hand, they can benefit from sharing. Mutually satisfying interactions are in fact possible. There is a trade-off, which can be solved on the grounds of economics. This can be applied to situations where public institutions, e.g., statistical offices, cannot share detailed data but can sell aggregated data. If the buyer wishes to obtain customized analyses, then there should be a way to make it possible. Privacy Enhancing Technologies (PET) can be one of the solutions. Such technologies should be *effective*, i.e., identification of individuals should not be possible or be prohibitively costly, and *efficient*, i.e., regular transactions could be completed without additional costs. They allow to reach equilibria where data holders can analyze anonymized and aggregated data while individual data remains concealed.

One of the possible approaches to deal with the above dilemma is using data de-identification methodologies (Li et al., 2007): k-anonymity, l-diversity, and t-closeness. The goal is two-fold: anonymized data should give a scientific guarantee that subjects cannot be re-identified and that data remains useful for drawing conclusions. Usually the granularity of a data representation is reduced, which results in decreased effectiveness of data analysis. "A release of data is said to have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least k-1 individuals whose information also appears in the release" (Li et al., 2007). In the case of 3-anonymity there are at least 3 records for each equivalence class of identifying attributes. Unfortunately, k-anonymity does not prevent attribute disclosure if we know the subject is in the dataset. It was improved by l-diversity, which additionally maintains the diversity of sensitive fields, e.g., disease, income. "An equivalence class is said to have l-

diversity if there are at least l 'well-represented' values for the sensitive attribute. A table is said to have l-diversity if every equivalence class has l-diversity" (Li et al., 2007). L-diversity still can reveal the distribution of sensitive attributes. T-closeness is a further refinement of l-diversity group-based anonymization. "An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness" (Li et al., 2007). These methodologies can mitigate the risk of revealing sensitive personal data. It should also not be harmful for data quality. Nevertheless, there are cases when anonymity may be difficult to achieve, for example, equaling the number of positive and negative records can result in a very small sample.

### *6.5.3 Value of Customer Data*

Tucker (2010) focused mostly on the advertising value of customer data. The value stemmed from *measurability*, the possibility to measure the impact of advertising, and from *targetability*, the ability to precisely select who will see the advertisement. Better online advertising consists in directing the message to people who are willing to listen. By better knowing of a customer, the company can increase the ability to address customer needs. Although targeted advertising is more expensive, it is also more efficient, allowing reduction of the costs of incorrectly targeted ads. Summarizing, this area is relatively well researched and used in practice. What is interesting is the economic value of online customer data outside of advertising.

The first area of this value is the website and communications design. The data can be used to tailor products automatically to consumers' needs as well as to get immediate feedback. The second area is product design. Users are able to generate their own content, e.g., online reviews allow optimization of product designs. The third area concerns improving recommendations. Recommender systems leverage a history of other customers' purchases to offer product recommendations to other customers with similar preferences. The fourth area covers organizational and operational efficiency. Information on purchase trends can be used to adjust the supply levels. Thus, supply chains may be managed more efficiently. Finally, there are also privacy and societal implications. Collected data can improve the product offering by reducing information asymmetry. Such an effect represents a real gain to society.

The asymmetry can be observed at least in the monetary valuation of the same data item by various market players. For example, according to economic experiments and surveys in the United States, people are ready to reveal their social security number for $240 on average. The same number can be obtained for less than $10 from data brokers. Thus, value of personal information can be estimated in different ways (see Fig. 6.12). There are two general approaches: market valuation vs. individuals' perception of value. The first group based on market valuation includes, among others, market capitalization. Value is estimated based on financial

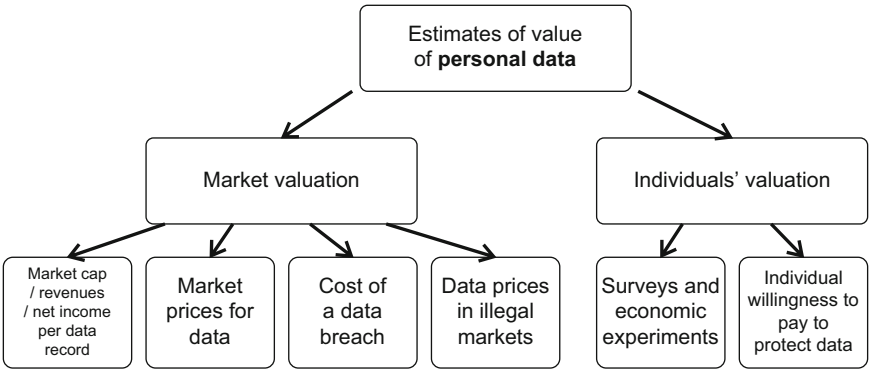



**Fig. 6.12** Estimates of value of personal data. Source: own work based on (OECD, 2013b, p. 19)

results (e.g., revenues, net income) per individual record. Such an approach is useful for business models where companies profit mostly from personal data. The second method bases directly on the market prices at which personal data is traded, both by data brokers and illegal markets. Third, the valuation can consider the economic cost of a data breach (both for companies and individuals). It captures the market costs of damage rather than the value of data itself. There are also two estimates based on individuals' valuation. First, we can use economic experiments and surveys, which inquire for how much individuals would be ready to sell their own personal data. Second approach is based on the idea of insurance, i.e., how much would an individual pay to protect own data.

Special kind of personal data is mobile data. Many telecom operators, researchers, and governments explored ways to deal with the right to privacy—anonymization, aggregation, opt-in/opt-out models, regulations, and legislation (Naef et al., 2014b). There are several publications concerning value of mobile data (Andrienko et al., 2013; Zafeiropoulou et al., 2012).

Protection of personal data is particularly difficult as new application environments are introduced. For example, a new generation of Facebook and mobile applications may gain access to users' most private details circumventing the privacy settings (Krasnova et al., 2014).

### *6.5.4 Benefits, Costs, and Externalities of Disclosed Data*

Externalities from disclosed data are understood as an impact beyond the initial effect of data collection. Acquisti (2010) studied costs and negative externalities of disclosed data. He mainly considered the costs of disclosed data and privacy intrusions but mentioned also the costs of protecting data and the benefits of protected data.

Analysis of a large amount of data on a macro scale allows to predict demand trends and preferences for products, thus minimizing the risk of incorrect resource allocation. Various agencies researching economic situation can profit from additional data. There is a trend of using alternative sources, i.e., sources traditionally not associated with a given domain (see Sect. 5.3.5). For example, Foursquare was able to predict the fall of income of Chipotle restaurant chain (Turck, 2017; Turner, 2016) based on their check-ins data. From such mobile data also other businesses can profit: credit reporting, real estate industry, raw materials, etc.

Better marketing information is not only about targeting correct customers but also about the development of niche products. Such products can only be developed when an inventing company can determine if there are potential customers, thus reducing the risk of development.

Macroeconomic benefits can also be observed thanks to the analysis and aggregation of individual data manifested in online behavior or measured with sensors. Web searches allow to identify the outbreak of infectious diseases (Wilson & Brownstein, 2009). Readings from GPS in mobile phones allow to monitor the traffic and recommend faster routes. Internet of Things can be used to report on environmental conditions.

There are also costs of not disclosing data. For example, lack of access to customer data can be a barrier for market entrance, and as a result, the competition can be limited. Companies with restricted access to market data cannot offer innovative products and new services. Costs of undisclosed data can also be born by a society as a whole. For example, when one country decides not to collect some data that was previously mandatory, then it can have a negative impact on researchers or policy makers.[8]

The willingness to disclose the data ultimately depends on the relative valuations of the parties interested in data. Additional data shared with the data collector will increase its profits only if the data donor can expect additional services in reward (Acquisti & Varian, 2005; Noam, 1997). Customers, as data donors, are increasingly willing to share personal data for direct or indirect benefits (Naef et al., 2014b). At the same time, they are more concerned and suspicious about how various organizations use their data. In fact, users are usually not aware of all uses and profits achieved by other entities from using the data posted by them.

## 6.6 Innovation as Value

Hansen and Birkinshaw (2007) recommended viewing innovation as a value chain comprising three phases: idea generation, conversion, and diffusion. Various frameworks can be used to capture value from innovation. When producing innovative outputs depends on data we speak of ‘data-driven innovation’ (see Sect. 6.6.2). Thus,

[8] The case of Canadian Ministry of Industry was given in (Acquisti, 2010).

data-driven innovation is usually relevant to big data. When open data is considered, the more appropriate framework to use is 'open innovation' (see Sect. 6.6.3).

As observed by Varian (2014), there are many organizations that have interesting data but have no internal expertise in data analysis. There are also data analysts that do have expertise but have no data. These are two sides of the market that can benefit from cooperation under open innovation umbrella.

### *6.6.1 Analytics as a Product*

Data is an asset, but it is rarely sold in a raw form. The value comes from the insights that are extracted from data. Therefore, data is sold rather as a service where data is already processed, integrated with other data, or linked with other data for a broader context.

Turck (2017) provided several examples of companies that successfully monetized their data assets. Facebook does not sell its user data in raw form; it profits from data-driven advertising products that enable to target Facebook users. The same strategy is followed by Foursquare, which allows to analyze real-world foot traffic using Foursquare Analytics.[9] Twitter evolved into a data platform, which allows measuring engagement and exploring the interest or locations of followers.[10] Twitter data can also be explored via an enterprise API platform, where services, such as making investment decisions or brand protection are offered.[11]

Data sources remain locked and useless without dedicated tools to navigate and discover the relationship between data. Particularly challenging are textual resources. Trivial matching of keywords may not be sufficient; therefore, a more sophisticated approach is needed. The content can be intelligently searched using the linked data approach, referred to as a semantic search. The semantic search goes beyond the mere textual representation of the content. In order to facilitate the creation of analytical products Barnaghi et al. (2013) proposed a value chain for processing models and learning mechanisms. (1) preprocessing of raw data, e.g., aggregation, summarization, filtering; (2) annotation and metadata modeling, e.g., data representation frameworks, languages; (3) data abstraction and pattern recognition; (4) semantic interpretation; (5) analytical processing; and (6) information visualization. The outcomes of the analysis are useless unless they can be incorporated into a complex decision making system (Shah et al., 2012). The outcomes should empower actions that can provide value (cf. *execution gap* in Sect. 4.2.2).

[9] https://enterprise.foursquare.com/solutions/analytics.

[10] https://analytics.twitter.com/about.

[11] https://developer.twitter.com/en/enterprise.

### 6.6.2 *Data-Driven Innovation*

Data-driven innovation (DDI) is an "innovation that creates real business value and stems from data processing and analysis" (Deloitte, 2016). It can be described as a cycle formed by a sequence of activities from data generation to decision making, as depicted in Fig. 6.13. This process is not a value chain, but the value creation process is repeated in many iterations, providing an opportunity for feedback. We assume that the possessed data forms a critical mass so that conclusions from the analysis are valuable. The following phases are distinguished within data-driven innovation (OECD, 2015):

1. Datafication and data collection—data is generated through the digitization of content, monitoring of activities, user provision, etc.
2. Big data—it is a state resulting from the first phase. Significant volumes of data are gathered and ready to be exploited through data analytics.
3. Data analytics—data is processed and interpreted. Nowadays this process is usually carried out in a cloud. It depends on infrastructure (software) but also on data science skills.
4. The knowledge base—the knowledge is accumulated over time through learning. For this purpose various machine learning algorithms are used. Knowledge is built and managed on enterprise level.
5. Data-driven decision making—value of data is first captured when data analytics returns insights. Here the value is captured from decision making. It contributes to value-added growth and well-being.

Companies use analytics not only for their own internal datasets but increasingly also for datasets extended with external data. Data-driven services are usually distributed and there are some complex interdependencies; therefore, assigning

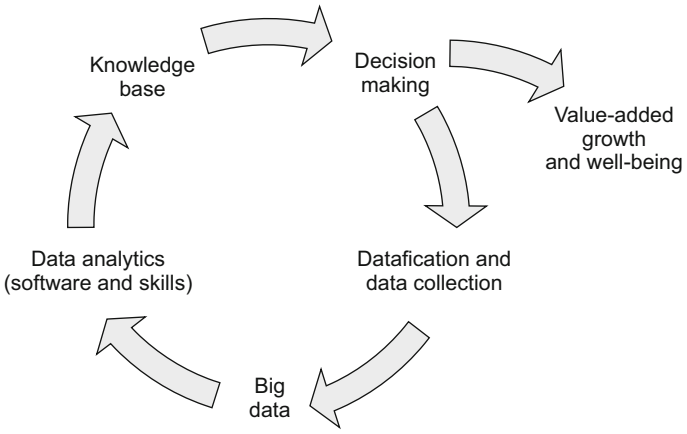



**Fig. 6.13** The data value cycle. Source: own work based on (OECD, 2015)

value to data. Data on a specific subject can be gathered from different sources, even for different purposes, and then combined in various ways to create value. The combination is not only used for visualization in a form of a mashup. Some businesses, particularly infomediaries, may decide to develop data-driven services based on existing services with APIs.

### *6.6.3 Open Innovation*

Continuing with Arrow's information paradox, we can show the motivation for open innovation. It also answers another question why people should care to share information and why they can profit anyway. Arrow (1962) examined incentives to innovation for monopolistic and competitive markets—profits of the invention were compared with the cost. He provided equations explaining that the monopolist's incentive is less than the inventor's incentive under competition. Moreover, the pre-invention monopoly power is a strong disincentive to further innovation.

The term "open innovation" refers to the "use of purposive inflows and outflows of knowledge to accelerate internal innovation, and expand the markets for external use of innovation" (OECD, 2015). It is very often related to such terms as licensing, collaboration, joint venture, etc. Open innovation has two interpretations, i.e., there are two models how innovation can be organized:

1. Market for intellectual property—where companies trade patents and other assets.
2. Distributed innovation system—individuals from around the world participate in an innovative processes of companies.

The first interpretation is quite straightforward—companies open their intellectual property and exchange ideas. One of the examples is the patent pool, presented in Sect. 6.2.3. It also refers to Arrow's information paradox, i.e., part of the information has to be revealed in order to be sold. The main issue here is what to reveal. As King and Lakhani (2013) concluded, managers must understand what to open, how to open it, and how to manage the resulting problems in order to benefit from open innovation. More details about these issues are contained in Chap. 7 devoted to business models.

The second interpretation is closer to the spirit of open movement. Instead of generating ideas in-house, companies can take advantage of ideas created using external expertise. Unlike firm-centered innovation, open innovation is decentralized and based on social motives (Lakhani et al., 2013). Researchers explored how a community could inform and shape a company but also how the company shaped and leveraged its community. Such exchange of ideas has to be institutionalized in some way (O'Mahony & Lakhani, 2011). Lakhani et al. (2013) claimed that there was no theory of the firm that took into account community innovation.

A community can be involved only in open idea generation, only in idea selection, or in both processes. The combination of these choices yields four

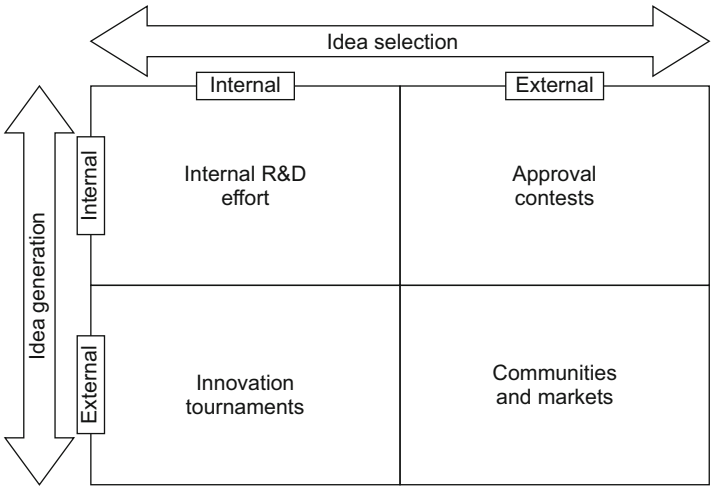



**Fig. 6.14** Various approaches to innovation. Source: own work based on (King & Lakhani, 2013)

different approaches to innovation, presented in Fig. 6.14. Typically, only the first process is followed and companies are usually not aware of the benefits of opening the second process. 'Innovation tournaments' are thus quite popular compared to 'communities and markets.' Companies rarely decide to open idea selection for internally generated ideas. The openness of the tournament can be graded as well. It can be completely open, meaning that competitors can see each others solutions. This requires a special design to allow or even encourage *coopetition* between tournament rivals. The latter strategy was followed by Netflix, which organized a contest to find the best algorithm for recommending movies to its customers.

There are several advantages related to open innovation. For example, a big number of ideas that become available results in a bigger chance for a high-quality idea. Moreover, the variability of the ideas received increases the value of the best idea. The advantages stem also from distinctive expertise and new perspectives. The open innovation can be particularly beneficial when it involves products that can be used in many ways, i.e., in different contexts. We have already presented data as a product, so open innovation can be applied to data and other intangible products as well. Nevertheless, one of the success factors is carefully designed process-oriented knowledge management (Gronau & Vladova, 2012).

Another group of advantages of open innovation concerns cost cutting and reduction of risk. First of all, a company knows for what ideas it pays. Moreover, it pays only after the design is completed. Therefore, there is no problem as with buying the idea in a traditional business, where the Arrow's information paradox can emerge. The other side of the coin is that the idea generator bears the cost of developing a solution. Therefore, incentives need to be offered so that external innovators are willing to join. Participants can be provided with dedicated tools or

offered access to a platform normally used by a contest organizer. In the case of data-related contest, it can be access to a sample of proprietary data, in the form of API, query framework or datasets to download, or access to organizer's computing infrastructure. The latter can lead to the creation of two-sided platforms.

The collaboration platforms were pioneered by IT sector and enable thousands of developers to create own applications working on a common infrastructure. Developed applications are then marketed by platform operators. Developers are satisfied because they get access to a standardized platform, can leverage the sales channel, and reach a broader market. Platform owners also profit from the relation—the higher is the turnover, the higher is the income from commission. The most successful companies are Google with its Google Play, an official application store for the Android operating system, and Apple with its App Store, a digital distribution platform for mobile applications on iOS operating system. Microsoft was less successful as it had problems with attracting developers. The number of people using a particular operating system translates directly to potential profits. Two-sided platforms are one of the examples how a company can make profit from open innovation. The platform operators do indeed change the focus of business from selling *products* to selling *access to a platform*.

One of the doubts was if the firm-based innovations could be done outside company. O'Mahony and Lakhani (2011) pointed at low-cost communication that facilitated the emergence of *self-organizing communities*, which are as effective as firms. Lakhani et al. (2013) emphasized that communities contained highly socialized people who were able to work collectively and come up with a solution to innovative problems.

Another issue—how open the company should be—was raised by Balka et al. (2014). In innovation development, companies need to consider a trade-off between openness to external value creation and closedness for internal value capture. They proposed the so-called 'selective openness' in software development. Companies should identify needs for openness within their community and decide on opening some modules and keeping essential modules closed. Selective openness can be beneficial both for the company and individual innovators. When carefully designed, the company can take advantage of both worlds: to increase the value of the system by openness and to capture the value from the system by closedness.

Open innovation creates several additional effects besides those involving entities mentioned above, which are usually directly involved. These are spillovers, defined as positive externalities, i.e., affecting additional actors beyond those pursuing the activity. For example, R&D investment of one company may open new opportunities for the whole branch. Other companies can benefit as well without any investment. Additionally, open innovation can also catalyze traditional intra-firm innovation. Open data also enables spillover effects.

There are some successful examples of applying open innovation in business practice, especially in high-tech companies. For example, IBM and its competitors (including Toshiba and Samsung) joined forces to develop integrated circuit fabrication. As a prerequisite, all companies released their intellectual property rights to the other members (cf. a patent pool). The consortium contained both

idea generators and selectors. This agreement between competitors seemed to be counterintuitive. Nevertheless, this cooperation showed that earning is possible in spite of competition. In fact, the competition between companies moves to another level: from competition based on a superior production process to competition based on better design of products. The whole industry developed thanks to the shared production capabilities they jointly developed (King & Lakhani, 2013).

The initiative taking advantage of open data is Open Data Incubator Europe.[12] The project develops an open data ecosystem, which shall generate economic, social and environmental impact. Open data is understood as a support for a rapidly developing market for innovative business ideas. As of October 2021, there were 57 startups and SMEs involved.

There are several incarnations of open innovation concept. For example, 'open science' is a term describing a movement that promotes greater transparency in the data collected and the scientific methodology used. It also advocates that data and tools should be available for reuse. Research results, particularly supported with public money, should be broadly disseminated. Open access refers to the possibility of accessing scientific literature and data online, for free, and free of licensing restrictions. This term is also applied to data provided by companies as part of their business model. Open access enables, supports, and encourages value-creating activities by users, leading to consumer-driven innovation. Open access is an optimal strategy for organizations "when they recognize that users may be best positioned to create value" (Frischmann, 2013).

The positive aspects of open innovation approach are widely discussed, but enterprises, particularly smaller ones, encounter various challenges and negative consequences. Some of them include costs of coordination or difficulties in finding a partner (Enkel et al., 2009). Problematic is also a knowledge spillover. Veer et al. (2012), based on a survey among 3956 German companies, found that companies engaged in open innovation were more exposed to imitation. The risk was lower in the last phase—testing and marketing. A surprising result was that all potential innovation partner types but competitors could expose to imitation risk. Nevertheless, the authors suggest to carefully design their open innovation strategy which does not reveal their core competencies neither to competitors nor to partners. Bigger risk was perceived in partners (predominantly suppliers) who can eventually become competitors, e.g., by vertical diversification.

Estimation of benefits and evaluation of risk is particularly challenging to SMEs, which usually have a limited budget for such activities. Ullrich and Vladova (2016) proposed framework and a software tool for assessment of open innovation project participation. The framework included an internal, external, and integrated analysis as well as a recommendation and decision phase. It was part of Open Darkness

[12] ODINE—Horizon 2020 Research and Innovation Programme, Grant Agreement 644683, https://opendataincubator.eu/.

project,[13] which aims at weighing the risks and benefits of open innovation participation.

## 6.7 Summary

Within the chapter, the following analogies were identified. First, sharing is a prerequisite for the success of two-sided markets. Second, linked data resembles value networks, which explain how value is generated. It is beneficial to integrate available linked data with non-open data that resides within enterprises. Internet is not only an efficient tool for data acquisition, but it is also a key platform for many emerging, innovative applications that can transform data into valuable services for businesses and consumers. The following scenarios for measuring value of data should be considered.

The first scenario considers *acquisition of data*. The common example is the integration of external data and combination with internal data. Value can be measured based on: (a) cost saving—use available data instead of investing in gathering or preparation; (b) new opportunities—offer existing resources to new customers or sell on new markets. Data that is to be integrated need not be open—similar effects can be observed in paid sources, but then the justification for investment must be more explicit.

The second scenario is *publication of data*. Change of value can be observed if data is made open. Data can be offered to other cooperating partners or shared with the community. There are different means of opening: standards (receipts), datasets (content), and applications (source code). Incentives for opening are related to specific business models. The main question that has to be answered is if the value is lost when data is made open.

The third scenario is *exchange of data*. Value is generated by the creation of the partner networks. They can be focused around commonly accepted standards, e.g., product catalogs, customer base. Growth of value is achieved when information is shared. Common standards should help in common maintenance of data.

There are benefits and positive externalities from disclosed data as well as costs and negative externalities. If patents can be shared in a patent pool, then why not to share data. In the chapter, we have provided a few arguments showing that a society can be more effective than a single company. The main message is that earning is possible in spite of competition. One of the most prominent examples is open innovation. As Tim Berners-Lee pointed out, "we need diversity of thought in the world to face the new challenges." Ideas can be more valuable than data. Further research should thus focus on how to develop valuable open data business models to foster data.

[13] OpenDarkness—Risk Assessment for Open Innovation, http://opendarkness.de/.

# References


Abramova, O., Shavanova, T., Fuhrer, A., Krasnova, H., & Buxmann, P. (2015). Understanding the sharing economy: The role of response to negative reviews in the peer to-peer accommodation sharing network. In *ECIS 2015 completed research papers*. ISBN: 9783000502842. (page 156)

Acquisti, A. (2010). The economics of personal data and the economics of privacy. In *The economics of personal data and privacy: 30 years after the OECD privacy guidelines* (pp. 1–24). Paris: OECD. (pages 163, 164, 166, 167)

Acquisti, A., & Varian, H. R. (2005). Conditioning prices on purchase history. *Marketing Science, 24*(3), 1–15. (page 167)

Adner, R. (2006). Match your innovation strategy to your innovation ecosystem. *Harvard Business Review, 84*(4), 98. (page 152)

Akerlof, G. A. (1970). The market for 'Lemons': Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics, 84*(3), 488–500. Issn: 0033-5533. https://doi.org/10.2307/1879431 (page 152)

Abramova, O., Shavanova, T., Fuhrer, A., Krasnova, H., & Buxmann, P. (2014). Designing a second generation of open data platforms: Integrating open data and social media. In *Electronic Government: 13th IFIP WG 8.5 International Conference, EGOV 2014, Dublin, Ireland, September 1–3, 2014. Proceedings* (pp. 230–241). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-44426-9_19 (page 151)

Allee, V. (2008). Value network analysis and value conversion of tangible and intangible assets. *Journal of Intellectual Capital, 9*(1), 5–24. ISSN: 1469-1930. https://doi.org/10.1108/14691930810845777 (pages 155, 156)

Andrienko, G., Gkoulalas-Divanis, A., Gruteser, M., Kopp, C., Liebig, T., & Rechert, K. (2013). Report from Dagstuhl: The liberation of mobile location data and its implications for privacy research. *SIGMOBILE Mobile Computing and Communications Review, 17*(2), 7–18. ISSN: 1559-1662. https://doi.org/10.1145/2505395.2505398 (page 166)

Arrow, K. (1962). Economic welfare and the allocation of resources for invention. In *The rate and direction of inventive activity: Economic and social factors* (pp. 609–626). Princeton University Press. ISBN: 0-87014-304-2. (pages 148, 156, 170)

Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. *Government Information Quarterly, 32*(4), 399–418. (page 146)

Baldwin, C. (2014). *Liverpool John Moores University launches big data healthcare initiative*. http://www.computerweekly.com/news/2240217724/Liverpool-John-Moores-University-launches-big-data-for-healthcare-initiative (visited on 2017-10-08). (page 158)

Balka, K., Raasch, C., & Herstatt, C. (2014). The effect of selective openness on value creation in user innovation communities. *Journal of Product Innovation Management, 31*(2), 392–407. ISSN: 0737-6782. https://doi.org/10.1111/jpim.12102 (pages 154, 155, 172)

Barnaghi, P., Sheth, A., & Henson, C. (2013). From data to actionable knowledge: Big data challenges in the Web of Things. *IEEE Intelligent Systems, 28*(6), 6–11. ISSN: 1541–1672. https://doi.org/10.1109/MIS.2013.142 (page 168)

Berners-Lee, T. (2017). *Three challenges for the web, according to its inventor*. http://webfoundation.org/2017/03/web-turns-28-letter/ (visited on 2017-04-18). (page 159)

Bolen, A. (2017). *Can data sharing help cure cancer? Collaborative analytics reveals hidden answers in clinical trial data*. https://www.sas.com/en_us/insights/articles/big-data/can-data-sharing-help-cure-cancer.html (visited on 2017-09-07). (page 157)

Brynjolfsson, E., Hitt, L. M., & Kim, H. H. (2011). Strength in numbers: How does data-driven decision-making affect firm performance? *SSRN Electronic Journal*. ISSN: 1556-5068. https://doi.org/10.2139/ssrn.1819486 (page 152)

Choice (2017). *Nine hours of 'conditions apply'*. https://www.choice.com.au/about-us/media-releases/2017/march/nine-hours-of-conditions-apply (visited on 2017-10-18). (page 159)

Chui, M., Farrell, D., & Jackson, K. (2014). How government can promote open data and help unleash over 3 million$ in economic value. In *Innovation in local government. open data and information technology* (pp. 4–23). McKinsey & Company. (pages 142, 143)

Croll, A. (2011). *Who owns your data*. http://news.yahoo.com/owns-data-20110112-030058-029.html (visited on 2017-05-27). (page 161)

Deloitte. (2013). *Market assessment of public sector information*. BIS/13/743. London: Department for Business Innovation & Skills. (pages 143, 144, 145)

Deloitte. (2016). *The value of DDI (data driven innovation)*. (pages 143, 169)

Dumbill, E. (Ed.). (2012). *Planning for big data. A CIO's handbook to the changing data landscape* (p. 84). O'Reilly Media. (page 150)

Edelmann, N., Höchtl, J., & Sachs, M. (2012). Collaboration for open innovation processes in public administrations. In Y. Charalabidis & S. Koussouris (Eds.), *Empowering open and collaborative governance* (pp. 21–37). Springer Berlin Heidelberg. ISBN: 978-36-422721-9-6. https://doi.org/10.1007/978-3-642-27219-6_2 (pages 146, 147, 152, 153)

Enabling the Data Revolution. (2015). Conference Report. Open Data for Development. (page 161)

Enkel, E., Gassmann, O., & Chesbrough, H. (2009). Open R&D and open innovation: Exploring the phenomenon. *R & D Management, 39*(4), 311–316. ISSN: 1467-9310. https://doi.org/10.1111/j.1467-9310.2009.00570.x. arXiv: 00178012 (page 173)

Esmeijer, J., Bakker, T., & de Munck, S. (2013). *Thriving and surviving in a data-driven society*. TNO. (page 150)

EU. (1995). Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal of the European Union, L 281*, 31–50. (page 163)

Eurostat. (2013). *Eurostat consumer price index from internet price data*. https://statswiki.unece.org/display/bigdata/Eurostat+-+Consumer+Price+Index+from+internet+price+data (visited on 2017-10-18). (page 161)

Frischmann, B. M. (2013). *Infrastructure: The social value of shared resources* (p. 436). Oxford University Press, ISBN: 978-0199975501. (pages 151, 173)

Gibson, J. (2009). *Intellectual property, medicine and health: Current debates*. Ashgate Publishing. (page 151)

Glick, B. (2014). *NHS England faces growing pressure to delay Care. Data medical records plan*. http://www.computerweekly.com/news/2240214577/NHS-England-faces-growing-pressure-to-delay-Caredata-medical-records-plan (visited on 2017-10-08). (page 158)

Gonzalez-Zapata, F., & Heeks, R. (2015). The multiple meanings of open government data: Understanding different stakeholders and their perspectives. *Government Information Quarterly, 32*(4), 441–452. ISSN: 0740-624X. https://doi.org/10.1016/jgiq.2015.09.001 (pages 139, 140, 141, 143, 145)

Goodwin, T. (2015). *The battle is for the customer interface*. Techcrunch. https://techcrunch.com/2015/03/03/in-the-age-of-disintermediation-the-battle-is-all-for-the-customer-interface/ (visited on 2017-10-18). (page 158)

Grant, A., Razdan, R., Shang, T. (2014). Coordinates for change: How GIS technology and geospatial analytics can improve city services. In *Innovation in local government. Open data and information technology* (pp. 32–43). McKinsey&Company. (page 163)

Gronau, N., Vladova, G. (2012). Wissensmanagement im Innovationsprozess. In A. Braun, E. Eppinger, G. Vladova, S. Adelhelm (Eds.), *Open Innovation in Life Sciences: Konzepte und Methoden offener Innovation-sprozesse im Pharma-Mittelstand* (pp. 99–120). Wiesbaden: Gabler Verlag. ISBN: 978-3-8349-7105-0. https://doi.org/10.1007/978-3-8349-7105-0_6 (page 171)

Hansen, M. T., & Birkinshaw, J. M. (2007). The innovation value chain. *Harvard Business Review, 85*(6), 121–130. ISSN: 0017-8012. (page 167)

Hermalin, B. E., & Katz, M. L. (2006). Privacy, property rights and efficiency: The economics of privacy as secrecy. *Quantitative Marketing and Economics, 4*(3), 209–239. (page 163)

IFB. (2016). *Crash for cash*. https://www.insurancefraudbureau.org/insurance-fraud/crash-for-cash/ (visited on 2017-09-07). (page 157)

Janssen, M., & van den Hoven, J. (2015). Big and open linked data (BOLD) in government: A challenge to transparency and privacy? *Government Information Quarterly, 32*(4), 363–368. (page 163)

Jetzek, T., Avital, M., & Bjørn-Andersen, N. (2012). The value of open government data: A strategic analysis framework. In *SIG eGovernment pre-ICIS workshop 2012*. Orlando. (pages 154, 156)

Jetzek, T., Avital, M., & Bjorn-Andersen, N. (2013). The generative mechanisms of open government data. In *Proc. of the 21st European Conference on Information Systems*. Utrecht: AIS. (page 154)

King, A., & Lakhani, K. R. (2013). Using open innovation to identify the best ideas. *MIT Sloan Management Review, 55*(1), 41–48. (pages 170, 171, 173)

Klenow, P. J., Rodriguez-Clare, A. (2005). Externalities and growth. In P. Aghion, S. N. Durlauf (Eds.), *Handbook of economic growth* (Vol. 1A, Chap. 11, pp. 817–861). Elsevier B.V. https://doi.org/10.1016/S1574-0684(05)01011-7 (page 156)

Krasnova, H., Eling, N., Abramova, O., & Buxmann, P. (2014). Dangers of 'Facebook login' for mobile apps: Is there a price tag for social information? In *Proc. of International Conference on Information Systems (ICIS 2014)*. Auckland, New Zealand. (page 166)

Lakhani, K. R., Lifshitz-Assaf, H., & Tushman, M. L. (2013). Open innovation and organizational boundaries: Task decomposition, knowledge distribution and the locus of innovation. In A. Grandori (Ed.), *Handbook of economic organization: Integrating economic and organization theoryic and organization theory* (Chap. 19, pp. 355–382). Northampton, MA: Edward Elgar Publishing. ISBN: 178254822X. https://doi.org/10.2139/ssrn.1980118 (pages 170, 172)

Lakomaa, E., & Kallberg, J. (2013). Open data as a foundation for innovation: The enabling effect of free public sector information for entrepreneurs. *IEEE Access, 1*, 558–563. ISSN: 2169–3536. https://doi.org/10.1109/ACCESS.2013.2279164 (page 148)

Landry, J. N., Webster, K., Wylie, B., & Robinson, P. (2016). *How can we improve urban resilience with open data?* Open Data for Development. (page 158)

Li, N., Li, T., & Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering* (pp. 106–115). https://doi.org/10.1109/ICDE.2007.367856 (pages 164, 165)

Loshin, D. (2001). *Enterprise knowledge management: The data quality approach* (p. 493). Morgan Kaufmann. ISBN: 978-0124558403. (page 160)

Ma, Y. (2016). The use of advanced transportation monitoring data for official statistics. PhD thesis. Erasmus Research Institute of Management. (page 161)

Manovich, L. (2012). Trending: The promises and the challenges of big social data. In *Debates in the digital humanities* (pp. 460–475). University of Minnesota Press. https://doi.org/10.5749/minnesota/9780816677948.003.0047. arXiv: arXiv:1011.1669v3 (page 164)

Manyika, J., Chui, M., Groves, P., Farrell, D., Van Kuiken, S., & Doshi, E. A. (2013). *Open data: Unlocking innovation and performance with liquid information*. McKinsey Global Institute. (page 159)

Marshall, C. C., & Shipman, F. M. (2017). Who owns the social web? *Communications of the ACM, 60*(5), 52–61. https://doi.org/10.1145/2996181 (pages 158, 159)

Morgan, L., Feller, J., & Finnegan, P. (2010). Value creation and capture with open source software: A theoretical model for understanding the role of value networks. In: *Proc. of the 18th European Conference on Information Systems (ECIS)*. (page 156)

Naef, E., Muelbert, P., Raza, S., Frederick, R., Kendall, J., & Gupta, N. (2014b). *Using mobile data for development*. Cartesian and Bill & Melinda Gates Foundation. (pages 166, 167)

Nejad, N. M., Scerri, S., Auer, S., & Sibarani, E. M. (2016). EULAide: Interpretation of end-user license agreements using ontology-based information extraction. In *Proceedings of the 12th International Conference on Semantic Systems* (pp. 73–80). ACM. (page 159)

Nielsen, J. (2006). *The 90-9-1 rule for participation inequality in social media and online communities*. Nielsen Norman Group. https://www.nngroup.com/articles/participation-inequality/ (visited on 2017-03- 22). (page 146)

Noam, E. M. (1997). Privacy and self-regulation: markets for electronic privacy. In *Privacy and Self-Regulation in the Information Age*. (page 167)

O'Mahony, S., Lakhani, K. R. (2011). Organizations in the shadow of communities. *Research in the Sociology of Organizations: Communities and Organizations, 33*, 3–36. ISSN: 0733-558X. https://doi.org/10.1108/S0733-558X(2011)0000033004 (pages 170, 172)

ODINE. (2017). *Who really owns data?* https://opendataincubator.eu/who-really-owns-data/ (visited on 2017-09-10). (page 159)

OECD. (2013b). Exploring the economics of personal data: A survey of methodologies for measuring monetary value. *OECD Digital Economy Papers, 220*, 40. https://doi.org/10.1787/5k486qtxldmq-en (page 166)

OECD. (2015). *Data-driven innovation. Big data for growth and well-being* (pp. 1–456). Paris. ISBN: 9789264229358. https://doi.org/10.1787/9789264229358-en (pages 150, 151, 157, 159, 160, 164, 169, 170)

Pentland, A. (2013). The data-driven society. *Scientific American, 309*(4), 78–83. (page 159)

Pentland, A. (2015). *Who should we trust to manage our data?* World Economic Forum. https://www.weforum.org/agenda/2015/10/who-should-we-trust-manage-our-data/ (visited on 2017-04-20). (page 163)

Pisano, G. P., & Verganti, R. (2008). Which kind of collaboration is right for you? *Harvard Business Review, 86*(12), 78–86. ISSN: 00178012. (page 153)

Poikola, A., Kola, P., & Hintikka, K. A. (2010). *Public data* (p. 80). Helsinki, Finland: Ministry of Transport and Communications. ISBN: 978-952-243-238-4. (page 140)

Posner, R. A. (1981). The economic of privacy. *The American Economic Review, 71*(2), 405–409. (page 163)

Reuters. (2011). *Financial times pulls its apps from Apple store*. https://www.reuters.com/article/us-apple-ft/financial-times-pulls-its-apps-from-apple-store-idUSTRE77U1O020110831 (visited on 2017-12-08). (page 157)

Roca, T., & Letouzé, E. (2016). *Open algorithms: A new paradigm for using private data for social good*. Devex. https://www.devex.com/news/open-algorithms-a-new-paradigm-for-using-private-data-for-social-good-88434 (visited on 2017-04-20). (page 162)

SAS. (2017). *Why it's so easy to steal from insurance companies—and what to do about it*. https://www.sas.com/en_us/insights/articles/risk-fraud/easy-to-steal-from-insurers.html (visited on 2017-09-10). (page 157)

Shah, S., Horne, A., & Capellá, J. (2012). Good data won't guarantee good decisions. *Harvard Business Review, 90*(4). (page 168)

Stigler, G. J. (1980). An Introduction to privacy in economics and politics. *The Journal of Legal Studies, 9*(4), 623–644. (page 163)

Taylor, C. R. (2003). Privacy in competitive markets. *SSRN Electronic Journal* 1–25. ISSN: 1556-5068. https://doi.org/10.2139/ssrn.419720 (page 163)

Tucker, C. (2010). The economics value of online customer data. In *The economics of personal data and privacy: 30 years after the OECD privacy guidelines* (pp. 1–23). Paris: OECD. (page 165)

Turck, M. (2017). *The new gold rush? Wall street wants your data*. http://mattturck.com/the-new-gold-rush-wall-street-wants-yourdata/ (visited on 2017-09-09). (pages 150, 167, 168)

Turner, M. (2016). *An unlikely source predicted Chipotle's disastrous quarter, and it says a lot about the future of investing*. http://www.businessinsider.com/foursquare-data-predicted-chipotle-results-2016-4 (visited on 2017-10-20). (page 167)

Ubaldi, B. (2013). Open government data—towards empirical analysis of open government data initiatives. *OECD Working Papers on Public Governance, 22*, 61. ISSN: 1993-4351. https://doi.org/10.1787/5k46bj4f03s7-en (page 141)

Ullrich, A., & Vladova, G. (2016). Weighing the pros and cons of engaging in open innovation. *Technology Innovation Management Review, 6*(4), 34–40. ISSN: 1927-0321. (page 173)

van Es, M. (2014). *The networked economy: Meeting the challenges*. https://www.linkedin.com/pulse/20140916133902-31672864-the-networked-economy-meeting-the-challenges (visited on 2017-09-07). (page 158)

Varian, H. R. (2014). Beyond big data. *Business Economics, 49*(1), 27–31. ISSN: 0007-666X. https://doi.org/10.1057/be.2014.1 (page 168)

Veer, T. H., Lorenz, A., & Blind, K. (2012). How open is too open? The 'dark side' of openness along the innovation value chain. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2148399 (page 173)

West, J., & O'Mahony, S. (2008). The role of participation architecture in growing sponsored open source communities. *Industry and Innovation, 15*(2), 145–168. (page 155)

Wilkerson, J., Abdallah, K., Hugh-Jones, C., Curt, G., Rothenberg, M., Simantov, R., Murphy, M., Morrell, J., Beetsch, J., Sargent, D. J., Scher, H. I., Lebowitz, P., Simon, R., Stein. W. D., Bates, S. E., & Fojo, T. (2017). Estimation of tumour regression and growth rates during treatment in patients with advanced prostate cancer: a retrospective analysis. *The Lancet Oncology, 18*(1), 143–154. ISSN: 1470-2045. https://doi.org/10.1016/s1470-2045(16)30633-7 (page 157)

Wilson, K., & Brownstein, J. (2009). Early detection of disease outbreaks using the Internet. *Canadian Medical Association Journal, 180*(8), 829. (page 167)

Woo, B. (2013). *A mind blowing big data experience: Notes from strata 2013*. Forbes. https://www.forbes.com/sites/bwoo/2013/02/27/amind-blowing-big-data-experience-notes-from-strata-2013 (visited on 2017-09-10). (page 152)

Yan, A., & Luo, Y. (2016). *International joint ventures: Theory and practice*. Routledge. (page 151)

Zafeiropoulou, A. M., O'Hara, K., Millard, D., & Webber, C. (2012). Location data and privacy: A framework for analysis. In B. Stiegler (Ed.), *Réseaux Sociaux: Culture Politique et Ingénierie des Réseaux Sociaux* (pp. 185–200). FYP éditions. (page 166)

Zagheni, E., Weber, I., & Gummadi, K. (2017). Leveraging Facebook's advertising platform to monitor stocks of migrants. *Population and Development Review*. https://doi.org/10.1111/padr.12102 (page 161)

Zuiderwijk, A., Janssen, M., & Davis, C. (2014). Innovation with open data: Essential elements of open data ecosystems. *Information Polity, 19*(1–2), 17–33. ISSN: 18758754. https://doi.org/10.3233/IP-140329 (pages 140, 141, 142)

Zuiderwijk, A., Janssen, M., & Dwivedi, Y. K. (2015). Acceptance and use predictors of open data technologies: Drawing upon the unified theory of acceptance and use of technology. *Government Information Quarterly, 32*(4), 429–440. ISSN: 0740-624X. https://doi.org/10.1016/j.giq.2015.09.005 (pages 139, 140)

# Chapter 7
# Business Models for Data


*A mediocre technology pursued within a great business model may be more valuable than a great technology exploited via a mediocre business model (Chesbrough, 2010)*

## 7.1 Introduction

The concept of a business model has already received much attention in the literature on management and strategy. Reducing to a single function, the business model should answer the question how the company makes money. The increase in the number of publications was caused mainly by development of the Internet. Besides the generic articles, there are also some domain-specific efforts to address specific areas of business modeling. Particularly interesting is addressing a challenge related to the exploitation of a business potential of innovations whenever a new idea or technology emerges. One of such ideas is linked data, a part of the semantic web technology. Although it is already 10 years old, it has not yet marked clear commercialization paths. No wonder that as linked data is gaining importance, talking about dedicated business models is also becoming increasingly popular.

Any technology has no single value unless it is commercialized. The same technology commercialized in two different ways can yield different economic outcomes (Chesbrough, 2010). Company's environment has a vital influence on the kinds of business models that can be invented and implemented to offer value in a given market context. Sometimes an innovation fits into an existing business model of a company, but breakthrough discoveries affecting whole industries necessitate the design of completely new business models. Even though the change seems non-disruptive, companies should not be blinded by the established market practices. A potentially valuable use of technology can be missed if it does not fit the current business model of the company.

A given technology rarely works independently from other technologies. An interoperability and integration with other technologies is necessary in order to create value (Baden-Fuller & Haefliger, 2013). This is particularly important when many technologies are already in place in the company and the number of



K. Węcel, *Big, Open and Linked Data*, Business Information Systems,
https://doi.org/10.1007/978-3-031-07147-8_7

connections seems to follow the Metcalfe's law (Shapiro & Varian, 1998). It should not be assumed that adding just another technology will lead to a steady increase of revenues. Companies should have in mind the "moderating influence of business model choice." In other words, this choice determines the value-generation potential of the technology (Baden-Fuller & Haefliger, 2013).

Linked data is not just a technology. It is inherent to open data, a philosophy to make government data available to all interested parties. Thus, it is a whole ecosystem, offering new opportunities for data reuse. Pellegrini et al. (2013) is referring to the entire spectrum of linked data assets: from instance data through ontologies to technology. Open data is a movement raising interest for its potential to improve the delivery of public services by changing how the government works. It can also empower citizens and create added value for enterprises. Further potential can be released by applying advanced analytics to combined proprietary and open knowledge.

So far, linked data is embraced mostly by governments. Although linked data is not yet well settled in managers' minds, it is a good time to discuss certain issues. For the majority of companies, the most important is *how income can be generated*. Value proposition is vital for linked data, because the resulting revenue stream can support further development.

In this chapter, we carry out the analysis of business models and their components that can be applied for commercialization of linked data. There is a vast generic literature on business models that tries to homogenize definitions and come up with a single one. This is not the goal of this chapter and we just refer to other research for details where necessary. Literature about linked data business models is still modest; therefore, we assume a broader scope of our survey—we believe that business models developed for data assets and the Web can find their applications in linked data.

## 7.2 Digital Disruption and Social Business Transformation

Traditional models of management and value creation are no longer valid in a competitive economy that we observe today. There are new needs that can be fulfilled by social business.[1] It usually starts with software and processes but still needs support from semantics and potentially linked data (Abramowicz et al., 2011). Sica (2014) observed the following discontinuities that companies are faced with:

- Turbulent and disruptive economy—competitive positions of various organizations are changed in a high speed.

[1] http://socialbusinessmanifesto.com/.

- Social and empowered customers—consumers use the power of the Internet, e.g., opinions shaping the purchase behavior.
- Service economy—product economies are transformed into service economies; weaker contact of companies with customers, replaced with bigger players that provide the front-end to customers (emergence of brokers), e.g., Booking.com—no more direct contact with a hotel is necessary before the stay.

There is a strong belief between enterprises that revolutionary and disruptive business models are now possible thanks to "real-time digital connections across people, businesses, and devices."[2] In the light of the potential benefits of big, open, and linked data there is a strong need for research on business models for these emerging resources.

## 7.3 Business Models Research

In this section, we present the general research related to business models. We start with overview papers, then present several interesting frameworks, and conclude with the evolution of business models.

### *7.3.1 Definition*

There are some simple definitions of a business model found in the literature. Chesbrough and Rosenblom (2002) defined a business model as "a blueprint for the way a business creates and captures value from new services or products.", and according to Osterwalder and Pigneur (2010), "a business model describes the rationale of how an organization creates, delivers and captures value."

There are several interesting publications that provide extensive surveys on business models (Chatterjee, 2013; Kinnari, 2013; Krcmar et al., 2011; Lindgren & Rasmussen, 2013; Malone et al., 2006; Novak, 2014; Teece, 2010; Weill et al., 2005; Weiner et al., 2010; Wirtz et al., 2015). The review of the literature reveals that in general, researchers do not agree on what business model is (Shafer et al., 2005; Zott et al., 2010). Not only the proposed frameworks differ but also explicit statements on disagreement can be found. Zott et al. (2010) claimed that although a business model was emerging as a new unit of analysis, "literature is developing largely in silos." Several trends, however, can be delineated regarding the focus of definition. Value is intrinsic to business models. Some authors focus on how revenue is generated, i.e., what are sources of value (Bekkelund, 2011; Pellegrini et al., 2013; Plé et al., 2010; Vafopoulos, 2011; Zott et al., 2010). Other group takes

[2] Vivek Bapat, SAP's global vice president for portfolio and strategic marketing (MIT Technology Review Custom, 2014).

customers' viewpoint and defines business models as a description of how value is created, i.e., what is valuable for customers (Ahmadi Zeleti et al., 2014; Casadesus-Masanell & Ricart, 2010; Krcmar et al., 2011). Sometimes a business model, as an architecture of revenue, is referred to as a business ecosystem (Zott et al., 2011). Finally, a business model can be a means to explain how an organization works (Krcmar et al., 2011).

The last should not be confused with strategy. Indeed, the phrase 'logic of the firm' can refer to both. Many authors explicitly distinguish a business model from a strategy (Casadesus-Masanell & Ricart, 2010; Magretta, 2002; Shafer et al., 2005). Casadesus-Masanell and Ricart (2010) argues that "a business model is a reflection of the firm's realized strategy." At the same time, a business model is also separated from a tactic.

Certain works focused on value creation as the most important topic. Amit and Zott (2001) explored the theoretical foundations of value creation in e-business. They developed *eValue* framework, a model of the sources of value creation, which are: efficiency, complementarities, lock-in, and novelty. This approach was later extended in (Zott & Amit, 2010), where a business model was conceptualized as a system of interdependent activities within a company. Chesbrough (2012) perceived business model through the functions that it should realize: to create value and to capture value. The first function refers to a production process where products and services are delivered to customers through value-adding activities. The second function has to assure a unique position of the company in order to achieve the competitive advantage.

Weiner et al. (2010) provided a very appealing visualization of research on business models. Their "Business Model Research Map" was prepared in analogy to a tube map. Another broad and comprehensive overview of business models is presented in (Krcmar et al., 2011). They analyzed many perspectives ranging from definitions through taxonomies to software-supported design. Trends and a research agenda are also discussed. Lindgren and Rasmussen (2013) provided a very detailed analysis of business model components and dimensions in Appendix 2 (also the table on page 144). Boons and Lüdeke-Freund (2013) reviewed the current literature on business models in the context of technological, organizational, and social innovation. They also provided a draft for a research agenda in the form of research questions. Several publications are devoted to the classification of business models by providing various taxonomies (Afuah & Tucci, 2000; Gassmann et al., 2013; Linder & Cantrell, 2000; Rappa, 2010; Timmers, 1998; Weill & Vitale, 2001).

### *7.3.2 Frameworks*

One of the best recognized frameworks is Business Model Canvas (Osterwalder, 2004; Osterwalder & Pigneur, 2010). It not only defines the components of a business model but also provides guidelines how it should be invented. The nine areas of a business model are as follows: key partnership, key activities, key

resources, value proposition, relationship with customers, customers, channels, revenue stream, and cost structures. Boons and Lüdeke-Freund (2013) combined models by Osterwalder (2004) and Doganova and Eyquem-Renault (2009) to come up with four elements model: value proposition, supply chain, customer interface, and financial model.

Hedman and Kalling (2003) proposed to extend the scope of a business model definition by introducing additional components. Several levels in value chain were distinguished: market, offering, activity, and resource. The conceptual business model consisted of the following components, assigned to different levels: (1) customers, (2) competitors, (3) offering, (4) activities and organization, (5) resources, (6) supply of factors and production inputs. Shafer et al. (2005) proposed framework based on four major categories, generalized from 42 business model components found in the literature: strategic choices, value creation, value network, and capture value.

Various frameworks for business model analysis add new dimensions, not found in other works. For example, Seppänen (2009) focused on resources for business models claiming that they were crucial for value creation. He defined the following top-level categories: human, organizational, informational, physical, financial, legal, and relational resources. Mason and Spring (2011) introduced network architecture, apparently for the first time, as one of the core elements of business models, supplemented by technology and market offerings. Changes observed in the music industry were the pretext to discuss the evolution of business models.

Lindgren and Rasmussen (2013) claimed that the majority of publications about business models were just conceptual and not grounded in empirical research. Their Business Model Cube[3] was introduced as a generic and empirically tested framework, consisting of seven generic dimensions: value proposition (VP), customers and users (CU), value chain functions (VC), competence (C), network (N), relations (R), and value formula (VF). Schallmo (2013) also proposed another set of dimensions, which are: financial (costs, sales), customer (channels, relations, segments), partner (channels, relations), value creation (resources, processes, capabilities), and usefulness (services, benefit). He also analyzed B2B business models by looking at diverse 'generic providers': market, finance, sustainability, product, service, people, and network providers.

Schief and Buxmann (2012) introduced a business model concept comprising 20 elements. Later, a very detailed analysis was carried out and presented in (Schief, 2014). The author has prepared a business model reference database with over 1000 software business models. In the book, he analyzed the characteristics of business models and their impact on a company's performance. He also offers a Business Model Wizard supporting the configuration and management of business models.

Gassmann et al. (2013) analyzed 250 business models and generalized them into a catalog of 55 patterns of business models. This is another approach to help in the

[3] Not to be confused with Linked Data Business Cube (Pellegrini et al., 2013).

identification of the appropriate business model. They defined who, what, how, and value of business models.

What was discussed until now can be called a horizontal decomposition of business models, i.e., no hierarchical relations were introduced. It is reasonable to consider also the vertical decomposition. Indeed, Casadesus-Masanell and Ricart (2010) observed that business models could be viewed from different levels of detail. This topic was also discussed by Wirtz et al. (2010), who proposed four *business model levels*: industry level, corporate (company) level, business unit level, and product level. Schallmo and Brecht (2010) added a fifth level: an abstract level. These five levels can thus facilitate analyzes of business models in big companies with a portfolio of diverse products. It is also relevant from the point of view of this work—division into levels can be important for large organizations that utilize linked data to support some of their processes but do not build the whole strategy on this (e.g. Wolters Kluwer, BBC).

### *7.3.3 Business Model Innovation and Evolution*

When a new paradigm emerges, like linked data, it necessitates changes in business models. New models are invented or old ones evolve. Chatterjee (2013) provided some clues on the first topic by elaborating the roadmap for business model development. In the analysis, two dimensions were considered: (1) *resource*—virtual/digital vs. physical and (2) *relationships*—few vs. many. This produces four types of generic business models as starting points for further development: efficiency-based business model, perceived value-based business model, loyalty-based (network value) business model, and network-efficiency.

The second topic, evolution, was covered, among others, by Magretta (2002), who stated that a business model should be adapted to changes in the environment, particularly to changes in a value creation process as well as changes in the technology. She perceived business modeling as the managerial equivalent of the scientific method—one starts with a hypothesis, then tests it in action, and revises when necessary. A good example for evolution was also provided by Mason and Spring (2011), who analyzed changes in the recorded music market since the 1870s.

The topic of business model innovation has also been undertaken by Chesbrough (2010). As a general conclusion, companies are characterized by inertia, i.e., business models change slowly because they are part of the bigger system and business processes must be adapted first. These findings can explain the slow adoption of linked data among enterprises. There are certainly many other factors, for example, reluctance to change, characteristic for public entities.

## 7.4 Literature Review Methodology

### *7.4.1 Research Objectives*

In previous Sects. 3.2 and 7.3 we have refined the definition of linked data as well as outlined the main concepts related to business models. The main research question that we now seek an answer for is: what are the business models that are or can be applied for linked data. As we expect that there is no single universal model to support all activities related to linked data, we also want to determine what components of business models can contribute to create new business models.

We pursues three targets:

1. Based on the in-depth structured review of literature on business models, to identify components of business models pertinent for linked data adoption or commercialization.
2. To analyze relevant business models in domains related to linked data, with emphasis on their sustainability.
3. To discover specific issues and potential further research areas concerning business models for linked data.

### *7.4.2 Data Collection and Search Process*

We have conducted a systematic structured literature review in the domain of business models for linked data in order to verify our research question. According to guidelines by Webster and Watson (2002) we have focused on relevant journals and conferences in the analyzed field. For this purpose we have used the ProQuest's ABI/INFORM Complete database, which indexes key business and economics periodicals. According to the publishers, it gives "a complete picture of companies and business trends around the world."

In the search process, we have combined several key phrases to narrow the selection of the articles to the most relevant ones. We were only interested in articles published in 2001 and later—this is the time the seminal paper about the Semantic Web was published. We tried full text search as well as search in selected fields, i.e., title, abstract, and keywords. Number of results for various search phrases is presented in Table 7.1. We were definitely interested in business models, hence its presence in all phrases. We first started with a full text search and tried to restrict a big number of articles by adding phrases related to linked data. The additional phrase 'linked data' was too specific, producing only 18 results, while 'information' and 'data' were too broad. When we switched to search in all fields except full text, the number of papers concerning business models for linked data dropped to two. Further modifications included restrictions to 'data or information' and 'Web or Internet' and extension of a business model to concepts related to value. Finally, as

**Table 7.1** Number of search result for various phrases

| Search phrase | # papers |
|---|---|
| *Full text* | |
| "business model" | 13,666 |
| "business model" AND "data" | 9846 |
| "business model" AND "information" | 11,409 |
| "business model" AND ("data" OR "information") | 12,157 |
| "business model" AND "linked data" | 18 |
| "business model" AND ("linked data" OR "semantic web") | 171 |
| *All fields except full text (title, abstract, keywords)* | |
| "business model*" | 5093 |
| "linked data" | 180 |
| "business model*" AND ("data" OR "information" OR "web") | 2069 |
| "business model*" AND ("data" OR "information") | 1954 |
| "business model*" AND ("data" OR "information") AND ("internet" OR "web") | 594 |
| "business model" AND ("linked data" OR "semantic web") | 2 |
| ("business model*" OR "value creation" OR "value generation" OR "value proposition") AND ("data" OR "information") | 2889 |
| ("business model*" OR "value creation" OR "value generation" OR "value proposition") AND ("data" OR "information") AND ("internet" OR "web") | 403 |
| ("business model*" OR "value creation" OR "value generation" OR "value proposition") AND ("data" OR "information") AND ("semantic*" OR "ontolog*" OR "vocabular*") | 40 |

Restriction on publication date: after 2001-01-01, only peer-reviewed publications
Search results as of 2016-02-27

**Table 7.2** Search criteria

| | |
|---|---|
| Database | Proquest ABI/INFORM complete |
| Search fields | Title, abstract, keywords |
| Period | After 2001-01-01 |
| Search query | All(("business model*") AND ("typology" OR "classification" OR "taxonomy" OR "design") AND ("internet" OR "web")) |
| Number of results | 198 |

we were interested in business model classification and design, these phrases finally shaped our search query. Final criteria used for the search are presented in Table 7.2.

This search resulted in 198 papers. We have then analyzed their abstracts and decided if the paper was relevant, i.e., contained a business model or value design process that could be adapted to linked data. At this stage we have excluded papers not dealing with information, e.g., about fiber infrastructure. We have also rejected papers focusing on detailed cases of single companies, e.g., Haier in China, as they did not provide the expected level of abstraction. This filtering lead to the collection of 33 papers that were then subject to detailed content analysis as advised by (Krippendorff, 2012).

**Table 7.3** Research methods used in the analyzed papers

| Research method | # papers |
|---|---|
| Action research | 1 |
| Analysis | 6 |
| Case study | 7 |
| Design science | 3 |
| Grounded theory | 1 |
| Interviews | 2 |
| Literature review | 5 |
| Modeling | 3 |
| Multi case study | 2 |
| Survey | 1 |
| *No method* | 4 |

**Table 7.4** Industries targeted by business models

| Industry | # papers |
|---|---|
| Mobile | 5 |
| E-business | 4 |
| Music | 3 |
| Media/publishing | 3 |
| Internet | 2 |
| Brokerage | 1 |
| Telemedicine | 1 |
| Education | 1 |
| In-memory computing | 1 |
| Internet carriers | 1 |
| Internet of things | 1 |
| *Generic* | 10 |

**Table 7.5** Business model elements

| Element | # papers |
|---|---|
| Business model definition | 8 |
| Business model proposition | 9 |
| Business model classification | 11 |
| Business model design | 7 |

During manual scanning of each paper, major findings have also been summarized. In order to structure the presentation, we have defined several categories, by which the analyzed papers were classified. We were interested in the following features: what research methods were used (cf. Table 7.3); what industry was covered by the paper (cf. Table 7.4); if the paper defined, proposed, or classified business models and if the design process was provided (cf. Table 7.5). After reading and annotation of the papers, only 14 were selected for inclusion in the detailed analysis, i.e., they contained germane findings.

For the literature concerning business models the collected corpus was extended by going backward and forward in citations. The journal *Long Range Planning*

turned out to be the source of many papers concerning strategy and business models; therefore, we also scanned tables of contents for additional relevant papers. The same period was assumed.

After the structured literature analysis based on Proquest ABI/Inform, we came to the conclusion that the literature about business models for linked data is scanty. Therefore, broader search of the Web was carried out. It lead to several less renowned publications (e.g., blog discussions), which in turn lead to several other interesting publications by analysis of forward and backward citations.

### *7.4.3 Results Overview*

Detailed description of the findings from the literature review is contained in the next two sections, structured in a way reflecting the main categories covered by various authors. This structuring should also allow easy reference to linked data. Sect. 7.5 analyzes business model components and Sect. 7.6 covers business models from related domains. Such an approach allows us to come up with a complete picture of the domain of interest. In Table 7.6, we present only the main findings of the selected papers, which can, by analogy, be applied in linked data domain.

## 7.5 Analysis of Business Model Components

Based on the analysis of business models carried out earlier, we identified their crucial elements, which provide a canvas for this section (Chesbrough & Rosenblom, 2002; Osterwalder & Pigneur, 2010; Peters et al., 2015; Shafer et al., 2005). Value creation and value capture are the two common denominators in most of the works. Many authors have observed the necessity to transfer value from producers to consumers, including interaction and communication channels. This phenomenon is called differently in diverse works: value delivery, distribution, communication, or just transfer. We have decided to combine them under the umbrella of the last concept. These three elements are used to structure the literature analysis in this section.

### *7.5.1 Value Creation*

In this subsection, we analyze how value is created and what is valuable for customers. In the context of linked data, value is bound to contents. This can be created by dedicated companies in formal processes with a purpose or by individuals and communities in a less co-ordinated approach (referred to as crowdsourcing). Data can also be provided by government.

**Table 7.6** Main contributions of papers with potential application in linked data *by analogy*

| Publication | Findings and inspirations |
|---|---|
| Katz and Shapiro (1994) | Main source of value of the linked data can be derived from the network effects: massive collaboration between users and creating bidirectional links between resources |
| Parker and Van Alstyne (2005) | Two-sided network externalities: "increased demand in a complementary premium-goods market more than covers the cost of investment in the free-goods market," i.e., linked data offered for free to earn elsewhere |
| Chang (2006) | Two-sided markets theory can provide hints how to make linked data profitable; solutions based on open access can be sustainable by: saving costs, increasing incomes, adopting innovative technology, controlling the quality of publication |
| De Reuver et al. (2009) | A viable business model should create both *customer value* and *network value* |
| Hougaard and Tvede (2010) | "No business model selling digital goods directly to customers is profitable"; optimal business models for selling non-excludable goods (e.g. knowledge, information, linked data), suggests alternative ways of making profit |
| Wirtz et al. (2010) | The perceived value of offered content can be increased by adding customization and personalization |
| Dadzie and Rowe (2011) | Linked data consumption is difficult for non-technical users, increase of value of liked data can be achieved by proper visualization |
| Zhang et al. (2012) | User-generated content as one of critical factors for competitive advantage (co-creation); the notion *value chain* should be replaced with *value network* |
| Hagen et al. (2013) | Linked data will not exist as a standalone technology |
| Hart and Dolbear (2013) | Freemium model is of the most interest for geographic information publishers (linked geodata) |
| Pellegrini et al. (2013) | Primary value proposition of linked data is to generate network effects at the data level; software as one of the linked data assets |
| Bärenfänger et al. (2014) | Technical features can be transformed into business benefits, in-memory computing as en example |

Source: own work

The first freely available data appeared when Web 2.0 was introduced. People got the tools to voluntarily provide data, starting the phenomenon of crowdsourcing (Howe, 2006). Actually, a collaborative production of contents by network communities preceded the open data movement. Wikipedia, the biggest open encyclopedia, was even analyzed in the context of economics and the term Wikinomics was coined (Tapscott & Williams, 2006). Quality of freely available content is usually questioned. Wikipedia is different as it implements special mechanisms for assessing the quality and there are methods for automatic assessment of article quality in various languages (Lewoniewski et al., 2016, 2017; Węcel & Lewoniewski, 2015).

One of the examples provided by Wirtz et al. (2010) also concerned the case of Wikipedia. Its business model is built on the idea of voluntary and cooperative value generation with users. The project is fully depending on the user-added value. It is also interesting how many people contribute not only content but trust—this is one of the phenomena of social networking and a visible example of network externalities. The case of Wikipedia is also interesting from a linked data point of view. The biggest source of linked data and the center of linked data cloud—DBpedia—derives its content from Wikipedia. Whatever has to be done with Wikipedia's data, it can be done effectively with DBpedia using SPARQL queries.

The user-generated content was identified by Zhang et al. (2012) as one of 14 critical success factors for achieving competitive advantage by companies. Moreover, user content generation was deemed as one of the next generation trends in e-business.

In more formal approaches, value can be created in a value chain. Taking into account the complex relations between various entities, it is more appropriate to speak about co-creation instead of creation. Organizations in a value network undertake different activities playing diverse roles.

According to Zwass (2010), "co-creation is the participation of consumers along with producers in the creation of value in the marketplace." The same author presents a typology of co-created value, with two classes at the top level: autonomous co-creation (activities by consumers) and sponsored co-creation (consumers and producers). The first class—autonomous co-creation—concerns consumer-side production and includes:[4] production of procedural contents (software); production of declarative contents (e.g., knowledge compendia,[5] consumer reviews, mash-ups[6]); development of social capital (e.g. web-based relationships on Facebook); trust creation (e.g. buyer and seller feedback on eBay); collective sense-making ('folksonomies'); collective ranking of importance; collective sentiment expression. The second class, sponsored co-creation is about contribution to the producer value chain. From the described examples, none can provide real-world examples for the linked data domain.

Similar ideas concerning value creation in the music market are promoted by Hougaard and Tvede (2010), who suggest to encourage fan participation. However, in their approach co-creation is a strategy to sell non-excludable goods, like digital audio.

User-generated content is not the only vehicle for value. Value can also be derived from technical characteristics. For example, Bärenfänger et al. (2014) discussed the business value of in-memory computing (IMC) technology. They analyzed its organizational impact based on various implemented application scenarios. Similarly, Westerlund et al. (2014) postulated to perceive Internet of Things (IoT) not as a technology platform but as a business ecosystem.

[4] Only relevant classification has been chosen, e.g., hardware co-creation has been skipped.

[5] Wikipedia as the most prominent example.

[6] usu. Google Maps as a base.

### 7.5.2 *Value Transfer*

A value transfer specifies how the value is distributed to goods and service consumers. It includes interaction and communication channels. In the context of linked data, there is an interesting remark about the role of the Internet. Porter (2001) observed that "better technology does not necessarily lead to profitability." By introduction of new technology, the question should be asked who will capture the benefits of this technology. There are always various types of entities interested in the distribution. Zwass (2010) identified economic beneficiaries: the world, community, sponsoring firm, aggregators, and contributors. He also mentioned who are performers, what are task characteristics, source of motivation, and governance of creation. Matt and Hess (2012) analyzed how the value between suppliers and intermediaries is split in the context of the digital music industry. It applies to other markets where the content is distributed in an electronic way. For Peters et al. (2015), value co-creation is important for services whose value creation depends on the integration of the consumer into the service provision.

One of the most important phenomena described in the analyzed literature is *free-riding problem* for digital goods. It is covered by several papers, particularly referring to the publishing and music industry (Amberg & Schröder, 2007; Denning, 2014; Hougaard & Tvede, 2010; Mason & Spring, 2011; Matt & Hess, 2012; Parker & Van Alstyne, 2005).

Parker and Van Alstyne (2005) observed that it is not necessarily harmful for a company to give away a good for free. The reason is that on two-sided markets there are bounds between various products—in the freemium model one group gets the product for free in order to attract the other user group. In other words, "increased demand in a complementary premium-goods market more than covers the cost of investment in the free-goods market." On such markets both content providers and consumers can offer a free good and they provide each other with network benefits. Parker and Van Alstyne (2005) also proved that in the case of low marginal costs the company can provide products for arbitrarily large market, which is the case of information and by inference also linked data.

Hougaard and Tvede (2010) demonstrated that no business model selling digital goods directly to customers is profitable. They analyzed the music market on the assumption that music is a non-excludable good. The paper provided a taxonomy of attempts to fight the free-riding problems (i.e., piracy). *Open source* was mentioned as one of the possibilities. In this case, the producer makes the main product freely available, focusing on related products and services to create profit. Two further models can also be applied for linked data. First, *bundling with private good*, where the non-excludable good is bundled with an excludable good in order to reduce free riding. Second, the *eyeball model*, where producer sells access to a good rather than selling the good itself, thus changing the primary non-excludable good into a secondary good. For the music industry, two models have been recommended: the business model should be a combination of the 'eyeball model' and the 'open source model.'

The music industry is not the only one providing analogies for data markets. Chang (2006) analyzed open access publishing models. Concluding, the two-sided markets theory can provide hints how to make linked data profitable.

Besides free-riding, there is another phenomenon that decreases revenues and forces changes in business models—shared access. It can be considered a special case for value distribution—distribution in time as an artifact is passed from one person to another. Although sharing concerns physical goods, some analogies can be drawn for digital goods as well. Denning (2014) asked a rhetorical question why one should buy expensive and quickly becoming outdated encyclopedias (e.g., Britannica), when a much larger and frequently updated Wikipedia is available for free. In the past, people used a printed statistical yearbook to obtain relevant data. Nowadays we can ask a similar question: why anyone would like to buy the yearbook, when it is more convenient to use linked open data. Denning (2014) also noted that sharing creates significant positive externalities: experience, learning, and relationships.

### *7.5.3 Value Capture*

A company is only sustainable by generating revenue. It is then important how the company positions itself in the value chain (Rappa, 2010). Value capture describes the way in which companies make profit within a value ecosystems (Zott et al., 2011). The company should charge for what the customers value the most, based on the revenues they generate (Pine II & Gilmore, 2016).

De Reuver et al. (2009) proposed to measure business model performance based on meta-theory of business model design. They explained the value captured by the organization by examining the impact of a company design on the success factors. A viable business model should create both *customer value* (how value is created) and *network value* (how the company makes money). There are approaches, particularly in e-business, that combine a revenue model with a customer value model (Sahut et al., 2013). From twelve 'value models for customers' presented there, three are relevant for linked data: content—satisfaction of information needs; search—targeting the needed information; entertainment—providing specific information to a particular area of interest.

For the value capture, it has to be distinguished what kind of market is targeted. Leem et al. (2004) showed that mobile business models are different from Internet business models and that distinction should be done between B2C and B2B/B2E.[7] The B2C models focus on value proposition for various needs of individuals. B2B are divided according to the scope of enterprise value chain and their main purpose is cost reduction of the enterprise.

It is important to note that not all created value is meaningful from the commercialization point of view (Westerlund et al., 2014). Then, before value is

[7] Business to employee.

captured, it has to be extracted first. From another perspective, the customers should pay for the effects. There are several ways how the value can be captured (Weill & Vitale, 2001): payments for transactions; for information and advice; for services and commissions; advertisement-generated income and payments for referrals.

To some extent, open data can be compared to open access publishing. Chang (2006) analyzed value capture in publishing. Two business models have been compared: traditional publishing and open access. The main difference was a source of financial support, e.g., print version subscriptions, advertising, licensing content. In the traditional model readers pay, while in the open access model authors are required to cover the costs of publication. Analogous situation is in the case of open data—here the payer is usually the government (in fact, tax payers).

Chang (2006) also specified four means of how solutions based on open access can be sustainable: saving costs (decreasing various expenditures); increasing income (collecting additional fees); adopting innovative technologies (utilization of new technology to make the publication sustainable); controlling the quality of the publication (people are willing to pay for better quality).

## 7.6 Analysis of Relevant Business Models

By simplifying the definitions of linked data discussed in Sect. 3.2 we can say that linked data is *data* available over the *Web*. Many of the business models identified in the structured literature analysis can be assigned to these categories. The extended search provided additional business models for linked data available. These three domains—data, the Web, linked data—were used to structure this section.

### *7.6.1 Business Models for Data Assets*

The value of internal data for a success of an enterprise is unquestioned (Fisher, 2009). Using new nomenclature, this traditional asset is now called closed data, in contrast to open data. For the closed data model, data is a value proposition according to Business Model Canvas. A good overview of data-driven business models is presented in (Hartmann et al., 2014).

The key activities for data-driven business models are collection and maintenance of data. These are costly activities—require both people and equipment. Open Data Institute (2015) identified strategic risks for businesses that charge for data delivery. People are less willing to pay for contents that can be easily copied and shared with others, e.g., music, films (Amberg & Schröder, 2007; Hougaard & Tvede, 2010; Matt & Hess, 2012). Competitors can disrupt the business by offering as open data the same data scope. Additionally, it is important to guarantee business sustainability so that other companies can build own solutions based on the offered data. Security of business can only be assured by achieving a critical mass of activities.

The growing volume of open data, considered as the main source for linked data, finally motivated questions about economic value. Cost arguments were also raised by government bodies. When the attention of open data community shifted to the economic value of open data assets, a growing number of open data business models appeared in the literature. Unfortunately, rigorous research concerning the models is still very limited (Ahmadi Zeleti et al., 2014).

Sheridan and Tennison (2011) proposed a very simple categorization into two types of business models: for organizations that publish but do not sell open data and organizations built on top of using open data. There are further model types distinguished:

- Freemium—entry level data is free and publishers charge for added value.
- Cross-subsidy—companies get extra benefit for their data, data is a key resource.
- Network effects—taking value from collaboration in a data-rich environment.

More sophisticated approaches usually recognized more types of data-related business models. For example, Tennison (2012) identified seven types of open data business models: cost avoidance, sponsorship, dual licensing, support and services, charging for changes, increasing quality through participation, and supporting primary business. Howard (2013) mentioned the following eight open data business models: premium, freemium, open source, infrastructural razor&blade, demand-oriented platform, supply-oriented platform, free as branded advertising, and white-label development. The same models, called *archetypal business models*, are enumerated and defined by Ferro and Osella (2013).

It is easily observed that above-mentioned 15 models are not coherently defined. Sometimes they combine concepts from various levels. Some of the models do not even deserve the name "business model." For example, *cost avoidance* is a tactics for cost saving rather than a fully fledged business model. For a practical application and analysis, there are too many models. Although various activities are described within each model, the economic effect is in many cases equivalent. Therefore, some efforts were observed to define common categories and reduce the number of models or at least present them in a more convenient way.

One of such proposals can be found in (Ahmadi Zeleti et al., 2014) who consolidated the elements of different business models and identified six core elements. The resulting framework is called 6-V and includes: value proposition, value adding process, value network, value return, value capture, and value management. Central to the framework is value proposition including product, services, distribution channel, information, and prices. The 6-V framework was used to analyze the mentioned 15 open data business models, which ultimately were grouped into five major categories: freemium, premium, cost saving, indirect benefit, and the razor-blade. The analysis of particular value propositions resulted additionally in four types of value disciplines for open data businesses: usefulness, process improvement, performance, and customer loyalty. Overall, the framework is logical and easy to understand.

Other business models dedicated for open data were introduced by Janssen and Zuiderwijk (2014). The so-called *infomediary business models* are located between

**Table 7.7** Business models for open data

| Publication | Business models |
|---|---|
| Sheridan and Tennison (2011) | Freemium, cross-subsidy, network effects (3) |
| Tennison (2012) | Cost avoidance, sponsorship, dual licensing, support and services, charging for changes, increasing quality through participation, supporting primary business (7) |
| Howard (2013) | Premium, freemium, open source, infrastructural razor&blade, demand-oriented platform, supply-oriented platform, free as branded advertising, white-label development (8) |
| Ahmadi Zeleti et al. (2014) | Freemium, premium, cost saving, indirect benefit, the razor-blade (5) |
| Janssen and Zuiderwijk (2014) | Intermediary models: single-purpose applications, interactive applications, information aggregators, comparison models, open data repositories, service platforms (6) |

Source: own work

open data providers and consumers. Six types of them were identified: single-purpose applications, interactive applications, information aggregators, comparison models, open data repositories, and service platforms. One can observe that they focus rather on the form of interaction and consider social media. Hybrid business models can be created, combining the features of the models mentioned above. Summary of business models for open data is contained in Table 7.7.

Some lessons concerning data assets can be learned from big data (Hagen et al., 2013; Hitzler & Janowicz, 2013; Muhtaroglu et al., 2013). Big data applications are deployed in different sectors. Core elements of the business are studied from the business model canvas perspective. The paper presents how various applications deliver value to customers and how profit is made of using big data. It is important to mention that competitive advantage can be gained by companies which can adapt their businesses to leverage big data. Business models should be reshaped accordingly. As a summary, we can repeat theses by Hagen et al. (2013), concerning big data: "big data is not likely to exist as a standalone technology for long" and "the distinction between big and traditional data is disappearing." It is highly probable that the same statements would actually apply to *linked data*.

### *7.6.2 Business Models and the Web*

Web standards are at the foundation of linked data; therefore, analysis of this domain can reveal interesting business models. Linked data can be interpreted as an enhanced version of the Web we know.

The Internet has been free since its inception. People got used to this elementary value and now it is very hard to change the habits. In the early years of the Internet, in the 1990s, companies were doing business without a business model (Shapiro

& Varian, 1998). Only after dot-com bubble, people learned that they needed to guarantee the revenue stream and the "business model" became one of the most important buzzwords at the time (Magretta, 2002). The "free" business models are still very popular among the biggest content players on the Internet, like Google or Adobe (Bryce et al., 2011). Taking a closer look, people pay anyway, but the currency is attention and privacy.

In the early work on business models for Internet, Timmers (1998) provided a framework for the classification of Internet electronic commerce business models. As the criterion, he assumed a degree of innovation and a functional integration. The following eleven models have been distinguished: e-shop, e-procurement, e-auction, e-mall, trust services, info brokerage, value chain service provider, virtual community, collaboration platform, third party marketplace, and value chain integrator. Main sources of revenue were subscription and pay-per-use payments, whereas advertising was found a less popular model. The closest model to the domain of this paper is *information brokerage*, with such examples as business opportunities brokerage and investment advice.

Weill and Vitale (2001) claimed to provide the first systematic and practical analysis of e-business models. They defined *e-business model* as the model describing the way a company does business electronically. The following viable business models have been specified: (1) content provider, (2) direct to consumer, (3) full-service provider, (4) intermediary, (5) shared infrastructure, (6) value net integrator, (7) virtual community and (8) whole of enterprise/government. They are the so-called atomic e-business models.

Wirtz et al. (2010) proposed 4C Internet business model typology. Four basic types of prototypical Internet business models are distinguished: content—collecting, aggregating, distributing, or presenting online content; commerce—fulfilling online transactions; context—organizing available online information; connection—providing virtual or physical infrastructure. Within 4C, *user-added value* is the most important trend for content-oriented companies. User-generated content is valuable both on input (sourcing) as well as on output (innovation). The perceived value of the offered content can be increased by adding customization and personalization.

Wirtz et al. (2010) also contributed to the evolution of business models by showing the differential effect of environmental changes on each type of internet business model. They argue that web trends still disrupt the effectiveness of established internet business models, thus forcing companies to continuous change. Unfortunately, the overall findings of this paper are too general. It refers to changes in the environment, but the changes are perceived mostly on the demand side (e.g., customer need), not necessarily technological capabilities, i.e., the growing relevance of user-added value and interaction orientation.

Rappa (2010) introduced a comprehensive taxonomy of business models observed on the Web: brokerage, advertising, infomediary, merchant, manufacturer (direct), affiliate, community, subscription, utility. Some of the models above are then accepted for linked data. Matt and Hess (2012) researched the music industry, where the content was distributed in an electronic way. More specifically, they

analyzed the competition between suppliers and intermediaries. They found that consumers were especially responsive to changes in price and product assortment.

Data is very often made available over the Web by the application programming interface (API), usually implemented as a web service. Such software components can be modified and configured to work with other components. Additionally, the software market is influenced by trends and developments in the area of data and information management: Web 2.0, semantic web, big data, and linked data to name a few. Therefore, we also include in our analysis business models in the software industry, initially studied by (Popp, 2011). He presented a series of matrices constructed as the intersection of two dimensions: business model archetypes (borrowed from Weill et al., 2005) and types of goods and services (financial, physical, intangible, human). When linked data is considered, the intangible good is actually offered. Software is becoming crucial in serving linked data and some authors already treat it as one of the linked data assets (Pellegrini et al., 2013).

When speaking about business models on the Web, one should not ignore the network effects. Here, collaboration is essential—effort can be distributed while at the same time benefits can be retained. The most prominent examples of collaboration platforms are Wikipedia, OpenStreetMap, and MusicBrainz. The network effects should be taken into account when deciding about the proper choice of technology. For example, Zynga, an online games provider, has developed their business by leveraging Facebook, which offered connections between friends (viral marketing) and advertisements (Baden-Fuller & Haefliger, 2013). It was using a well-know platform, without the need for heavy marketing investments. Zynga offers games for free and still earns from a small number of paying customers.

### *7.6.3 Business Models for Linked Data*

Design issues for linked data were published in July 2006, but only around 2009 the economic issues started to gain momentum. Latif et al. (2009) first came up with a lightweight business model for this domain—Linked Data Value Chain. The model defined four roles: raw data provider, linked data provider, linked data application provider and end-user. Linked data is not directly consumed but transformed into human-readable data at the end of the chain. The idea was demonstrated in a case study involving BBC.

In 2010, the discussion on business models in general was already mature, when several proposals for business models for linked data appeared (Brinker, 2010a, 2010b, 2010c; Dodds, 2010; Erickson, 2010; Groth, 2010; Hellman, 2010). Initially, there were mostly various blog posts and ideas were collected in the discussions, without scientific rigor. Recommendations were based on practice rather than research. Research approaches appeared later and embraced both systematization of the approaches as well as classification of the existing business models.

Brinker was the first to come up with eight business models, which he classified by how revenue is generated (Brinker, 2010a, 2010c):

- subsidized—funded by government or a non-profit organization, the most direct revenue source;
- subscription—selling access to raw data; linked data facilitates access to data, standardized format makes reuse even more attractive;
- advertising—selling placements inside data feeds or advertising around data driven-applications;
- authority—charging for "authority stamps," e.g., official reviews, certifications, compliant services;
- affiliate links—e-commerce affiliate links embedded inside data feeds or data-driven applications;
- value-add—selling access to applications that creatively use data: enhanced on-demand or as a product;
- traffic—generating bigger traffic via linked open data networks, e.g., attracting search engines by using GoodRelations ontology;
- branding—positioning a company as valuable data provider, the most indirect revenue source.

Brinker believed that data branding would become one of the most significant forces behind the widespread adoption of linked data. In the second update, he extended his framework to three dimensions (who pays, who gets services, and how direct is the revenue) and proposed 15 business models (Brinker, 2010b).

Cobden et al. (2011), who introduced the notion of *linked closed data*, proposed to categorize business models by how their costs are recovered:

- advertising supported—costs are covered by revenue from advertisements within content;
- loss-leader—costs are to be recovered in the future;
- subsidized—costs are covered by a third party (no consumer nor producer).

Several extensions to Brinker's proposition can be found. Vafopoulos (2011) distinguished eleven business models that are a combination of the models described in (Brinker, 2010b) and (Dodds, 2010): public service, community service, subscriptions (including full, timely, on-demand, block, archival and convenient access plus freemium), customized service, sponsorship, advertising, marketplace, affiliate program, multi-sided platform, traffic generation, data branding. Vafopoulos (2011) also proposed a framework for linked data including two core value sources of linked data and critical functional components cited after (Stähler, 2002): value proposition, products/services, architecture, revenue model. It is not stated how the elaborated framework was applied to come up with the proposed categories of business models. As it is missing rigor, it is also less interesting from the research point of view.

Another more mature framework for linked data business models is Linked Data Business Cube (Pellegrini et al., 2013). Here three dimensions were defined (hence the name 'cube'): linked data assets, revenue models, and stakeholders. Revenue models are taken directly from (Brinker, 2010c) but their relevance is discussed with regard to other dimensions. The linked data assets have been divided into:

low-incentive assets (instance data, metadata) and high-incentive assets (contents, service, technology). Ontologies belong to both categories offering the widest scope of commercialization possibilities. Low-incentive assets can be commercialized using an indirect revenue model, i.e., branding, traffic, or commission. High-incentive assets are better suited to direct revenue models, i.e., subsidy, licensing, subscription, or advertising. There is also a third dimension in this model—stakeholders. It enumerates entities that can potentially apply the framework: internal stakeholders (for own purposes), partners, B2B, B2G, B2C, and B2Co.[8] The initial mapping of stakeholders to revenue models proposed by Pellegrini et al. (2013) helps enterprises in developing own business models as it reduces the number of combinations that have to be analyzed within Linked Data Business Cube.

Archer et al. (2013) discussed, from a business model point of view, how public organizations can provide their data to third parties. Although the scope is restricted to Linked Open Government Data (LOGD), conclusions are universally applicable to linked data as well. As the report points out, the dominating business model is the one where the costs of running linked data services are covered by public funding. The authors emphasized the following four value propositions of LOGD: flexible data integration, increase in data quality, creation of new services, and reduction of data integration costs.

The last item is a particular example of positive network effects—costs of resolving interoperability conflicts are born only once. Usually no other entity than just a public body is willing to cover these costs.

Certain kinds of information can be more useful for reuse in enterprises, thus creating bigger value. This is definitely true for the so-called reference data and geographic information is one of the good examples.

Geographic information has probably the biggest potential for linking datasets, as it is, beside time, the most popular dimension in statistical datasets. Hart and Dolbear (2013) tried to answer the question why geographic information publisher should invest in transforming its data into linked data. They observed that linked data was very often confused with *open* data. Collection of geographic information and maintaining its quality is a long-term investment, so no company would decide to give its datasets for free. Therefore, making data available needs a strong support from the revenue stream. Basing on (Brinker, 2010c) and (Dodds, 2010), they came up with examples for each of 15 business models. Only some of them are noteworthy: subsidy model—producing linked data for public benefit (e.g., GeoNames); internal savings—improving the precision of collected location-related data; branding—providing data free as a means of brand promotion; freemium—providing some data for free whereas enriched dataset is paid.

Even though branding and subsidy prevailed among publishers, according to Hart and Dolbear (2013), freemium model is of the most interest for geographic information publishers: old or overview maps can be given for free while accurate, detailed, and up-to-date maps can be sold for money.

[8] Business-to-community.

The freemium model has also been followed by Linked Life Data.[9] It is a service that offers various data from the life sciences domain concerning: genes, proteins, molecular interactions, drugs, diseases, and clinical trials. Altogether, there are 10 billion statements about 1.5 billion entities from 25 biomedical databases. Data can be accessed in a data-as-a-service manner by means of SPARQL queries. The service offers two different access levels: *LLD Public*, which is completely free and available for proof-of-concept purposes, *LLD Enterprise*, which is a premium service (subscription necessary) for real-world applications, with additional features. The extra features include: unlimited query time, higher update frequency, technical support, and access to sources with commercial licenses. Complex analytical queries usually require the enterprise license.

## 7.7 Discussion

We start a discussion with a simple question—"What should be first: technology or business model?" The former approach was commercialization of new technologies and the main question was "How to sell technology?" (Chesbrough & Rosenblom, 2002). The newer approach moves towards value creation and evaluation as reflected in the question "How to create value and for whom?" (Casadesus-Masanell and Ricart 2010). Technology has no single objective value and has to be commercialized via a business model to show the economic value. Still, the same technology can be commercialized in different ways, leading to various economic outcomes (Chesbrough, 2010).

From the technological point of view, linked data is not a breakthrough. Underlying specifications are mostly unchanged. What changes is the approach of users, who agreed to interpret certain facts in a specific way (semantics). Linked data is characterized by openness (undisturbed flow of data), modularity (identifiers managed locally), and scalability (can embrace web scale). It can be compared to a new programming language that expresses certain intentions in a more convenient way. Nevertheless, linked data is not only about representation—it is a whole ecosystem that is created when we change our approach to data reuse.

Companies need access to data from various sources and with different structures in order to carry out their business efficiently. Depending on how simple the access is, three levels are distinguished. The first level comprises internal, structured data with direct access. It can be enriched with external but still structured data, forming the second level. Integration consists in providing appropriate mappings. The least accessible data is on the third level, which is both unstructured and external. Leveraging this data requires the development of dedicated methods for information extraction. Linked data can usually be located on the second level and therefore it

[9] http://linkedlifedata.com/.

is very attractive for enterprises. Thanks to linked data, the integration step can be simplified or skipped at all. This is the added value derived from linked data.

The value proposition of linked data is not to be perceived only in *uniqueness* of data but also in a way this data is *made available*. Hence, *linked* open data can be priced higher than open data. Moreover, a service is more important than a static dataset without a query mechanism.

### *7.7.1 Study of Business Models*

For the analysis of various cases, very often the Business Model Canvas has been employed (Osterwalder & Pigneur, 2010). In the value proposition area, the most important observation is that little effort is currently spent on measuring the costs and benefits of linked data. As a result, the cost structure remains unknown and enterprises do not separately account for this. Increased data quality should be mentioned as a main value contribution. Key resources focus on the reference data, which is the most successfully applied. For revenue streams, the dominant model is public funding. As a consequence, data is usually provided free of charge and licenses are either open or not explicitly defined.

A relatively big number of proposed business models stem from empirical research. One can argue that there are as many business models as companies. Therefore, some authors take the challenge to find common characteristics of business models. The typical approach is to collect various cases and later try to generalize the findings by the introduction of the very own business model framework. The analysis of the literature suggests that it is more about the classification of business models rather than their potential to make profit or at least ensure sustainability. While technological challenges have been resolved in a bigger part, the economic viability of linked data initiatives is to be proven.

The profit achieved from the application of a given business model depends on many additional factors. If we took a random sample of brick&mortar enterprises and linked data companies, the distribution of revenue, income, and workforce would be similar and probably follow the power law. It is not about 'if' linked data is applied but 'how.' Linked data definitely is not the must-have-to-survive and the returns can be very unpredictable. It should be treated as a philosophy built into the enterprise architecture so that some processes become easier or are enabled.

New models are necessary, but it is not easy to build them 'on demand,' i.e., for a particular case, no matter what technology is involved. Mostly empirical and heuristic models are created for linked data (Brinker, 2010a; Wirtz et al., 2010) and the role of experimentation and evolution is emphasized. There are no convincing models besides some obvious ones, known from the publishing/media industry. Still, the dominating model for revenue is subsidy and interesting applications are mostly stemming from EU funded projects.

### 7.7.2 Real-World Applications

Cobden et al. (2011) pointed out the validation of business models as one of the important challenges. Potential for a success of a business model cannot be verified theoretically. Only the real-world company running a certain model can confirm its viability.

An interesting point of view on the Internet is presented by Porter (2001), who argued that "the Internet is not an industry but an enabling technology." It is not appropriate to speak about 'Internet companies' but companies using the Internet, which in turn has the power to influence the competitive forces in various established industries. Nevertheless, the changes are not necessarily disruptive. The impact is much bigger when a good is information. Porter supports the idea of business model division into levels. No company, except very small ones, makes linked data the only foundation for a business. Instead, relevant processes are supported.

Buyle (2014) researched real-world online open data applications. He surveyed communities of developers developing city applications using open data in Barcelona, Rome, Helsinki, Amsterdam, and Berlin. Findings were rather pessimistic: the application ecosystem failed as a whole, although single applications achieved social success. The author provided the following explanation: lack of managerial knowledge of application developers, so they were not aware of the possible business models supporting their applications. It is not sufficient to just open data—he postulated to invest in the ecosystem protecting various stakeholders and to train potential developers.

Overall, three scenarios for linked data usage in enterprises can be distinguished (Blumauer, 2014):

- internal purposes—enterprises use linked data to organize their own information assets; this is particularly advantageous when heterogeneous databases are in use;
- inbound purposes—organizations use external resources to supplement their own information; this trend is increasingly popular as also the number of open data sources is increasing;
- outbound purposes—organizations apply linked data principles to publish own data on the Web; good knowledge of licensing strategies is necessary.

### 7.7.3 Intellectual Property Issues

Many researchers put emphasis on intellectual property and licensing issues when writing about business models (Chesbrough, 2010; Pellegrini & Ermilov, 2013; Pellegrini et al., 2013; Tennison, 2012). Taking into account various data flow scenarios, it is necessary to master different kinds of licenses. The ability to combine these licenses is one of the most crucial competencies in developing linked data business models.

This is particularly important when a community of a B2Co model is considered. Various open source initiatives create value in a collaborative manner. Data collection very often requires human expertise, at least at the design stage of data acquisition. Additionally, communities by forming a critical mass can in fact decide about future standards adoption. Data quality remains the main concern in crowdsourcing, thus a vivid community has to be involved in order to keep up with the expectations of other stakeholders.

Publication of data becomes complex when many kinds of stakeholders are involved. It has to be clear what kind of licenses are applied and usually dual licenses are offered. As part of the license, a type of responsibility for data should be included. This is probably one of the reasons why provenance became so important (Dezani-Ciancaglini et al., 2012; Eckert, 2013; Hartig & Hartig, 2009).

### *7.7.4 Markets and Ecosystems*

A market is the place where a value proposition can be realized. It is an institution and as such should also be considered in business models. Both traditional and electronic markets serve three main functions (Bakos, 1998): (1) matching buyers and sellers, (2) facilitation of transactions and (3) providing institutional infrastructure. These functions are further divided into sub-functions. The matching function of the market is particularly interesting and relevant in the context of linked data, as it consists of the following sub-functions: determination of product offerings, searching and price discovery. The purpose of linked data is twofold: it can facilitate these functions or be a subject of exchange on the market.

In the first case, the role of linked data is similar to that of intermediaries, analyzed in (Giaglis et al., 2002). Linked data, like intermediaries, can provide a single point of contact for information gathering, both for buyers and sellers. It allows for monitoring, alerting, reducing search costs, and facilitating price discovery. Dedicated intelligence agents can be replaced with reasoning engines utilizing linked data.

In the second case, linked data is a digital good, similar to the publishing and music industry, where a specific kind of information is traded. Bakos (1998) distinguished two major trends characterizing electronic markets: increased personalization and aggregation of information-based product components. The first trend, personalization, is particularly visible for digital products, where relevant information can be delivered based on user profiles. Sometimes dedicated intermediaries, called infomediaries (Hagel & Rayport, 1997; Janssen & Zuiderwijk, 2014; Rappa, 2010; Tapscott et al., 2000), match buyer's needs to the sellers' offer. The second trend, aggregation, is likewise suitable for linked data. Infomediaries create value by aggregating *products* (data) and *services* (access to data), thus maintaining a product mix. Both trends can be handled simultaneously for linked data, taking into account it is susceptibility to transformation.

Taking into account the complex relations between various actors in value networks, some authors propose to consider not only the company level in business model development (Westerlund et al., 2014). They argued that business models should not be broken down into components. On the contrary, they proposed to extend the scope of business models to the whole business ecosystem and to speak about "value design." Not only a company but also the ecosystem can create and capture value. During value design for the ecosystem, the following value pillars should be considered (Westerlund et al., 2014): value drivers—individual and shared motivations; value nodes—various actors, activities, and processes; value exchanges—how value is exchanged and how revenues are generated and distributed; value extracts—the meaningful value that can be monetized. The related benefits can then be obtained at the higher level of the ecosystem.

## 7.8 Summary

Linked data is not yet widely adopted; therefore, it is hard to speak about the viability of existing business models for linked data. Nevertheless, we need to keep in mind the following analogy: "Linked Data is as essential for the Semantic Web as hypertext has been for the Web" (Latif et al., 2009). Hypertext, which by the way is about linking of pieces of information, was not commercialized, at least not alone. It can be used freely, as part of both non-profit and commercial solutions. There are currently millions of applications that have been built thanks to hypertext. Perhaps this is the reason there are only few works about just linked data business models, but the applications can flourish soon. Moreover, there are already similar domains that enclose linked data, for example data-driven and web business models can be applied as well.

It is apparently not feasible to build a business purely on linked data. Companies should rather focus on certain contributing parts—value creation, transfer, and capture processes supporting classical revenue models. Modular approach makes it possible to 'creatively combine' various elements as necessary, which is in line with the spirit of linked data itself. Combined can be both elements of models as well as business model patterns.

Concerning the research agenda for business models, there is an inspiring article by Veit et al. (2014), who proposed three research pillars: (1) business models in IT industries; (2) digital business models; (3) IT support for developing and managing business models. Business models for linked data are the special case of the first pillar. All sub-areas are relevant: identification of elements, classification, and performance implications of business models. Digital business models, the second pillar, are particularly interesting for research. They are defined as models that change when changes in digital technologies occur (Venkatraman, 1994). Data management, which is now reinforced with linked data, is definitely one of such technologies and changes in how companies carry out their business should follow. 'New product and service models' are the relevant sub-area. The last pillar on the

surface is not relevant; however, if we focus on the sub-area 'design of software tools for supporting business model development,' we then realize that linked data, along with big data, can be used to develop new or enhance existing business models, not necessarily concerning digital goods. Thus far we have considered 'business models for linked data' and the change towards 'linked data for business models' opens completely new and exciting opportunities.

The pillar 'digital business models' should be extended to include two issues that for linked data seem to be particularly important. Future research trends should then embrace privacy and intellectual property aspects because easy linking of information can be a curse if personal data is concerned. There are already certain regulations in place, but they are still subject to improvements.[10] Second issue, licensing of data, is much more complex and requires organizational changes within stakeholders offering and consuming data. The question is also if it is possible to motivate companies to share more data in a linked data format without rising concerns for losing a competitive advantage.

## References

Abramowicz, W., Koschmider, A., Stein, S., Węcel, K., Kaczmarek, M., & Filipowska, A. (2011). Social software and semantics for business process management—alternative or synergy? *Journal of Systems Integration 2*(3), 54–69. ISSN: 0922-6389. https://doi.org/10.20470/jsi.v2i3.95 (page 182)

Afuah, A., & Tucci, C. L. (2000). *Internet business models and strategies: Text and cases* (Vol. 2, p. 358). McGraw-Hill Higher Education. ISBN: 0072397241. (page 184)

Ahmadi Zeleti, F., Ojo, A., Curry, E. (2014). Emerging business models for the open data industry: Characterization and analysis. In *Proceedings of the 15th Annual International Conference on Digital Government Research* (pp. 215–226). ISBN: 978-1-4503-2901-9. https://doi.org/10.1145/26127332612745 (pages 184, 196, 197)

Amberg, M., & Schröder, M. (2007). E-business models and consumer expectations for digital audio distribution. *Journal of Enterprise Information Management, 20*(3), 291–303. ISSN: 1741-0398. https://doi.org/10.1108/17410390710740745 (pages 193, 195)

Amit, R., & Zott, C. (2001). Value creation in e-business. *Strategic Management Journal, 22*(6–7), 493–520. ISSN: 1097-0266. https://doi.org/10.1002/smj.187 (page 184)

Archer, P., Dekkers, M., Goedertier, S., & Loutas, N. (2013). *Study on business models for linked open government data*. European Commission. (page 201)

Bärenfänger, R., Otto, B., & Österle, H. (2014). Business value of in-memory technology multiple case study insights. *Industrial Management & Data Systems, 114*(9), 1396–1414. ISSN: 0263-5577. https://doi.org/c10.1108/IMDS-07-2014-0212 (pages 191, 192)

Baden-Fuller, C., & Haefliger, S. (2013). Business models and technological innovation. *Long Range Planning, 46*(6), 419–426. ISSN: 0024-6301. https://doi.org/10.1016/j.lrp.2013.08.023 (pages 181, 182, 199)

[10] For example, the reform of the data protection legal framework in the EU, http://ec.europa.eu/justice/data-protection/reform/index_en.htm.

Bakos, Y. (1998). The emerging role of electronic marketplaces on the Internet. *Communications of the ACM, 41*(8), 35–42. ISSN: 0001-0782. https://doi.org/10.1145/280324.280330 (page 205)

Bekkelund, K. J. (2011). Succeeding with freemium. Specialisation project (p. 57). NTNU, Trondheim. (page 183)

Blumauer, A. (2014). Linked Data in Unternehmen. Methodische Grundlagen und Einsatzszenarien. In T. Pellegrini, H. Sack, & S. Auer (Eds.), *Linked enterprise data*. Berlin: Springer. (page 204)

Boons, F., Lüdeke-Freund, F. (2013). Business models for sustainable innovation: State-of-the-art and steps towards a research agenda. *Journal of Cleaner Production, 45*, 9–19. (pages 184, 185)

Brinker, S. (2010a). *7 business models for linked data* http://chiefmartec.com/2010/01/7-business-models-for-linked-data/ (visited on 2015-10-24). (pages 199, 203)

Brinker, S. (2010b). *Business models for linked data and web 3.0* http://chiefmartec.com/2010/03/business-models-for-linked-data-and-web30/ (visited on 2015-10-24). (pages 199, 200)

Brinker, S. (2010c). *The 8th linked data business model*. http://chiefmartec.com/2010/01/the-8th-linked-data-business-model/ (visited on 2015-10-24). (pages 199, 200, 201)

Bryce, D. J., & Dyer, J. H., & Hatch, N. W. (2011). Competing against free. *Harvard Business Review, 89*(6), 104–111. ISSN: 0017-8012. (page 198)

Buyle, R. (2014). *Business models for open data applications* Apps4EU. (page 204)

Casadesus-Masanell, R., Ricart, J. E. (2010). From strategy to business models and onto tactics. *Long Range Planning, 43*(2–3), 195–215. ISSN: 0024-6301. https://doi.org/10.1016/j.lrp.2010.01.004 (page 184)

Chang, C. C. (2006). Business models for open access journals publishing. *Online Information Review, 30*(6), 699–713. ISSN: 1468-4527. https://doi.org/10.1108/14684520610716171 (pages 191, 194, 195)

Chatterjee, S. (2013). Simple rules for designing business models. *California Management Review, 55*(2), 97–124. ISSN: 0008-1256. https://doi.org/10.1525/cmr2013.55.2.97 (pages 183, 186)

Chesbrough, H. (2010). Business model innovation: Opportunities and barriers. *Long Range Planning, 43*(2–3), 354–363. ISSN: 0024-6301. https://doi.org/10.1016/j.lrp.2009.07.010 (pages 181, 186, 202, 204)

Chesbrough, H. (2012). Why companies should have open business models. *MIT Sloan Management Review, 48* (p. 23).Winter. ISSN: 1532-9194. (page 184)

Chesbrough, H., & Rosenblom, R. (2002). The role of the business model in capturing value from innovation: Evidence from Xerox Corporation's technology spin-off companies. *Industrial and Corporate Change, 11*(3), 529–555. (pages 183, 190, 202)

Cobden, M., Black, J., Gibbins, N., Carr, L., & Shadbolt, N. (2011). A research agenda for linked closed data. In *Proc. of the 2nd International Workshop on Consuming Linked Data (COLD2011)* Bonn, Germany. (pages 200, 204)

Dadzie, A. S., & Rowe, M. (2011). Approaches to visualising linked data: a survey | www.semantic-web-journal.net. *Semantic Web, 2*(2), 89–124. https://doi.org/10.3233/SW-2011-0037 (page 191)

de Reuver, M., Bouwman, H., & Haaker, T. (2009). Mobile business models: Organizational and financial design issues that matter. *Electronic Markets, 19*(1), 3–13. ISSN: 1019-6781. https://doi.org/10.1007/s12525-009-0004-4. (pages 191, 194)

Denning, S. (2014). An economy of access is opening for business: Five strategies for success. *Strategy & Leadership, 42*(4), 14–21. ISSN: 1087-8572. https://doi.org/10.1108/SL-05-2014-0037. (pages 193, 194)

Dezani-Ciancaglini, M., Horne, R., & Sassone, V. (2012). Tracing where and who provenance in linked data: A calculus. *Theoretical Computer Science, 464*, 113– 129. ISSN: 03043975. https://doi.org/10.1016/j.tcs.2012.06.020 (page 205)

Dodds, L. (2010). *Thoughts on linked data business models*. http://blog.ldodds.com/2010/01/10/thoughts-on-linked-data-business-models/ (visited on 2015-10-24). (pages 199, 200, 201)

Doganova, L., & Eyquem-Renault, M. (2009). What do business models do? Innovation devices in technology entrepreneurship. *Research Policy, 38*(10), 1559–1570. (page 185)

Eckert, K. (2013). Provenance and annotations for linked data. In *International Conference on Dublin Core and Metadata Applications* (pp. 9–18). ISSN: 1939-1366. (page 205)

Erickson, J. (2010). *The evolution of linked data business models*. http://bitwacker.com/2010/01/11/evolution-linked-data-business-models/ (visited on 2015-10-24). (page 199)

Ferro, E., & Osella, M. (2013). Eight business model archetypes for PSI re-use. In *"Open Data on the Web" workshop* (p. 13). London. (page 196)

Fisher, T. (2009). *The data asset. How smart companies govern their data for business success* (p. 240). John Willey & Sons. ISBN: 978-0470462263. (page 195)

Gassmann, O., Frankenberger, K., & Csik, M. (2013). *The St. Gallen business model navigator* 3. (pages 184, 185)

Giaglis, G. M., Klein, S., & O'Keefe, R. M. (2002). The role of intermediaries in electronic marketplaces: Developing a contingency model. *Information Systems Journal, 12*(3), 231–246. ISSN: 13501917. https://doi.org/10.1046/j.1365-2575.2002.00123.x (page 205)

Groth, P. (2010). *Another 5 linked data business models*. https://thinklinks.wordpress.com/2010/01/25/another-5-linked-data-business-models/ (visited on 2015-10-24). (page 199)

Hagel, J., & Rayport, J. F. (1997). The coming battle for customer information. *Harvard Business Review, 75*(1), 53–5, 58, 60–1 passim. ISSN: 0017-8012. (page 205)

Hagen, C., Ciobo, M., Wall, D., Yadav, A., Khan, K., Miller, J., & Evans, H. (2013). *Big data and the creative destruction of today's business models*. AT Kearney. (pages 191, 197)

Hart, G., & Dolbear, C. (2013). *Linked data: A geographic perspective* (p. 290). CRC Press. ISBN: 9781439869956. (pages 191, 201)

Hartig, O., & Hartig, O. (2009). Provenance information in the web of data. In *Proceedings of the Linked Data on the Web LDOW Workshop at WWW* (pp. 1–9). (page 205)

Hartmann, P. M., Zaki, M., & Feldmann, N. (2014). Big data for big business? A taxonomy of data-driven business models used by start-up firms. *Paper Submitted*. (page 195)

Hedman, J., Kalling, T. (2003). The business model concept: Theoretical underpinnings and empirical illustrations. *European Journal of Information Systems, 12*(1), 49–59. ISSN: 0960-085X. https://doi.org/10.1057./palgrave.ejis.3000446. (page 185)

Hellman, E. (2010). *8 one-way business models for linked data*. http://go-to-hellman.blogspot.com/2010/01/8-one-way-business-models-for-linked.html (visited on 2015-10-24). (page 199)

Hitzler, P., & Janowicz, K. (2013). Linked data, big data, and the 4th paradigm. *Semantic Web Journal, 4*(3), 233–235. (page 197)

Hougaard, J. L., & Tvede, M. (2010). Selling digital music: Business models for public goods. *NETNOMICS: Economic Research and Electronic Networking, 11*(1), 85–102. ISSN: 1385-9587. https://doi.org/10.1007/s11066-009-9047-0 (pages 191, 192, 193, 195)

Howard, A. (2013). *Open data economy: Eight business models for open data and insight from Deloitte UK*. http://radar.oreilly.com/2013/01/open-data-business-models-deloitte-insight.html (visited on 2015-10-31). (pages 196, 197)

Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine, 14*(6), 1–5. ISSN: 10006788. https://doi.org/10.1086/599595 (page 191)

Janssen, M., & Zuiderwijk, A. (2014). Infomediary business models for connecting open data providers and users. *Social Science Computer Review, 32*(5), 0894439314525902. ISSN: 0894-4393. https://doi.org/10.1177/0894439314525902 (pages 196, 197, 205)

Katz, M. L., & Shapiro, C. (1994). Systems competition and network effects. *Journal of Economic Perspectives, 8*(2), 93–115. ISSN: 0895-3309. https://doi.org/10.1257/jep8.2.93 (page 191)

Kinnari, T. (2013). Open data business models for media industry—Finnish case study (p. 95). Master. Aalto University (page 183)

Krcmar, H., Böhm, M., Friesike, S., & Schildhauer, T. (2011). Innovation, society and business: internet-based business models and their implications. In *1st Berlin Symposium on Internet and Society* (pp. 1–33). Berlin. https://doi.org/10.2139/ssrn.2094222 (pages 183, 184)

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology* (3rd ed., p. 456). SAGE Publications. ISBN: 978-1412983150. (page 188)

Latif, A., Saeed, A. U., Hoefler, P., Stocker, A., & Wagner, C. (2009). The linked data value chain: A lightweight model for business engineers. In *Proceedings of ISE- MANTICS09 International Conference on Semantic Systems*, Graz (pp. 568–575). ISBN: 9783851250602. (pages 199, 206)

Leem, C. S., Suh, H. S., & Kim, D. S. (2004). A classification of mobile business models and its applications. *Industrial Management & Data Systems, 104*(1), 78–87. ISSN: 0263-5577. https://doi.org/10.1108/02635570410514115 (page 194)

Lewoniewski, W., Węcel, K., & Abramowicz, W. (2016). Quality and importance of wikipedia articles in different languages. In G. Dregvaite & R. Damaševičius (Eds.), *Information and Software Technologies: Proc. of 22nd International Conference, ICIST 2016* (Vol. 639, pp. 613–624). Communications in computer and information science. Springer International Publishing Switzerland. https://doi.org/10.1007/978-3-319-46254-7_50 (page 191)

Lewoniewski, W., Węcel, K., & Abramowicz, W. (2017). Relative quality and popularity evaluation of multilingual wikipedia articles. *Informatics, 4*(4). ISSN: 2227-9709. https://doi.org/10.3390/informatics4040043 (page 191)

Linder, J., & Cantrell, S. (2000). Changing business models: Surveying the landscape. In *Accenture Institute for Strategic Change* (pp. 1–15). (page 184)

Lindgren, P., & Rasmussen, O. H. (2013). The business model cube. *Journal of Multi Business Model Innovation and Technology, 1*(3), 135–182. (pages 183, 184, 185)

Magretta, J. (2002). Why business models matter. *Harvard Business Review, 80*(5), 3–8. ISSN: 0017-8012. https://doi.org/10.1016/j.cub.2005.06.028. (pages 184, 186, 198)

Malone, T. W., Weill, P., Lai, R. K., D'Urso, V. T. (2006). *Do some business models per form better than others?* MIT Working Paper 4615-06. (page 183)

Mason, K., & Spring, M. (2011). The sites and practices of business models. *Industrial Marketing Management, 40*(6), 1032–1041. ISSN: 0019-8501. https://doi.org/1.1016/j.indmarman.2011.06.032 (pages 185, 186, 193)

Matt, C., & Hess, T. (2012). Competing against electronic intermediaries: The case of digital music. In *Proc. of the 20th European Conference on Information Systems (ECIS 2012)* Barcelona, Spain. (pages 193, 195, 198)

MIT Technology Review Custom. (2014). *Revolution in progress: The net- worked economy*. https://www.technologyreview.com/s/530241/revolution-in-progress-the-networked-economy (visited on 2017-09-07). (page 183)

Muhtaroglu, F. C. P., Demir, S., Obali, M., & Girgin, C. (2013). Business model canvas per spective on big data applications. In *Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013* (pp. 32–37). ISBN: 9781479912926. https://doi.org/10.1109/BigData.2013.6691684 (page 197)

Novak, A. (2014). Business model literature overview. *Financial Reporting, 1*, 79–130. ISSN: 2036-671X. https://doi.org/10.3280/FR2014-001004 (page 183)

Open Data Institute. (2015). *How to make a business case for open data*. https://theodi.org/guides/how-make-business-case-open-data (visited on 2015-10-31). (page 195)

Osterwalder, A. (2004). The business model ontology—a proposition in a design science approach. Doctoral thesis (p. 169). University of Lausanne. ISBN: 978- 1-4503-0160-2. https://doi.org/10.1111/j.1467-9310.2010.00605.x. arXiv: z0037 (pages 184, 185)

Osterwalder, A., & Pigneur, Y. (2010). *Business model generation* (p. 288). Wiley. ISBN: 9780470876411 (pages 183, 184, 190, 203)

Parker, G., & Van Alstyne, M. (2005). Two-sided network effects: A theory of information product design. *Management Science, 51*(10), 1494–1504. ISSN: 0025-1909. https://doi.org/10.1287/mnsc.l050.0400 (pages 191, 193)

Pellegrini, T., & Ermilov, I. (2013). *Guide and best practices to licensing interlinked data. Deliverable D7.4* FP7 project - LOD2 Creating Knowledge out of Inter linked Data. (page 204)

Pellegrini, T., Dirschl, C., & Eck, K. (2013). Linked data business cube—modelling semantic web business models. In *Proc. of Share-PSI 2.0 Krems Workshop: A Self Sustaining Business Model for Open Data*. (pages 182, 183, 185, 191, 199, 200, 201, 204)

Peters, C., Blohm, I., & Leimeister, J. M. (2015). Anatomy of successful business models for complex services: Insights from the telemedicine field. *Journal of Management Information Systems, 32*(3), 75–104. ISSN: 0742-1222. https://doi.org/10.108/07421222.2015.1095034 (pages 190, 193)

Pine, II B. J., & Gilmore, J. (2016). Integrating experiences into your business model: Five approaches. *Strategy & Leadership, 44*(1), 3–10. ISSN: 1087-8572. https://doi.org/10.1108/SL-11-2015-0080 (page 194)

Plé, L., Lecocq, X., & Angot, J. (2010). Customer-integrated business models: A theoretical framework. *Management, 13*(4), 226–265. ISSN: 1286-4892. https://doi.org/10.3917/mana.134.0226 (page 183)

Popp, K. M. (2011). Software industry business models. *IEEE Software, 28*(4), 26–30. ISSN: 0740-7459. https://doi.org/10.1109/MS.2011.52 (page 199)

Porter, M. E. (2001). Strategy and the Internet. *Harvard Business Review, 79*(3), 62–78, 164. (pages 193, 204)

Rappa, M. (2010). *Business models on the web*. http://digitalenterprise.org/models/models.html (visited on 2015-10-31). (pages 184, 194, 198, 205)

Sahut, J. M., Hikkerova, L., & Khalfallah, M. (2013). Business model and performance of firms. *International Business Research, 6*(2), 64–76. ISSN: 1913-9012. https://doi.org/10.5539/ibr.v6n2p64 (page 194)

Schallmo, D. (2013). *Geschäftsmodell-Innovation: Grundlagen, bestehende Ansätze methodisches Vorgehen und B2B-Geschäftsmodelle* (pp. XXV 335). Gabler Verlag. ISBN: 978-3-658-00245-9. https://doi.org/10.1007/978-3-658-00245-9 (page 185)

Schallmo, D., & Brecht, L. (2010). Business model innovation in business-to-business markets—procedure and examples. In *Proceedings of the 3rd ISPIM Innovation Symposium: Managing the Art of Innovation* (pp. 1–24) Quebec City: The International Society for Professional Innovation Management (ISPIM). (page 186)

Schief, M. (2014). *Business models in the software industry. The impact on firm and M&A performance* (pp. XXVI, 231) Gabler Verlag. ISBN: 978-3-658-04351-3. https://doi.org/10.1007/978-3-658-04352-0 (page 185)

Schief, M., & Buxmann, P. (2012). Business models in the software industry. In *2012 45th Hawaii International Conference on System Sciences* (pp. 3328–3337). ISBN: 978-0-7695-4525-7. https://doi.org/10.1109/HICSS.2012.140 (page 185)

Seppänen, M. (2009). Empirical classification of resources in a business model concept. *Intangible Capital, 5*(2), 102–124. ISSN: 1697-9818. https://doi.org/10.3926/ic.103 (page 185)

Shafer, S. M., Smith, H. J., & Linder, J. C. (2005). The power of business models. *Business Horizons, 48*(3), 199–207. ISSN: 0007-6813. https://doi.org/10.1016/j.bushor.2004.10.014 (pages 183, 184, 185, 190)

Shapiro, C., & Varian, H. R. (1998). *The information economy* (p. 368). Harvard Business Review Press. ISBN: 978-0875848631. (pages 182, 198)

Sheridan, J., & Tennison, J. (2011). Linking UK government data. In *Proceedings of the Linked Data on the Web Workshop (LDOW2010)* (pp. 1–4). ISBN: 978-1- 4614-1766-8. https://doi.org/10.1007/978-1-4614-1767-5 (pages 196, 197)

Sica, R. (2014). *Digital disruption & social business transformation*. https://www.linkedin.com/pulse/20141104152353-1601518-digital-disruption-social-business-transformation (visited on 2017-09-07). (page 182)

Stähler, P. (2002). Business models as an unit of analysis for strategizing. In *Proceedings of the International Workshop on Business Models*. Lausanne. (page 200)

Tapscott, D., & Williams, A. D. (2006). *Wikinomics* (p. 320). Portfolio Hardcover. ISBN: 9781591841388. (page 191)

Tapscott, D., Ticoll, D., & Lowy, A. (2000). Digital capital: Harnessing the power of business webs (hardcover). *Harvard Business School Press Books, 1*. ISSN: 0360-8581. https://doi.org/10.1145/336228.336231 (page 205)

Teece, D. J. (2010) Business models, business strategy and innovation. *Long Range Planning, 43*(2–3), 172–194. ISSN: 0024-6301. https://doi.org/10.1016/j.lrp.2009.07.003 arXiv: z0024. (page 183)

Tennison, J. (2012). *Open data business models*. http://www.jenitennison.com/2012/08/20/open-data-business-models.html (visited on 2015-10-31). (pages 196, 197, 204)

Timmers, P. (1998). Business models for electronic markets. *Electronic Markets, 8*(2), 3–8. ISSN: 1019-6781. https://doi.org/10.1080/10196789800000016 (pages 184, 198)

Vafopoulos, M. (2011). A framework for linked data business models. In *2011 15th Panhellenic Conference on Informatics* (pp. 95–99). IEEE. ISBN: 978-1-61284-962-1. https://doi.org/10.1109/PCI.2011.74 (pages 183, 200)

Veit, D., Clemons, E., Benlian, A., Buxmann, P., Hess, T., Kundisch, D., Leimeister, J. M., Loos, P., & Spann, M. (2014). Business models. *Business & Information Systems Engineering, 6*(1), 45–53. ISSN: 1867-0202. https://doi.org/10.1007/s12599-013-0308-y (page 206)

Venkatraman, N. (1994). IT-enabled business transformation: From automation to business scope redefinition. *Sloan Management Review, 35*(2), 73–87. ISSN: 0019-848X. (page 206)

Węcel, K., & Lewoniewski, W. (2015). Modelling the quality of attributes in wikipedia infoboxes. In W. Abramowicz W (Ed.), *Business information systems workshops* (Vol. 228, pp. 308–320). Lecture notes in business information processing. Springer Inter national Publishing. ISBN: 978-3-319-26761-6. https://doi.org/10.1007/978-3-319-26762-3_27 (page 191)

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly, 26*(2), xiii–xxiii. ISSN: 0276-7783. https://doi.org/10.1.1.104.6570. arXiv: 02767783 (page 187)

Weill, P., & Vitale, M. R. (2001). *Place to space: Migrating to eBusiness models* (p. 372). Harvard Business Review Press. ISBN: 157851245X. https://doi.org/10.1109/EMR20031267030 (pages 184, 195, 198)

Weill, P., Malone, T. W., D'Urso, V. T., Herman, G., & Woerner, S. (2005). *Do some business models perform better than others? A study of the 1000 largest US firms* (Vol. 226). MIT Sloan School of Management. (pages 183, 199)

Weiner, N., Renner, T., & Kett, H. (2010). *Geschäftsmodelle im Internet der Dienste* (Vol. 64, pp. 1130–1137). 1. Stuttgart: Fraunhofer Verlag. ISBN: 978-38-396010-9-9. (pages 183, 184)

Westerlund, M., Leminen, S., & Rajahonka, M. (2014). Designing business models for the Internet of Things. *Technology Innovation Management Review, 4*(7), 5–14. (pages 192, 194, 206)

Wirtz, B. W., Schilke, O., Ullrich, S. (2010). Strategic development of business models. implications of the web 2.0 for creating value on the internet. *Long Range Planning, 43*(2–3), 272–290. ISSN: 0024-6301. https://doi.org/10.1016/j.lrp.2010.01.005 (pages 186, 191, 192, 198, 203)

Wirtz, B. W., Pistoia, A., Ullrich, S., & Göttel, V. (2015). Business models: Origin, development and future research perspectives. *Long Range Planning* https://doi.org/10.1016/j.lrp.2015.04.001 (page 183)

Zhang, X., Williams, A., & Polychronakis, Y. E. (2012). A comparison of e-business models from a value chain perspective. *EuroMed Journal of Business, 7*(1), 83–101. ISSN: 1450-2194. https://doi.org/10.1108/14502191211225392 (pages 191, 192)

Zott, C., & Amit, R. (2010). Business model design: An activity system perspective. *Long Range Planning, 43*(2–3), 216–226. ISSN: 0024-6301. https://doi.org/10.1016/j.lrp2009.07.004 (page 184)

Zott, C., Amit, R., & Massa, L. (2010). *The business model: Theoretical roots, recent developments, and future research*. Barcelona: University of Navarra. (page 183)

Zott, C., Amit, R., & Massa, L (2011). The business model: Recent developments and future research. *Journal of Management, 37*(4), 1019–1042. ISSN: 0149-2063. https://doi.org/10.1177/0149206311406265 (pages 184, 194)

Zwass, V. (2010). Co-creation: Toward a taxonomy and an integrated research perspective. *International Journal of Electronic Commerce, 15*(1), 11–48. ISSN: 10864415. https://doi.org/10.2753/JEC1086-4415150101 (pages 192, 193)

# Chapter 8
# Geographical Profiling with Linked Data


## 8.1 Introduction

Maps were originally used to get from one location to another. Nowadays, the proliferation of online maps triggers a number of other applications serving various sophisticated purposes. Maps are more often used for visualization of natural, social, and economic phenomena. For example, it is possible to indicate the average air pollution in a given city, the levels of crime, or particularly interesting cultural spots. Such visualization requires to combine external information with an underlying map and is referred to as *mashup*. The prerequisite is that the data has a spatial dimension. Things that are mainly perceived through latitude and longitude, like real estate, are best browsed on the map. The first mashup was created in 2005 by Paul Rademacher who put Craigslist homes and rentals on Google Maps (DuVander, 2010). It spurred a number of other similar initiatives and demonstrated the great potential of mashups both on the visual part as well as on the idea of stitching together pieces from various datasets.

New ways of querying and visualizing data are possible thanks to the development of geographic information systems (GIS). The advances were related to the following trends (Grant et al., 2014):

- an increase in the amount of geospatial information available, created mainly through smartphones,
- an increase in the accuracy of data used to determine locations,
- advancement of methods used to analyze geospatial information, partly enabled by the greater standardization of data and databases,
- improvements in hardware, such as GPS receivers,
- maturation of open-source software, making it more accessible to a broader audience.

The progress in the spatial information domain is observed through various measures. Manyika et al. (2011) estimated size of the global collection of personal



K. Węcel, *Big, Open and Linked Data*, Business Information Systems,
https://doi.org/10.1007/978-3-031-07147-8_8

geo-location data to be at least 1 petabyte in 2009, and to grow by about 20% a year. By 2020, this data pool was expected to provide $500 billion in value worldwide in the form of time and fuel savings. Individual companies, like the ones offering navigation hardware and software, also collect large volumes of spatial data. The database of the leading company—TomTom—already in 2012 contained more than 5000 trillion data points from its navigation devices and other sources, and collected 5 billion new data points every day. The data points included time, location, direction, and speed (OECD, 2013).

This chapter is focused on deriving user profiles from mobile data available for telecom operators and leveraging the linked geographical data to characterize the user's environment. This environment may be understood in terms of surrounding people (the people contacting a given user), geographical locations visited (in terms of base transceiver stations, including a specification of the visited places), mobility patterns, as well as services used by the user.

## 8.2 Spatial Information

Spatial information is information that contains a georeference, which can be coordinates, an address, a postcode, or any other area code defined for statistical purposes (ACIL, 2011). It can be used for fusing of different types of data through a common location in order to analyze and understand linkages between people, location, demographics, economic development, and social services. It further facilitates the identification of the spatial dimension of phenomena, spatial characteristics of needs, and location-based offers.

Geographic information and socio-economic data form the most valuable parts of databases held by public sector institutions. According to the assessment of the European Commission (European Commission, 2004), spatial information was embedded in up to 80% of all the datasets held in public sector institutions. The importance of this asset was confirmed by Fornefeld et al. (2008) who mentioned geographic information as one of the three categories of economically relevant data within PSI (see Sect. 5.2.2). This publication is particularly important as it studied in detail the geographic information. Although PSI is held by public bodies, many researchers believe that the major value and potential of spatial information lies in the private sector. There are many established business activities of high potential, including navigation, location-based services, and geomarketing (Koski, 2011).

Data analysis and visualization very often use geospatial and mapping data to provide geospatial intelligence. One of the activities can include the identification of geographical patterns in data. Spatial information is critical to better decision making by providing user-friendly map-based interfaces. It can support advanced analysis and city planning, environmental policy-making, forestry, and agriculture. It is also particularly interesting for the real estate market.

Spatial information can improve decision making in three areas (ACIL, 2011):

- Integration—each event or phenomenon can be assigned a location 'signature,' which then provides a mechanism for linking sources of data that cannot be easily associated using conventional approaches.
- Analysis—the consequences of decisions can degrade with distance; therefore, the analysis needs to be enriched with location and consider the location criteria. For example, optimizing bus routes requires spatial analysis.
- Visualization—patterns and trends can be illustrated in a form that can be easily understood by individual stakeholders.

Open data is already widely used in the real estate market. It can be leveraged on various levels. At the lowest level, open data can be used to generate search options for individuals like home buyers, sellers, and real estate agents. At the middle layer, open data can be used by planners and city advocates to monitor land use and analyze real estate markets. At the highest level, open data can be used by policy makers to develop strategies by combining with other open data related to geography, education, transportation, and the environment (Center for Open Data Enterprise, 2016).

The European Commission report (European Commission, 2004) mentioned also a high fragmentation of spatial-related markets—there were about 6000 organizations, both public and private, that dealt with spatial information. Knowing the potential of linked data, several researchers investigated how it could be leveraged to improve the potential of geographic information in linking and combining datasets through shared locations. Linked open data was usually used as an additional knowledge source. For example, Abel et al. (2012) aggregated knowledge from GeoNames and DBpedia to recommend points of interest (POI) based on social web. They demonstrated that user modeling quality improved when LOD-based background information was included in the process. According to Hart and Dolbear (2013), the potential was still not realized as data was not well organized and the technology was not mature and accessible enough to users.

## 8.3 Mobile Data

Mobile devices generate a lot of user-related data of various types. Data can be generated by a customer alone or by a network operator (see Fig. 8.1). The first encompasses various data concerning behavior, for example, clickstream, application usage, or personal data. The latter concerns mostly the technical layer, with the most prominent *call detail records* (CDR). Generated data is then stored and accessed in IT systems, as necessary and relevant for business activities, like billing and offering value-added services. Overall, information about identity, location, social interactions, movements, finance, and even ambient environmental conditions can be derived from the data logged in mobile systems. Mobile telecom operators have access to uniquely detailed and tractable data, not found in other

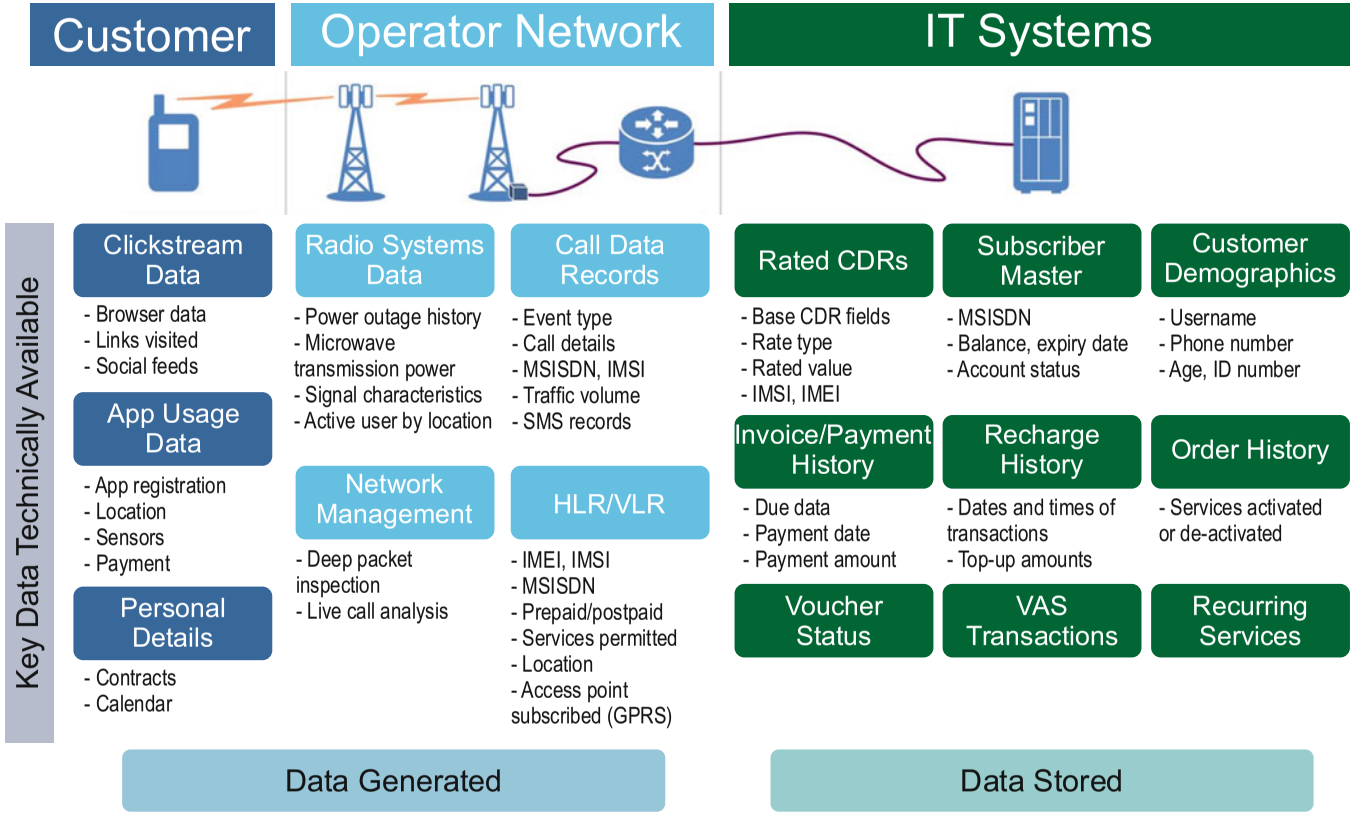



**Fig. 8.1** Key datasets related to mobile networks and associated systems. Source: (Naef et al., 2014a, p. 14) © Cartesian, Inc.

information sources (Naef et al., 2014a). Taking into account also the number of subscribers, such a scale is interesting enough for deriving additional value from mobile data.

Overall, Naef et al. (2014b) distinguished the following categories of mobile data:

- Identity and demographic data—basic information about a user: name, age, gender, nationality; useful for segmentation.
- Location and movement data—user location and movement patterns; it can include accuracy depending on the location method.
- Financial and economic data—information about purchases, including value-added services and transaction patterns; useful for understanding the subscriber's economic situation.
- Social/browser data—information inferred from social network analysis (call records, browser usage).
- Usage—mobile services used by a user.
- Sentiment and trends—language used by a subscriber; emerging popular trends.
- Diagnostic/ambient conditions—measurements of environmental conditions via sensors available in smartphones.

A lot of information can be derived from call detail records (CDR), which have already generated groundbreaking insights in various domains. For instance, mobile phone data has already been used to study the geographical partitioning of countries (Ratti et al., 2010), the spread of information in social networks (Miritello et al., 2011), and human mobility and behavior in cities (Toole et al., 2015). Further opportunities emerge from enriching CDR with open data.

In developing countries, mobile phone data has already been used to tackle social and economic challenges. Orange has contributed anonymous datasets based on detail records of phone calls and text exchanges between more than 9 million of Orange's customers in Senegal in 2013 (de Montjoye et al., 2014). It is a foundation of an open innovation data challenge—the D4D-Senegal challenge. The goal of the challenge is to help address society development questions in novel ways by contributing to the socio-economic development and well-being of the Senegalese population.

CDRs are used mostly for understanding of human movements. Mobile phones have become increasingly ubiquitous; therefore, it is now possible to collect individual-level, longitudinal data on human movements on a massive scale. This is an important data, for example, for disease transmission analysis. Wesolowski et al. (2012) analyzed mobile phone call detail records representing the travel patterns of 15 million mobile phone owners in Kenya over one year in order to build a model of malaria spreading. Later, a similar approach was applied for the dengue epidemic in Pakistan (Wesolowski et al., 2015). The epidemiological model based on mobility data from about 40 million mobile phone subscribers and predicted the timing and geographic spread of epidemic throughout the country.

Deville et al. (2014) used datasets of more than 1 billion mobile phone call records from Portugal and France to estimate population densities on a national

scale. The comparison with official human population mapping methods showed that the results were satisfactory. The advantage over the classical method is that not only estimations can be done spatially but also regarding temporal dimension. Their method made it possible to map contemporary and changing human population distributions over relatively short intervals. That paved the way for new applications and a near real-time understanding of patterns and processes in human geography.

The usefulness of mobile data is particularly visible in developing countries. Data that is good on a macroeconomic scale is not always useful for supporting single users of phones. The specific challenge in developing countries is the fact that most phones are prepaid, which means that mobile operators lack key information about the user like gender and other demographic variables (Jahani et al., 2017). A mobile phone is sometimes used by the whole family or even a village. Such data is useless for social science and development economic research.

There are, however, some research initiatives that aim at the reconstruction of demographic variables based on CDR. For example, Jahani et al. (2017) developed a framework to extract more than 1400 features from standard mobile phone data and used them to predict useful individual characteristics and group estimates. They validated their framework by showing how it could be used to reliably predict gender and other information for more than half a million people in two countries. For this purpose, standard machine learning algorithms were used on a sample of only 10,000 users. Individual's gender was predicted with an accuracy ranging from 74.3 to 88.4% in a developed country and from 74.5 to 79.7% in a developing country using only metadata.

As can be seen, there is a broad range of opportunities to leverage mobile data for development efforts. Additional indirect value can be created for mobile users through the appropriate use of mobile data in projects to improve public services (e.g., health, transportation). The positive externalities can only be unlocked by sharing anonymized data with trusted third parties in order to execute development projects. It is important to remember that companies are only data custodians, and data is still owned by customers who should retain control over their data (see Sect. 6.4 on data ownership).

## 8.4 Geographical Linked Data-Based Profiling

This section describes user profiling based on base transceiver stations (BTS) characteristics derived from the geographical linked data. To deliver the method, some arbitrary decisions were made and will be described in this section. Nevertheless, they do not affect the method itself and were necessary to illustrate the approach in a graphical manner.

The profiling process was split into several steps. First, the information about BTS location and its neighborhood had to be retrieved. This is described in Sect. 8.4.1. Then, based on this information, a summary of BTS profiles was prepared. In the next step, we characterized users that log in to specific BTS stations, what is described in Sect. 8.4.3.

### 8.4.1 Characteristics of Base Transceiver Stations

For building characteristics of base transceiver stations (BTS), we used one of the most complete open source of geographical data—OpenStreetMap.[1] Besides typical geographical data like the shapes of objects on the map, it also contains a taxonomy of concepts used to annotate these objects. This classification can be used to reason about more general categories of analyzed objects. The selected categories are then used to construct a profile.

At the beginning, the information about BTS locations was retrieved. There were about 8000 unique locations. The number of antennas was much higher, but for profiling only unique coordinates were necessary. Our source here was MySQL database and we used a Python script to get all locations along with latitude and longitude.

For each location of BTS, we had to retrieve a list of objects in the neighborhood along with their categories. As a source of data for our queries we used LinkedGeoData,[2] which is a derivative of OpenStreetMap. These sources distinguished two main categories of objects: nodes (just a point according to GIS terminology) and ways (lines or polygons in GIS terminology). As there are different ways of measuring the distance and similarity between different kinds of objects, at least it was found in our experiments, we had to separately retrieve nodes and ways. This distinction is retained to the end of this section.

Retrieval of a geographical object from a technical point of view has been conducted using SPARQL queries with LGD endpoint.[3] As there were about 8000 locations, two kinds of objects, two means for object capturing (a bounding box and a circle), and 3 various distances, we had to post over 90 thousand queries. The data collection was the most time consuming step in this method. Fortunately, once the data was retrieved, it could be reused when necessary.

The following figures present various approaches for retrieving geographical objects from linked open data sources. Visualizations are prepared using Leaflet,[4] based on the layers provided by OpenStreetMap. We considered various approaches to select objects in the neighborhood of BTS using spatial queries. The selection can base on a box or a circle. The decision on the method is particularly important for cities, where locations are usually dense with annotated objects. We have tried boxes and circles of various sizes (1 km, 2 km, 5 km). Box of 5 km size was too big—it retrieved too many objects (Fig. 8.2a). Box of 2 km still captured too many objects. Considering the propagation of GSM signal, a circle seemed to be a more natural shape. Finally, the circle of 1 km diameter has been chosen as a mean to query the LGD (Fig. 8.2b).

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

[2] http://linkedgeodata.org/.

[3] http://linkedgeodata.org/sparql.

[4] Leaflet is a JavaScript library for interactive maps, http://leafletjs.com/.


**Fig. 8.2** Nodes selection strategy. (**a**) Box of 5 km size. (**b**) Circle of 1 km diameter

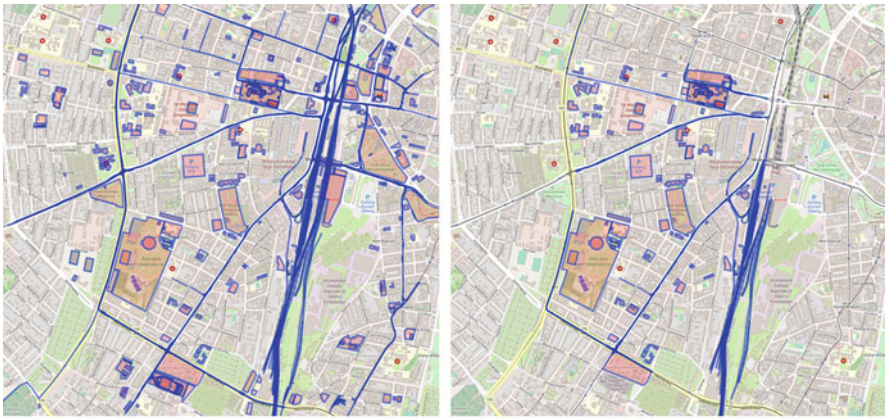



**Fig. 8.3** Ways selection strategy. (**a**) Box of 5km size. (**b**) Circle overlap of 0.01 toleration

Retrieval of ways was more complicated as the distance calculated by Virtuoso server executing SPARQL queries was not intuitive and momentarily looked incorrect. In the documentation, we could read that it was not fully implemented for some kinds of GIS objects. Therefore, several methods have been tested. Similarly to nodes, ways within the box of 5 km covered too wide area (Fig. 8.3a). Finally, the circle overlap after toleration parameter[5] tuning has been chosen as a method to query for neighboring ways (Fig. 8.3b).

[5] A technical parameter of SPARQL function `st_within`.

| loc_id | class | count |
|---|---|---|
| 6 | geo:Feature | 133 |
| 6 | lgdm:Node | 133 |
| 6 | lgdo:Amenity | 48 |
| 6 | lgdo:ManMadeThing | 35 |
| 6 | lgdo:Surveillance | 34 |
| 6 | lgdo:BarrierThing | 27 |
| 6 | lgdo:Bench | 13 |
| 6 | lgdo:LiftGate | 10 |
| 6 | lgdo:Shop | 10 |
| 6 | lgdo:HighwayThing | 9 |
| 6 | lgdo:BusStop | 8 |
| 6 | lgdo:WasteBasket | 8 |

**Fig. 8.4** A profile of a sample BTS location

Sample results obtained from the above procedure are given in Fig. 8.4. First of all, we need to clarify the namespaces used. Prefix `lgdm` stands for http://linkedgeodata.org/meta/, `lgdo` for http://linkedgeodata.org/ontology/, and `geo` for http://www.opengis.net/ont/geosparql#. The `loc_id` specifies that we look at the location with the identifier id=6. The ontology concepts in the column `class` were created via mapping from OSM artifacts.[6] The last column `count` specifies the number of occurrences of a particular class. From the table we can read that in the neighborhood of the BTS there are 133 geographical objects, all of them are nodes (points), and they are assigned respective categories as follows: 48 amenities, 35 objects made by humans, 34 surveillance cameras, 10 shops, and so on.

In order to best characterize the BTS stations, we have arbitrarily selected 30 categories that can provide meaningful information about a user when used in the profile:

- Atm,
- Bank,
- BusStop,
- Cafe,
- Cinema,
- FastFood,
- FuelStation,
- HistoricThing,
- Hospital,
- Hotel,
- Leisure,
- Library,
- Monument,
- Office,
- Parking,
- Pharmacy,
- Pitch,
- PlaceOfWorship,
- PostOffice,
- Pub,
- RailwayStation,
- Restaurant,
- School,
- Shop,
- SportsCentre,
- SportThing,
- Supermarket,
- Theatre,
- TramStop,
- University.

[6] For example, for an amenity the following specification is provided https://wiki.openstreetmap.org/wiki/Key:amenity.

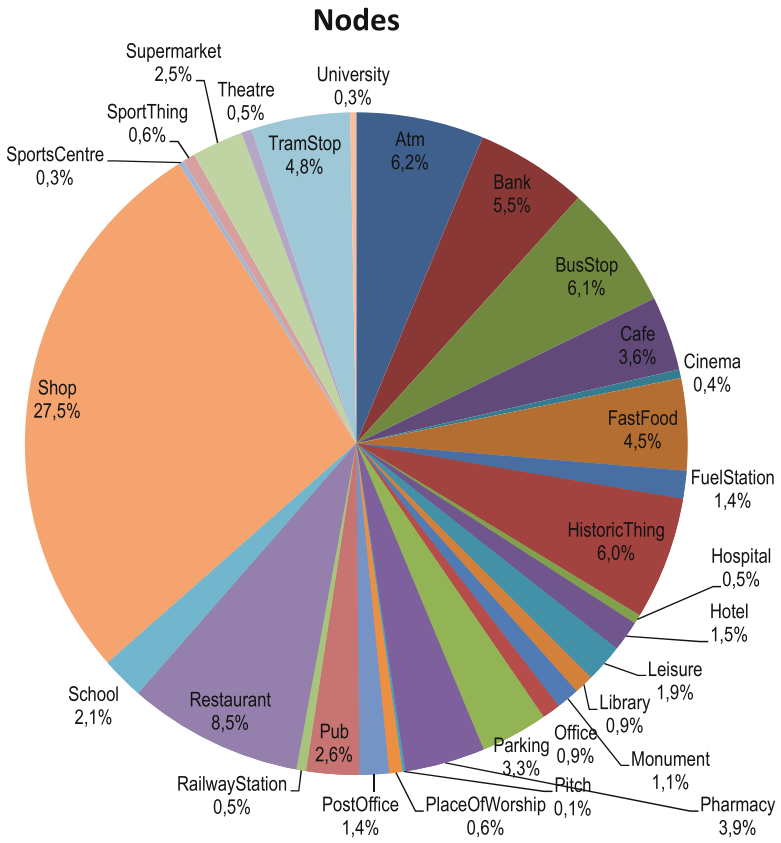



**Fig. 8.5** Average distribution of objects of type node among predefined 30 categories of objects

A profile of BTS is calculated based on the categories of objects located in the neighborhood. Simple counting of objects is not appropriate for determining the characteristics of the neighborhood. Not only the share of categories is important but also their distribution among locations. For example, there are 15 venues of type parking and four shops around an average BTS. Figures 8.5 and 8.6 show the average distribution of categories of objects, for nodes and ways, respectively. Effectively, they represent the profiles of the average location. Distributions are very different although they characterize the same locations. This effect can be explained by looking at the patterns followed by the OpenStreetMap community. For example, shops are usually small and thus are marked as points and they dominate in the 'nodes' dataset—27.5% of all object in the neighborhoods of BTS stations have the category 'shop.' Objects like parkings or leisure areas are better represented by showing the area; hence, they are more popular in the 'ways' dataset. Here 33.2% of objects are annotated as 'parking' and 22.3% as 'leisure.'

Various aspects of the profiles can be analyzed. For example, Fig. 8.7 presents stations that have hotels in their neighborhood. The center of each circle reveals the

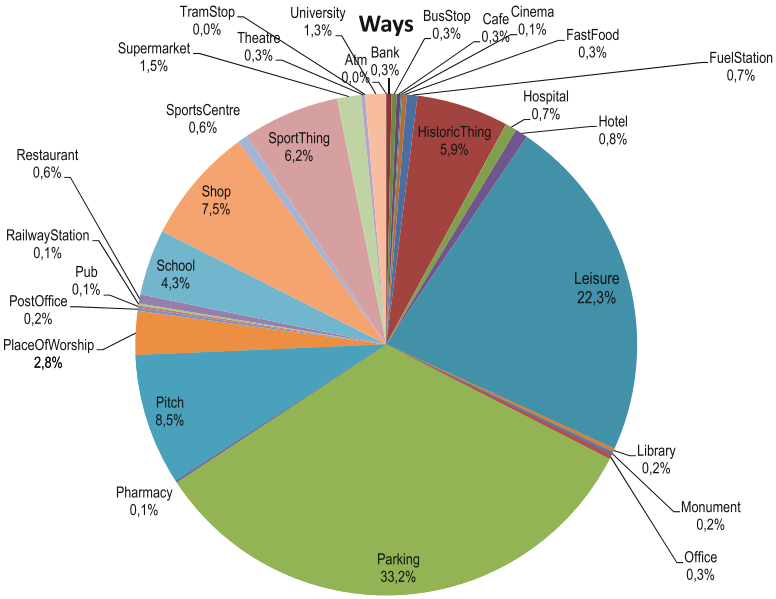



**Fig. 8.6** Average distribution of objects of type way among predefined 30 categories of objects

location of BTS. The radius of the circle is proportional to the number of hotels found in the neighborhood of BTS. It is interesting to observe that according to OSM, the highest density of hotels in Poland is in Gdańsk. Figure 8.8 shows a closer look on this city.

Based on this list, we have prepared another interesting visualization. Each BTS location is annotated with the most popular object in the neighborhood. If there are a lot of schools in the vicinity, the BTS is annotated as 'school.' The most important objects are coded as follows: red—shops, blue—historical objects, orange—schools, green—public transport. Visualization for the whole country is presented in Fig. 8.9 and for Poznań in Fig. 8.10.

### *8.4.2 TF-IDF Weighting Schema*

Simple aggregation of geographical categories assigned to places is not sufficient to correctly profile locations. Relative values are more important than absolute values. For example, as there are many more shops than libraries, the results would be biased if we had not included this correction in the characteristics of locations.

In order to alleviate the effect of uneven distribution of categories, we propose to use TF-IDF weighting schema known from information retrieval (Abramowicz, 2008). TF-IDF is actually a product of two statistics: term frequency (TF) and inverse document frequency (IDF). Details of the method are presented in (Węcel, 2015).

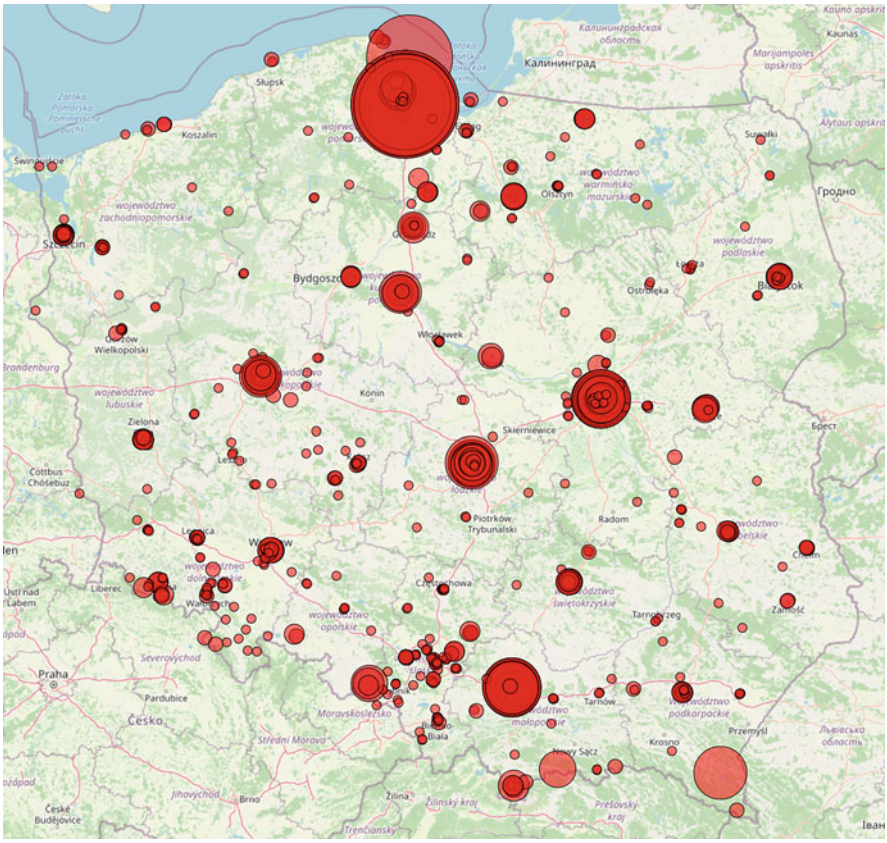



**Fig. 8.7** Number of hotels located within BTS stations—Poland

### *8.4.3 Characteristics of Users*

The idea for profiling of users is based on the following assumption—people visit places according to their interest or occupation. For example, people interested in sports can usually be met near stadiums, gyms, etc. Similarly, people interested in shopping will usually be spotted close to malls. In fact, we are interested in information, if given users visit a type of object more often than a typical user.

The calculation of the characteristics of users is rather straightforward. The profile of a user is prepared as a weighted sum of the profiles of visited BTS locations, where the weight is the number of connections initiated with a given BTS. Similarly, we can visualize the profile as a pie chart. The following Figs. 8.11 and 8.12 present sample profiles for users.

In order to emphasize the differences between user profiles, we put them together on a radar chart. We normalized the values on each axis in such a way

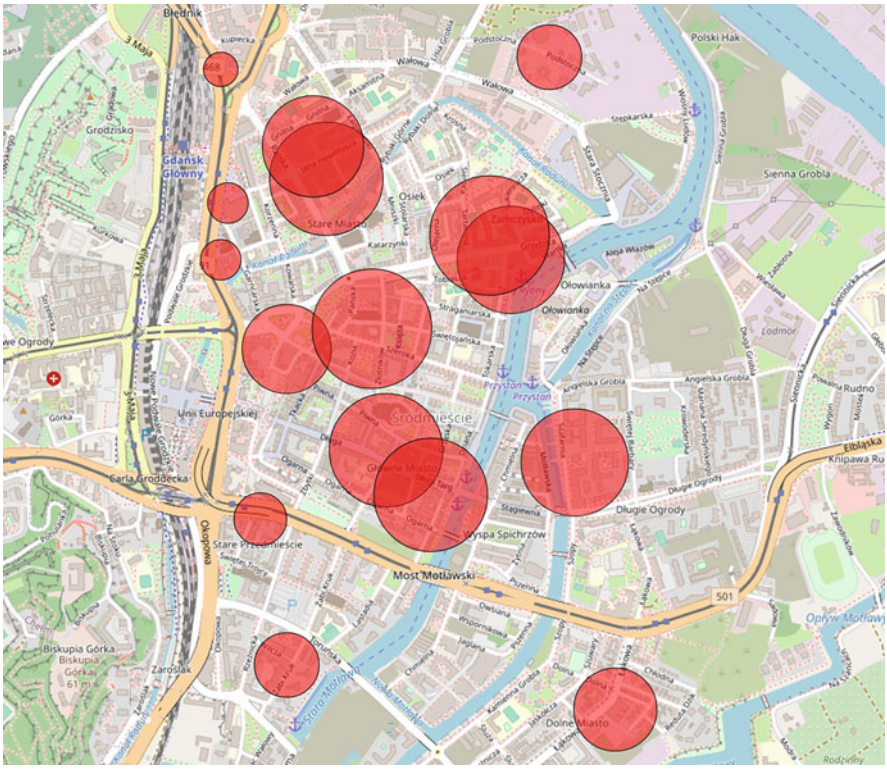



**Fig. 8.8** Number of hotels located within BTS stations—Gdańsk

that user with the highest value of a given object type in the profile has value 1. Figure 8.13 compares two sample users already presented above. Moreover, reducing the number of categories should help in homogenizing of the distribution and easing the comparison process.

## 8.5 Advanced Geographical Profiling with Latent Variables

In some cases, the vector space stemming from annotations can have a high number of dimensions. This makes the calculations more computationally intensive. In order to avoid this problem, number techniques to reduce the dimensionality of the feature space can be employed (Cunningham, 2007). There are two approaches: either the most informative variables are selected or variables that are correlated are replaced by a new 'compound' variable without significant loss of information. The latter is more interesting. Additionally, the process can be supervised or unsupervised. Figure 8.14 summarizes the possible approaches in dimension reduction research.

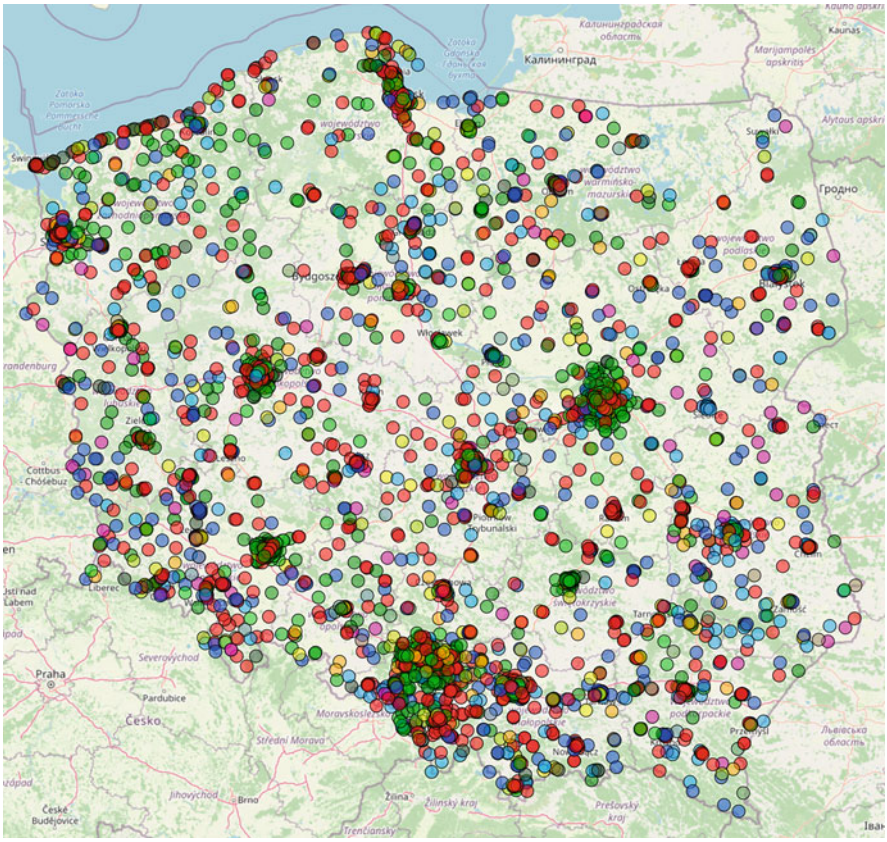



**Fig. 8.9** Most popular annotations of BTS locations—Poland

The content of this section builds on the previous work presented in (Węcel, 2015). Weighting schema was already improved by the introduction of TF-IDF weighting. In this section, we make further improvements through the use of Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA).

### *8.5.1 Data Flows*

Building profiles of users requires several data transformations. During data collection, certain decisions had to be made. For example, we allowed many sources and transformation variants, and this caused the multiplication of possible data processing pipelines. Details of possible combinations are illustrated in Fig. 8.15.

First, there are two types of geographical objects that can be annotated in OpenStreetMap—nodes and ways. *Nodes* represent single-point objects, e.g., bus

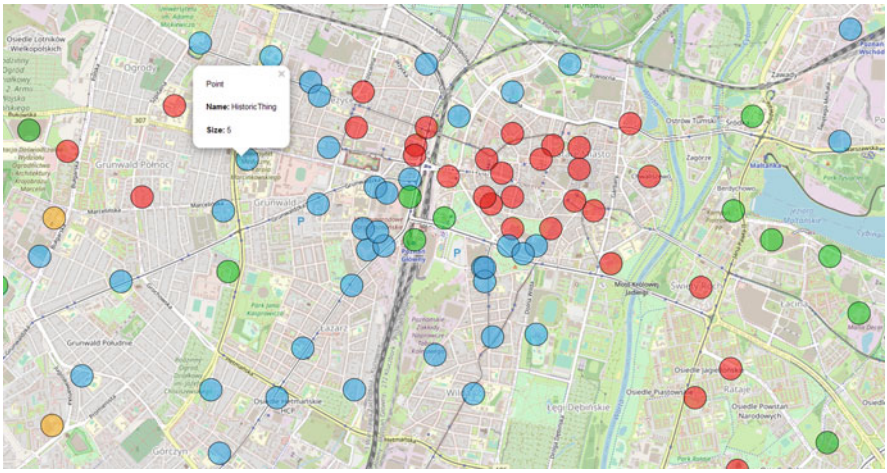



**Fig. 8.10** Most popular annotations of BTS locations—Poznań

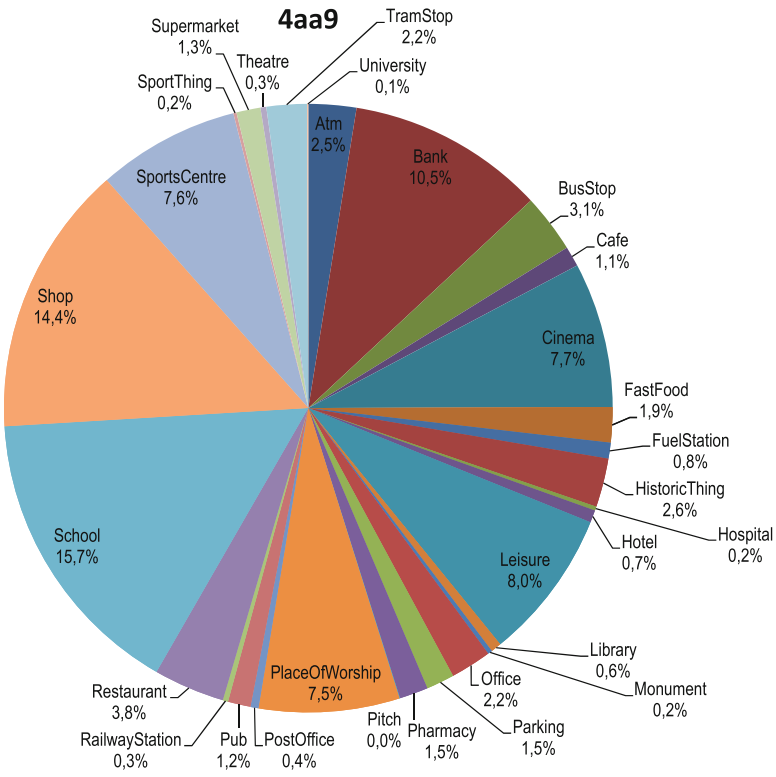



**Fig. 8.11** Geographical Linked Data-Based profile of the sample user 4aa9

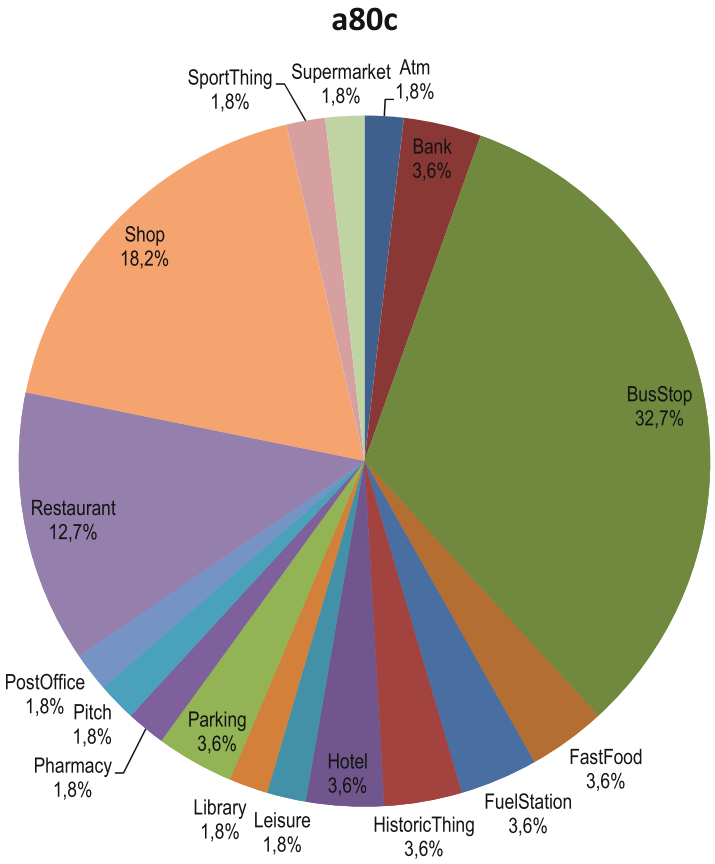



**Fig. 8.12** Geographical Linked Data-Based profile of the sample user a80c

stop, ATM. *Ways* are used for objects occupying bigger space that can be represented as polygons, e.g., malls, parks. As they are retrieved using different functions, they form two separate sources. During our experiments, we have retrieved 127,830 nodes and 212,881 ways.

Second, objects can be described with different term spaces. In essence, within OSM there are almost 500 categories used to describe venues. They are organized in a hierarchy and in fact there is some redundancy in the annotations, i.e., when a category is added, its parents are also added for query efficiency reasons (easier counting of occurrences). In order to simplify the initial experiments, we restricted the list of terms to 30 most common, useful, and moderately specific categories. We also defined mappings so that it was possible to properly describe all geographical objects. This dataset is called *filtered* as not all categories are taken into account. The models that we apply at a later stage were designed to cope with the big number of dimensions; therefore, we also use *full* data space in our experiments. At this

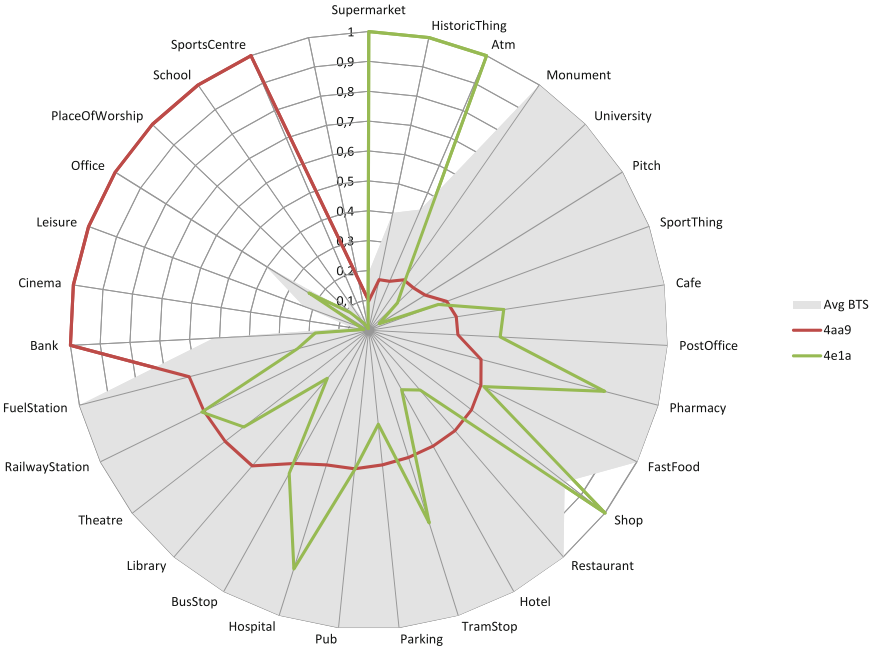



**Fig. 8.13** Comparison of user profiles on radar charts—restricted to two users

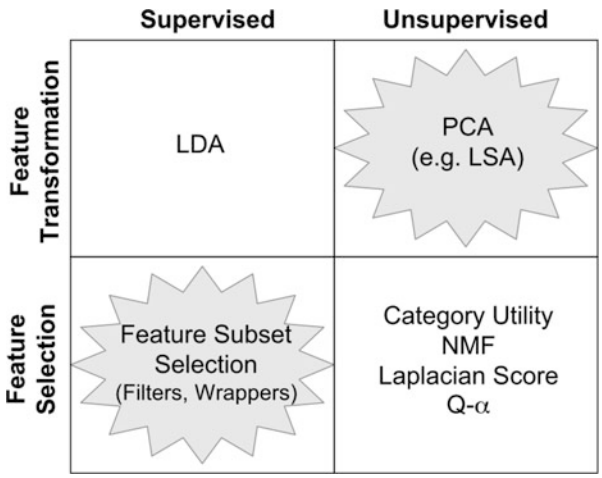



**Fig. 8.14** Key distinctions in dimension reduction research. Source: (Cunningham, 2007)

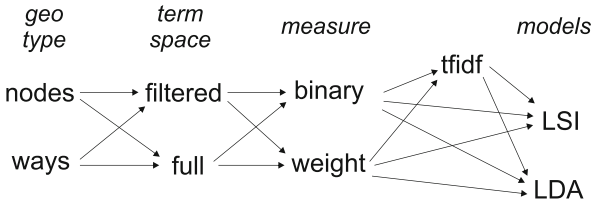



**Fig. 8.15** Data flow variants

point, we have four datasets: `nodes_full`, `nodes_filtered`, `ways_full` and `ways_filtered`.

Third, different measures can be assumed for the vector space. In the indexing phase, we note how many objects of a given category are found in the neighborhood of the analyzed location. In the *weighted* variant we take the number of objects—the higher the term frequency, the more important it is for the location. In the *binary* variant we only mark the presence of a given object type. The second approach can be justified for analysis of the coverage of locations with certain object types. Now we have eight datasets.

Fourth, the prepared eight datasets can be fed directly to the models or further transformation of weights can be applied using TF-IDF schema. We built five models combining various transformations, which were then applied to our data to transform the profiles of locations from the raw format to the ones calculated with the model. It is important to note that in the case of LSI and LDA we obtained a reduced term space. We assumed five topics in the experiments. This is actually a parameter to our scripts that build models, so further variants can be considered, for example, the initial 500-term space can be reduced to 10- or 30-dimensional space. It is also important to note that the prepared tools are universal, i.e., the general goal is user profiling, but detailed methods can be replaced as necessary with new implementations.

### *8.5.2 Tools*

Several tools have been tried to carry out calculations. The requirement was that the library implements LDA and is available in Python. One of the most convenient is Gensim,[7] the library for vector space and topic modeling. It makes heavy use of efficient matrix operations contained within NumPy package. It is also important to note that Gensim is an open-source toolkit.

In this section, we describe details of the prepared scripts in Python so that the results can be reproduced and used in another configuration. Whereas the location profiles are rather stable, new groups of users can be analyzed and parameters of the model fine-tuned.

[7] https://radimrehurek.com/gensim/.

**Corpus Preparation** The output from SPARQL query for retrieving categories of objects in the neighborhood is formatted as CSV file (see Listing 1). For example, BTS with id=1 has one annotation (a bus stop), and there is a railway station nearby BTS with id=4. BTS with id=20 has objects of four categories: two shops, one fuel station, one supermarket, and one office. Some BTS did not have any annotated objects in the neighborhood. The input file is converted using appropriate Python scripts—one produces corpus in *weighted* variant, the other reduces to *binary* observations.

```
1;BusStop;1
4;RailwayStation;1
11;School;1
20;Shop;2
20;FuelStation;1
20;Supermarket;1
20;Office;1
```

**Listing 1** Listing of nodes_filtered.csv

The output is stored in the dedicated folder in multiple formats for further reference and reuse. The most popular in similar applications is the market matrix format as it is human readable and compact at the same time. It is suitable for representing sparse matrices. Only nonzero entries need to be encoded and the coordinates of each entry are given explicitly. For example, '4 7 1' means that cell (4,7) has value 1. As can be read from Listing 2 there are 4802 rows (indexed BTS stations), 30 columns (categories), and 32,042 entries.

```
%%MatrixMarket matrix coordinate real general
4802 30 32042
1 1 1
2 2 1
3 3 1
4 4 1
4 5 1
4 6 2
4 7 1
```

**Listing 2** Listing of nodes_filtered.mm

**Building Models** Based on a given corpus, different models can be built. Such models are then applied to new data to make it consistent with the model representation and to allow inferences, e.g., searching for similar BTS profiles in new space. Five models are built at the same time (as defined in Fig. 8.15):

- tfidf—model with original term space but modified weights, reflecting the popularity of certain categories;
- lsi—latent semantic indexing, reduced-space model;

- tfidf+lsi—same as above but the learning is conducted on a modified dataset, i.e., using tfidf weights instead of raw weights;
- lda—latent Dirichlet allocation, reduced-space model;
- tfidf+lda—analogously to tfidf+lsi, i.e., lda using tfidf weights.

The next Python script requires a corpus name as an input. It runs on eight corpora producing altogether 40 models.

**Visualizations** Automating the chart drawing was one of the goals of the whole pipeline design. Taking into account the number of possible analysis variants, it was indispensable to generate charts quickly and name them correctly. The naming convention was also well thought-off, stemming from the variants in Fig. 8.15.

The prepared scripts address the differences in the source matrices used for visualization. In the case of raw data and tfidf model, we have one matrix: document-term. Such a matrix can easily be presented in a single bar chart, where labels of horizontal axis are just categories used for indexing. The idea of both LSI and LDA is to decompose this matrix into two separate matrices: document-topic and topic-term. This necessitates to prepare two visualizations for each model, hence split into two scripts. Examples of visualizations can be found throughout the next sections.

**User Profiling** Having prepared data about BTS stations, we can move on to user profiling. For this purpose, we need information about which BTS stations, also referred to as locations, have been visited by a certain user. We also note the intensity of visits in certain places by observing the number of actions conducted within the station: SMSes sent, calls initiated.

The final step—user profiling—is conducted by script that requires two parameters. The first provides the list of BTS stations visited by a user. The second parameter is the name of the corpus to use, which is equivalent to a weighting scheme. The script applies all models built for the corpus and yields CSV files with user profiles and visualizations both in PNG and PDF formats.

### *8.5.3 Latent Dirichlet Allocation*

Frequent labels can generate "false positives," i.e., if shops are located in the neighborhood of over 60% of BTS stations, they are not very specific for them. Then, the subscribers can be suspected to visit shops just because there are mostly shops in many areas in the city. Our goal is to adjust the weights of categories describing the geographical location so that they properly reflect the intent of subscribers. Profiles should then be based not on frequency but on the profiling power of certain categories. Here the solution based on probabilistic approaches (latent variables) will be studied.

**Dirichlet distribution**, denoted Dir($\boldsymbol{\alpha}$), is the multivariate generalization of the beta distribution. It is a family of continuous multivariate probability distributions

parameterized by a vector $\boldsymbol{\alpha}$ of positive reals. Variable $K$ denotes number of categories (integer), and $\boldsymbol{\alpha} = \alpha_1, \cdots, \alpha_k > 0$ are the concentration parameters—the bigger the value, the bigger "the concentration of the probability".

Probability density function is defined as follows:

$$f(x_1, \cdots, x_K; \alpha_1, \cdots, \alpha_K) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}, \quad \text{where } x_i \in (0, 1), \sum_{i=1}^{K} x_i = 1 \tag{8.1}$$

The normalizing constant is the multinomial Beta function, which can be expressed in terms of the gamma function:

$$\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}, \qquad \boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K). \tag{8.2}$$

**Latent Dirichlet allocation** (LDA) is a generative model that allows sets of *observations* to be explained by *unobserved groups* that explain why some parts of the data are similar. It is actually an example of Bayesian unsupervised learning model.

It is assumed that documents belonging to the analyzed collection are the result of two random processes. Each document is a collection of words (bag of words assumption). This is further decomposed into two parts: each document is represented as a mixture of latent topics, and each topic is represented as a mixture over words. These mixture distributions are assumed to be Dirichlet-distributed random variables that must be inferred from the data.

The generative process works as follows. First, one of the topics is chosen accordingly to a given distribution over topics. Second, one word from this topic is chosen accordingly to a given distribution over words. These steps are repeated until a document (a bag of words) is created. When we already have a collection of documents, we need to reverse the process and reconstruct the distributions, which happen to be Dirichlet distributions (the best approximation according to theory). We need to "guess" the best possible distributions (in fact, their parameters) that could produce the observed set of documents with the highest probability. We also need to learn the set of topics, the topic's word distribution, and the document's topic distribution (particular topic mixture of each document). The above is a problem of Bayesian inference and several approaches for providing the solution are available: a variational Bayes approximation of the posterior distribution, Gibbs sampling, and expectation propagation.

More formally, the generation process can be described in terms of probabilities.

$$P(word) = \sum_{k=1}^{K} P(Topic_k) P(word|Topic_k), \tag{8.3}$$

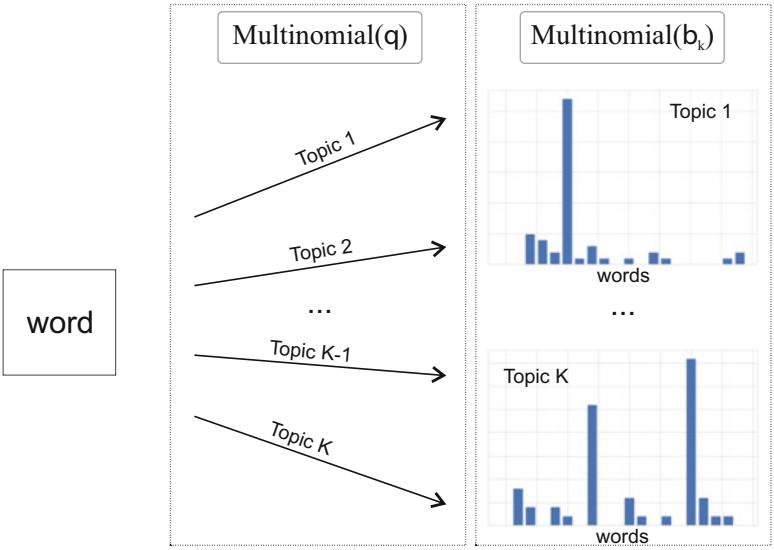



**Fig. 8.16** Explanation of a mixture model for LDA. Source: own work

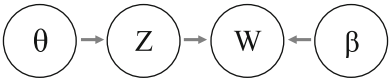



**Fig. 8.17** Dependencies between variables in LDA

where:

$P(Topic_k)$—a mixture weight for topic $k$
$P(word|Topic_k)$—multinomial distribution over all words based on topic $k$.

In Fig. 8.16 $\theta$ describes document composition from topics, it determines how topic $Z$ is selected, $\beta_k$ describes terms contained in a given topic $k$.

It can be further simplified using the plate notation. With the plate notation, the dependencies among many variables can be captured concisely. Random variables are represented as circles, while fixed parameters as small boxes. The first step is observing the basic dependencies between variables (Fig. 8.17).

In the second step, we add indexes that describe all elements in detail. We also add boxes, the so-called "plates", to represent replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document. LDA in plate notation is presented in Fig. 8.18.

Variables are used as follows:

$D$—number of documents,
$N$—number of words in a document,
$\alpha$—the parameter of the Dirichlet prior on the per-document topic distributions,

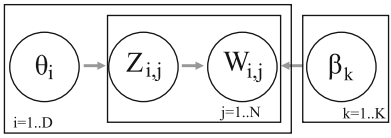



**Fig. 8.18** Plate notation for LDA

$\beta$—the parameter of the Dirichlet prior on the per-topic word distribution,
$\theta_i$—the topic distribution for document $i$,
$\phi_k$—the word distribution for topic $k$,
$z_{ij}$—the topic for the $j$th word in document $i$,
$w_{ij}$—the specific word ($j$th word in document $i$).

The $w_{ij}$ are the only observable variables, and the other variables are latent variables.

LDA can be confused with LSI—Latent Semantic Indexing, as the application area is indeed the same. There are, however, fundamental differences in the method designs. LSI uses a mathematical technique called singular value decomposition (SVD), and the space reduction is deterministic. In principle, words that are used in the same context tend to have similar meanings, and thus occupy the same "dimension". There is no need for simulation to find the best fit. Results of LSI are thus considered less precise than of LDA.

When thinking about applications, LDA can also be thought of as a clustering algorithm. In this interpretation, topics correspond to cluster centers and documents correspond to examples (rows) in a dataset. Topics and documents both exist in a feature space, where feature vectors are vectors of word counts. There is, however, a difference to the classical clustering approach: clustering uses a distance function and LDA uses a function based on a statistical model of how text documents are generated.

LDA is originally used for working with text; therefore, there are such notions as topic, word, and document. In this context, geographical profiling of BTS stations LDA can be applied in two settings:

- Endogenous (setting 1):
  - word—category from LGD, also referred to as term,
  - document—BTS station,
  - topic—how categories are generalized, it can also be interpreted as an intermediate summary.
- Exogenous (setting 2)—additional data concerning the usage of BTS by users is considered:
  - word—category from LGD,
  - document—profile of a user,
  - topic—BTS visited.

The exogenous approach is interesting as it depends on BTS usage. Such dependency can be a better indication of the real interest of users, thus leading to better their profiles.

### 8.5.4 BTS Profiling Results

In this section, we present the results concerning BTS profiling. We take the first settings of LDA application, i.e., the endogenous approach defined above.

Let us first observe real-world effect of TF-IDF weighting schema. Examples will be studied on *ways*—both filtered and full term space as well as binary and weighting schema will be considered.

Figure 8.19 compares the effect of two weighting schemes for specifying term occurrences. For *location 1* it is particularly important to distinguish the frequencies of different objects as there are so many of them. By looking at the term frequencies, it becomes apparent that *location 1* is devoted to *leisure*.

Figure 8.20 presents the weights obtained with TF-IDF transformation. We can observe now that *leisure* is not that important—it is the overall weight has been decreased by low discriminant power, i.e., many locations share this category. Leisure was popular in *location 1* because it is popular in the whole dataset.

Consideration of all categories is not always useful as some of the annotations carry no meaning. For example, there are some locations annotated as having *power lines* or *administrative boundaries* in the scope. This is not improving the accuracy and value of profiling. Therefore, as already mentioned, we decided to use an arbitrary set of 30 categories, removing unnecessary ones. In fact, we not only remove the terms, but as a consequence, some locations have to be removed as well as having no annotation. Figure 8.21 compares the term frequencies and their respective TF-IDF weights in the filtered set. The effect of TF-IDF weighting is visible for the terms ‘parking’ and ‘post office’ in location 4. Although ‘parking’ was 3 times more frequent, the final weight of ‘post office’ is significantly higher. Similar effect is observed for ‘hotel’ and ‘hospital.’

It is obvious that a certain reduction of space is necessary, but instead of relying on intuition, we can employ methods that can automatically suggest an appropriate mixture of categories. This is the goal of the introduced earlier latent semantic methods.

We first analyzed the effect of latent semantic indexing (LSI). We assumed that the term space is reduced from 30 dimensions to 5 dimensions. There is no specific interpretation of topics—the histograms should be perceived as a way to compare locations in a new reduced space. Figure 8.22 presents the topic weights in LSI. The left figure is based on the raw term frequencies and in this case locations 1 and 3 are the most similar. The right figure is based on the weights transformed with TF-IDF and here locations 4 and 5 are the most similar. We can conclude that the application of TF-IDF significantly influenced the LSI model, what is justified as the importance of certain categories is reduced or increased in the transformation.

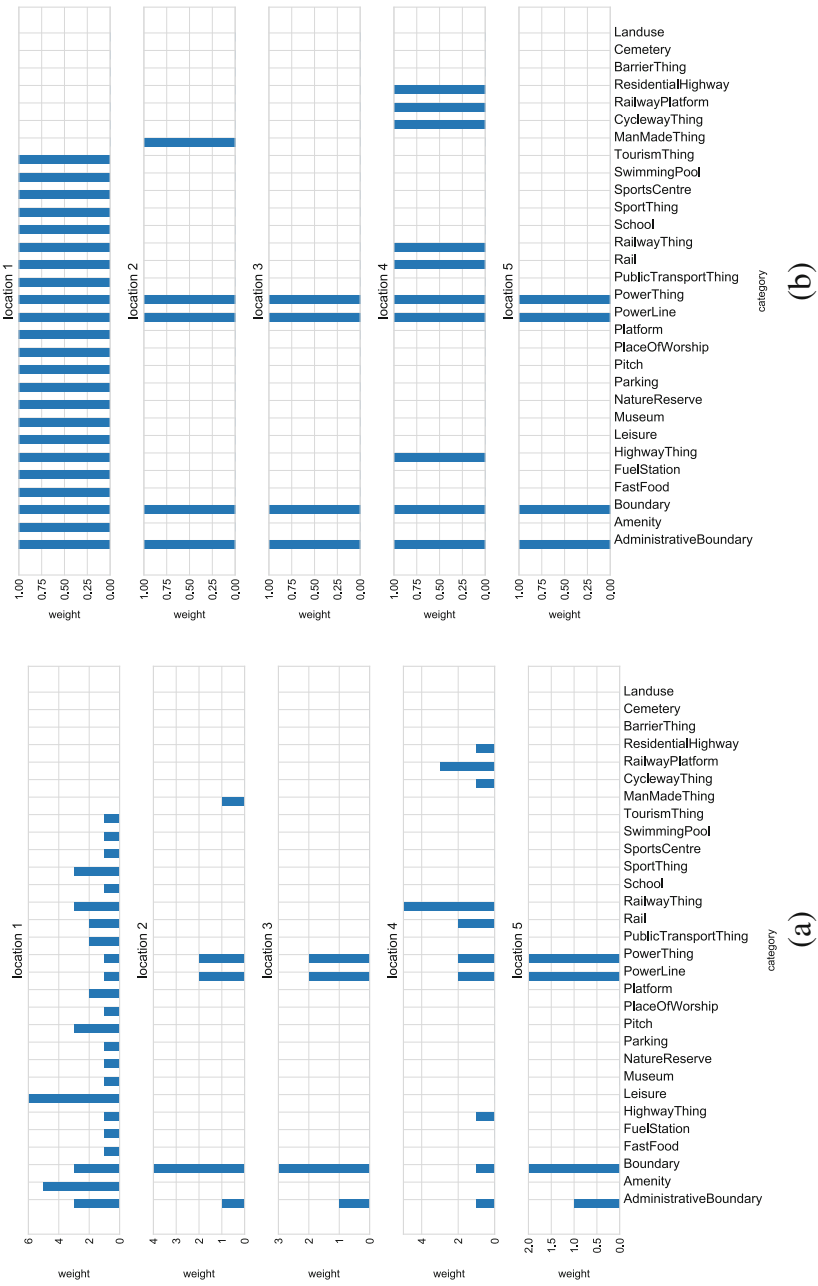



**Fig. 8.19** Term occurrences for sample locations in `ways_full` dataset. (**a**) Term frequencies. (**b**) Binary

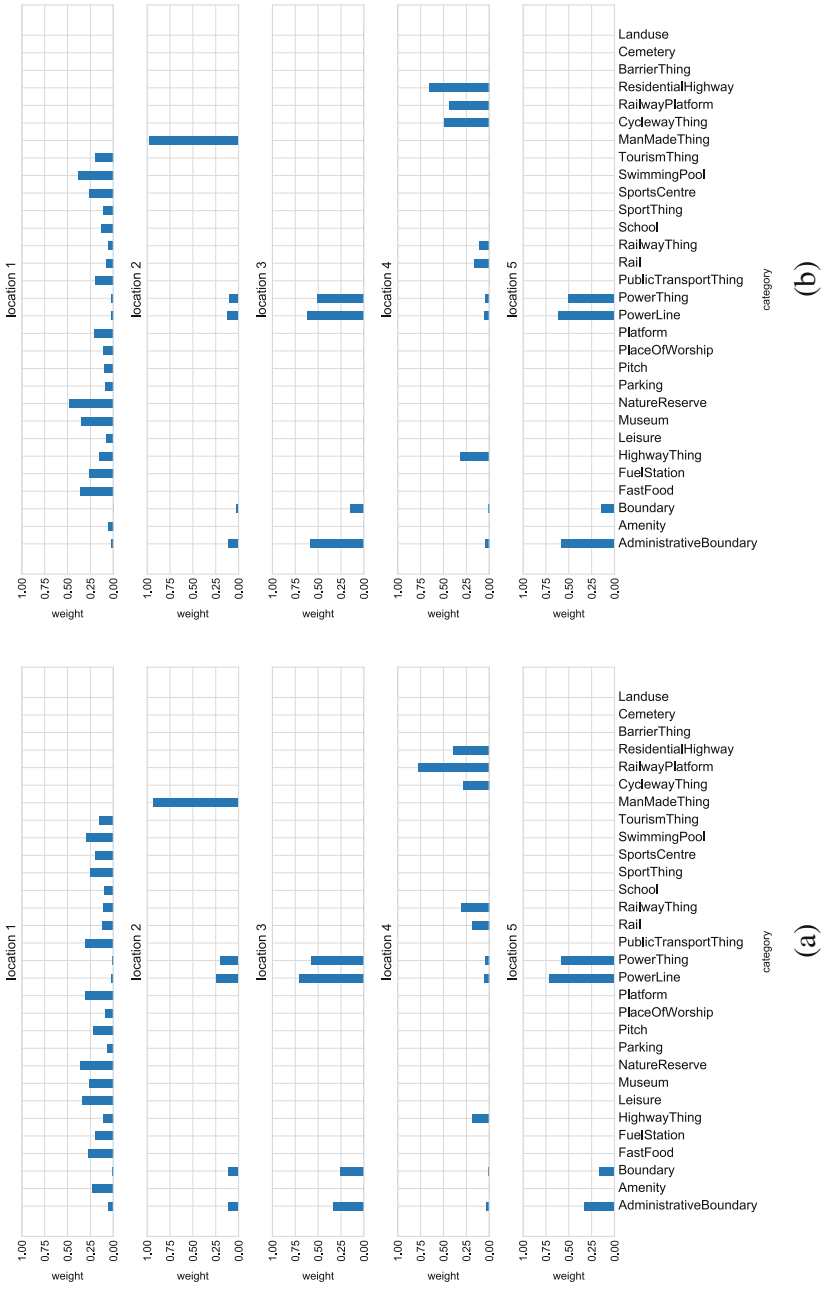



**Fig. 8.20** Term weights calculated with TF-IDF transformation for sample locations in `ways_full`. (**a**) Term frequencies. (**b**) Binary

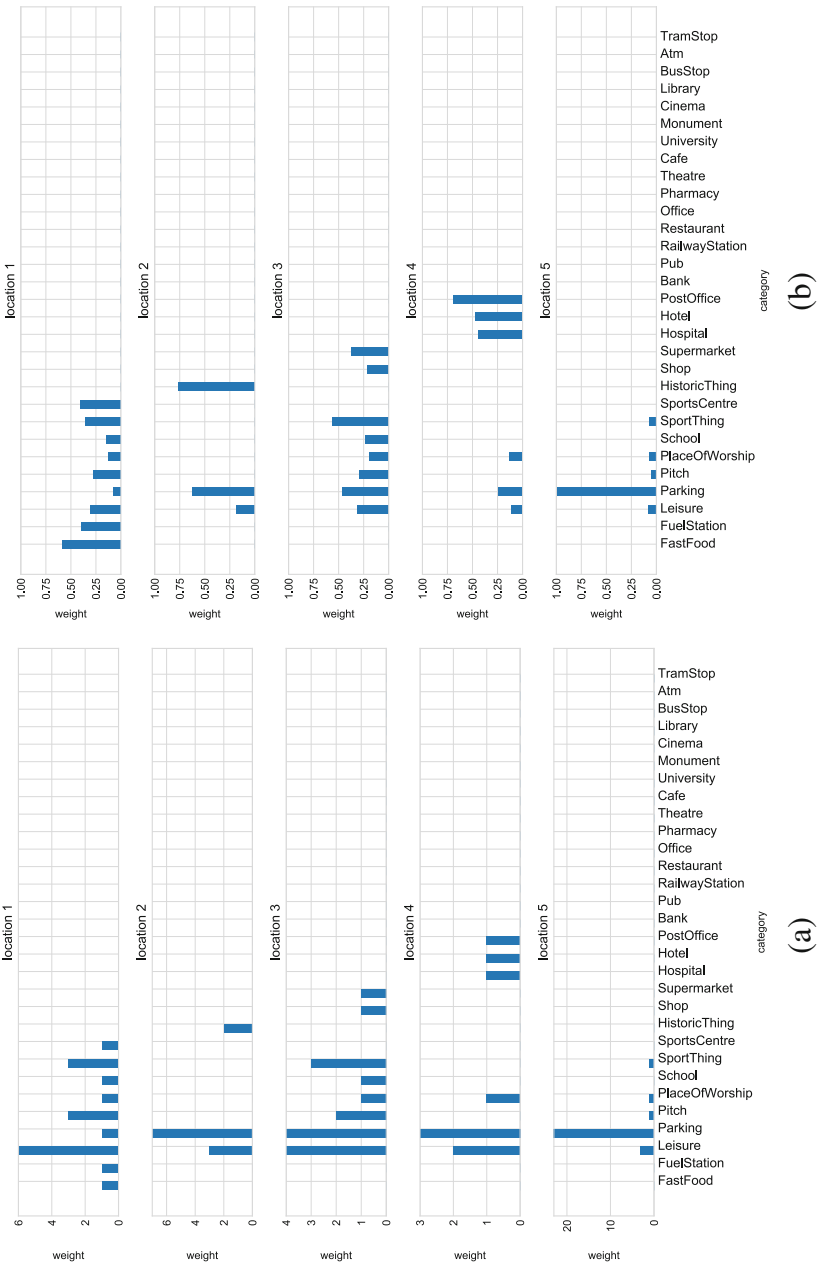



**Fig. 8.21** Term frequencies and weights for sample locations in `ways_filtered` dataset. (**a**) Term frequencies. (**b**) tfidf weights

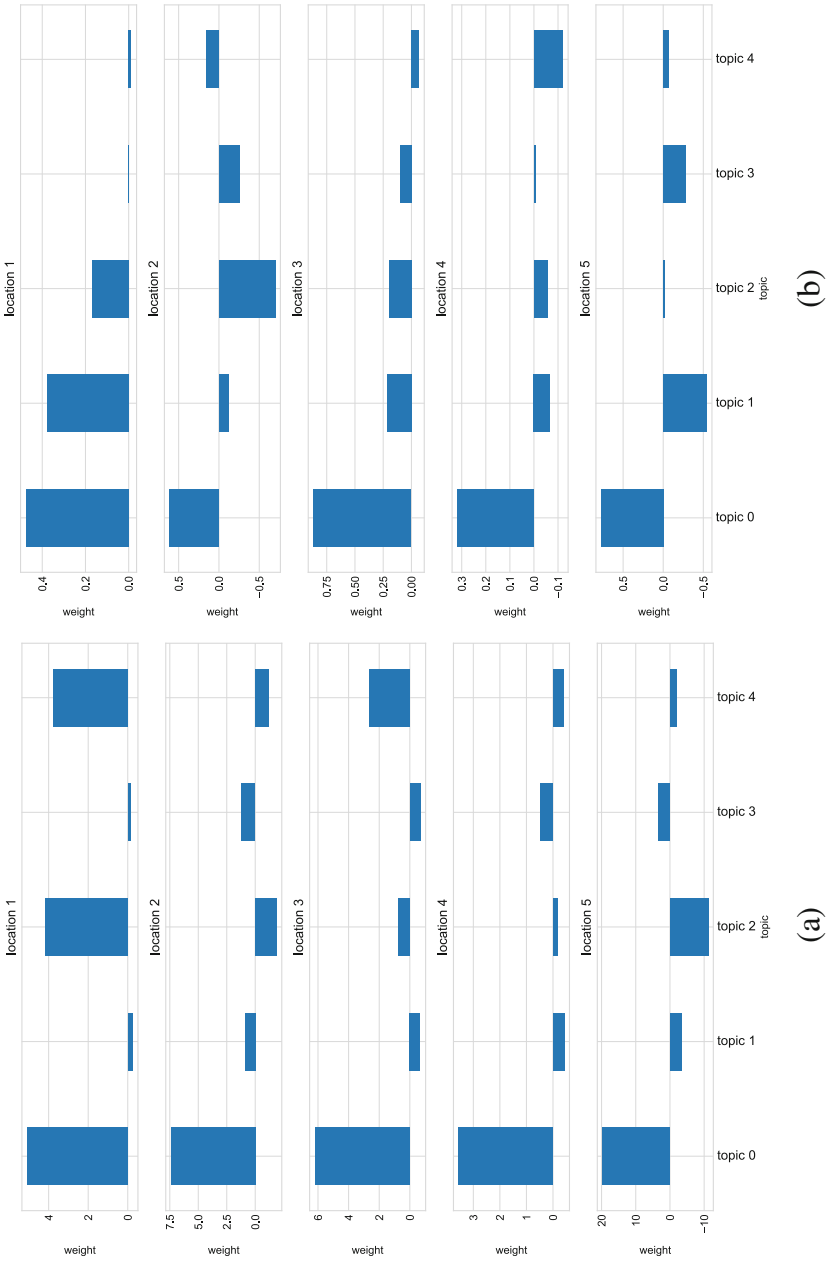



**Fig. 8.22** Topic weights for sample locations in `ways_filtered` dataset in LSI model. (**a**) On raw weights. (**b**) On tfidf weights

Another method for space reduction is latent Dirichlet allocation (LDA). Once more, a reduction to 5 dimensions was assumed. Figure 8.23 presents the same locations as previously but with LDA model space reduction. As can be observed, LDA is much more intuitive because there are no negative weights and the weights are normalized. Comparison of locations is then more intuitive. Differences between locations are also more visible than in the case of LSI.

It is also interesting to compare the weights before and after application of LDA and its impact on location similarity. Figure 8.24 presents the effect of category aggregation, which is discovered in an uncontrolled learning. Let us have a look at locations 3 and 5. It is hard to guess from the left figure if they are similar. Location 3 could be as well similar to location 1. Application of LDA model reveals a different picture—locations 3 and 5 are virtually identical, whereas location 1 is clearly distinct. It can be concluded that categories *school*, *shop*, and *supermarket* are not relevant for the distinction of locations.

This last issue has to do with the transformation of terms into topics. Not only locations but also a mixture of topics can be visualized, and the interpretation is more problematic (in most cases). Figure 8.25 compares the topic mixture characteristics built on raw term frequencies and TF-IDF weights. In the first case, it is easier to spot the topics' contents. Thus, *topic 1* is about historic things, *topic 2*—locations convenient for shopping, *topic 3*—addressing higher needs (places of worship, universities), *topic 4*—parking and finally *topic 5*—leisure, including sport. In the latter case, the topics consist of a bigger number of terms (only historical things in *topic 3* remained as something specific). We can conclude that TF-IDF weighting makes the interpretation of topics more difficult.

Other parameters have similar implications for the clarity of topic mixture interpretation. Binary weighting schema is also more dense (see Fig. 8.26a). This can be explained by equal weights (=1.0); hence, the topics also should be more homogeneously distributed. For the full term space (about 500 terms), the topics are also more dense (see Fig. 8.26b). This can be explained by the fact that more terms have to be reduced to only 5 dimensions. At the same time, significantly lower weights of single categories are observed. Choosing another number of topics would change the overall image.

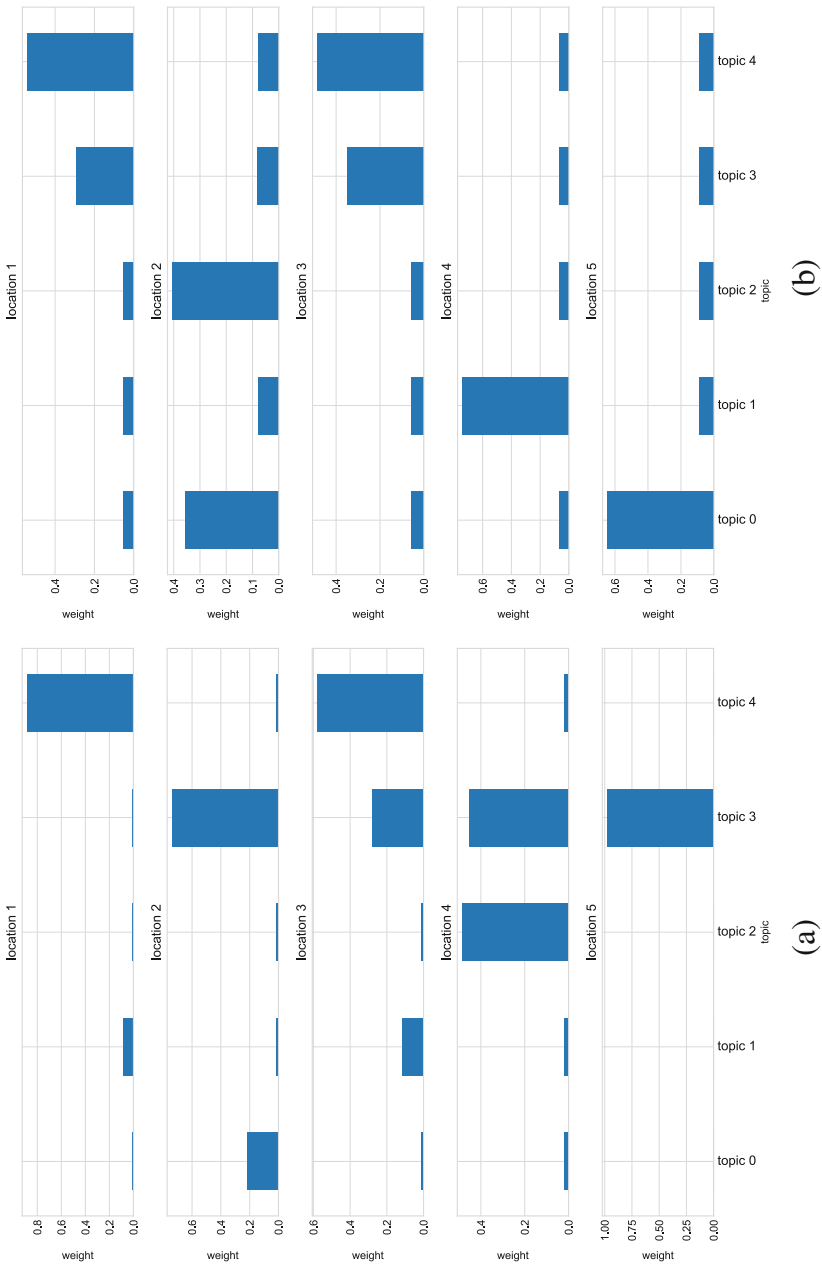



**Fig. 8.23** Topic weights for sample locations in `ways_filtered` dataset in LDA model. (**a**) On raw weights. (**b**) On tfidf weights

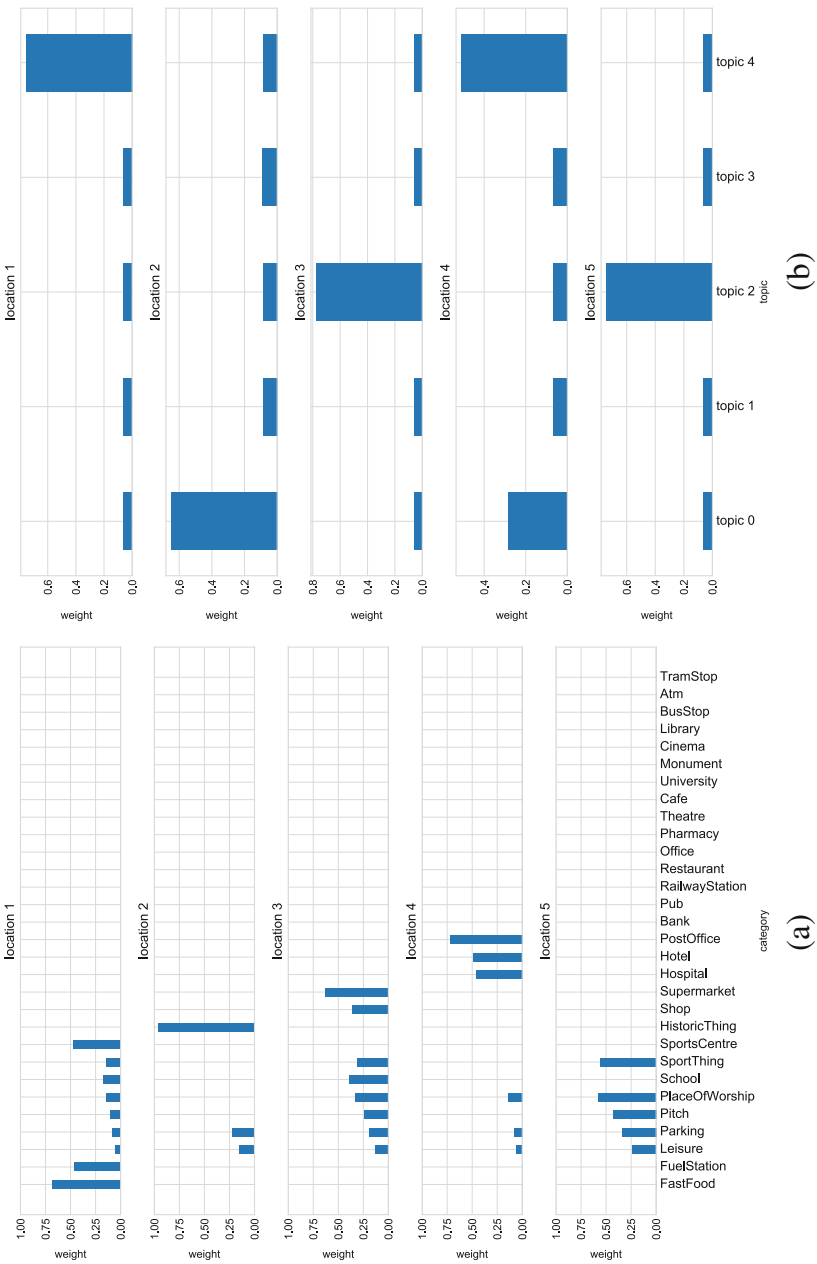



**Fig. 8.24** Term and topic weights for sample locations in `ways_filtered_bin` dataset in LDA model. (**a**) tfidf term weights. (**b**) LDA topic weights

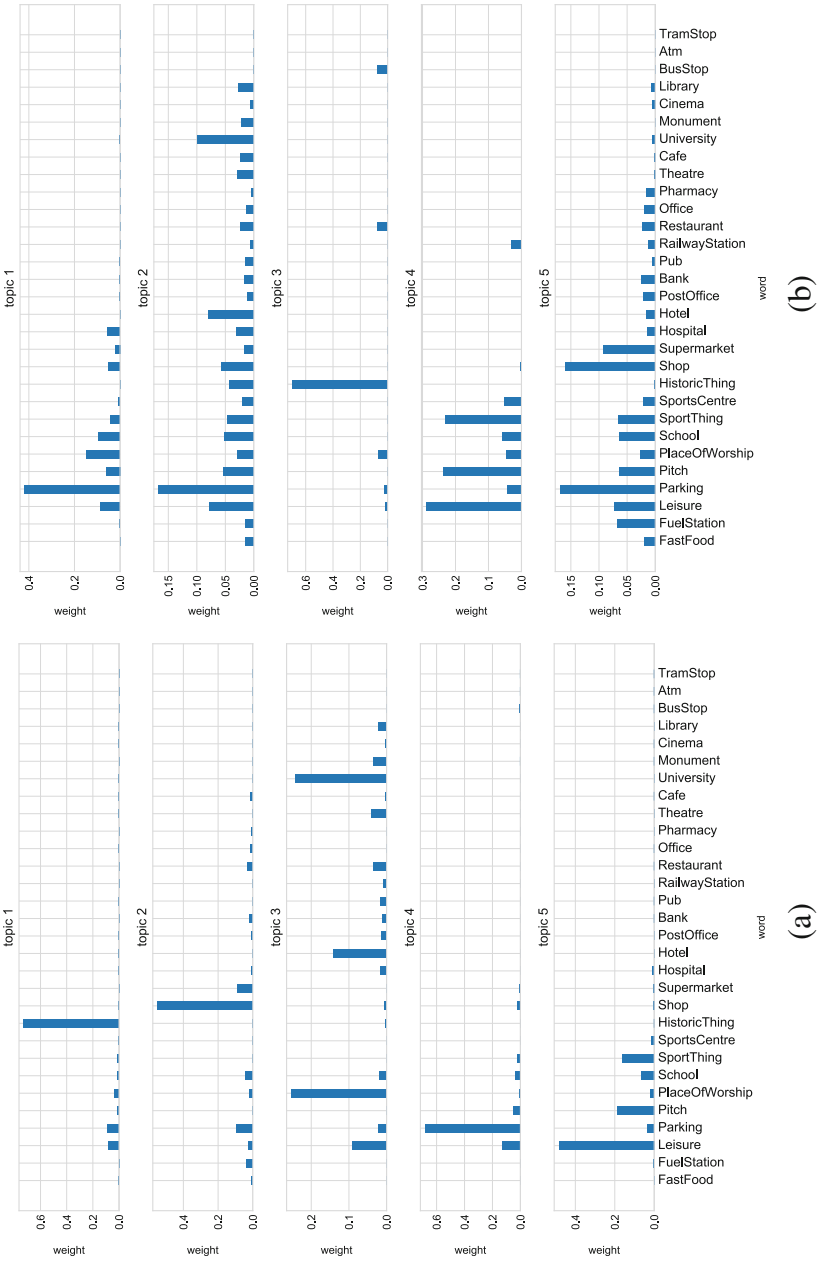



**Fig. 8.25** Topic mixture in LDA model for the corpus `ways_filtered`. (**a**) On raw term frequencies. (**b**) On tfidf weights

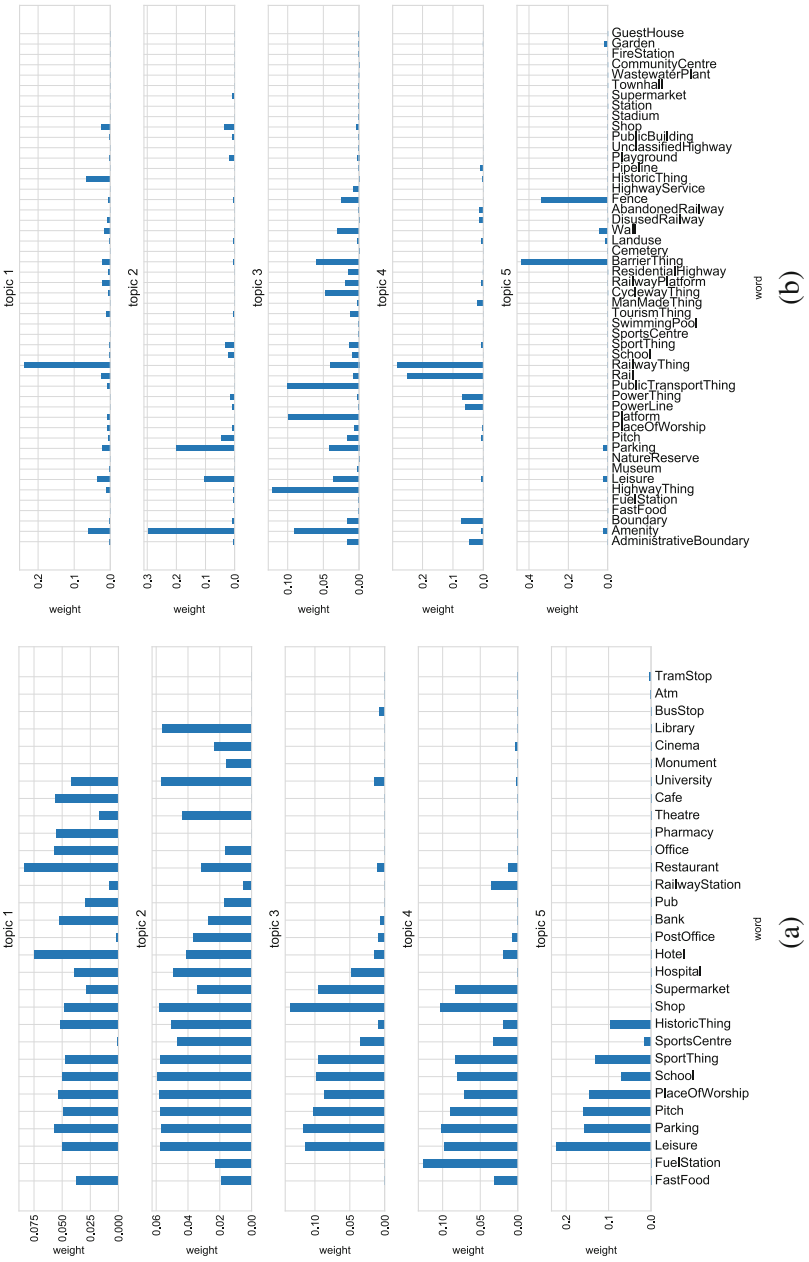



**Fig. 8.26** Topic mixture in LDA models based on term frequencies for various settings. (**a**) Binary weights, selected terms. (**b**) Term frequencies, all terms

### 8.5.5 User Profiling Results

User profiling is done by aggregating the profiles of visited BTS locations weighted with the number of actions performed within BTS (outgoing SMS, call). Therefore, for the final profile, additional data concerning the usage of BTS by users and various models of BTS profiling were considered: word—category from LGD; topic—BTS visited; and document—profile of a user.

For the purpose of demonstration this time we base on BTS profiles built using single-point geographical objects (=nodes). We start with the presentation of the simple profile based on term frequencies, presented in Fig. 8.27. There is a significant difference between binary and weighted schema profiles. Both profiles have certain drawbacks—we just present them as a baseline, motivating our work. The left figure shows the domination of frequent categories, e.g., *shop*. We also observe the high frequency of shops (35,000) for the third user. The right figure reveals the problem of binary weighting schema, i.e., assuming that we count only one shop even though there are 30 of them in the neighborhood. If one user is not moving around (i.e., connecting through one BTS), her profile cannot distinguish certain categories, e.g., *user 101354* has similar weights for all categories.

As in the case of BTS profiles, the importance of certain categories can be fine-tuned automatically by the application of TF-IDF weighting. Figure 8.28 shows the same users as above with profiles transformed with TF-IDF method. Here, the importance of shops is reduced by applying this weighting schema.

Let us now move on to space reduction models. Similarly to BTS profiles, we will analyze user profiles not in the term space (30 dimensions for the filtered variant) but in the topic space (5 topics have been assumed for learning, hence 5 dimensions). Figure 8.29 presents the latent semantic indexing model. Once more, by analogy to BTS profiles, also these profiles are hard to interpret—big or negative weights. Negative weights can reduce the impact of other BTS profiles. Additionally, it is hard for the human eye to distinguish users, particularly in the right chart.

The most sophisticated user profiles are built using LDA models, presented in Fig. 8.30. LDA much better reflects the differences between users. We also observe that in the case of LDA it is necessary to apply TF-IDF transformation. Unfortunately, without TF-IDF the dominating effect of the *shop* category is visible—*topic 3* has very high score in all user profiles.

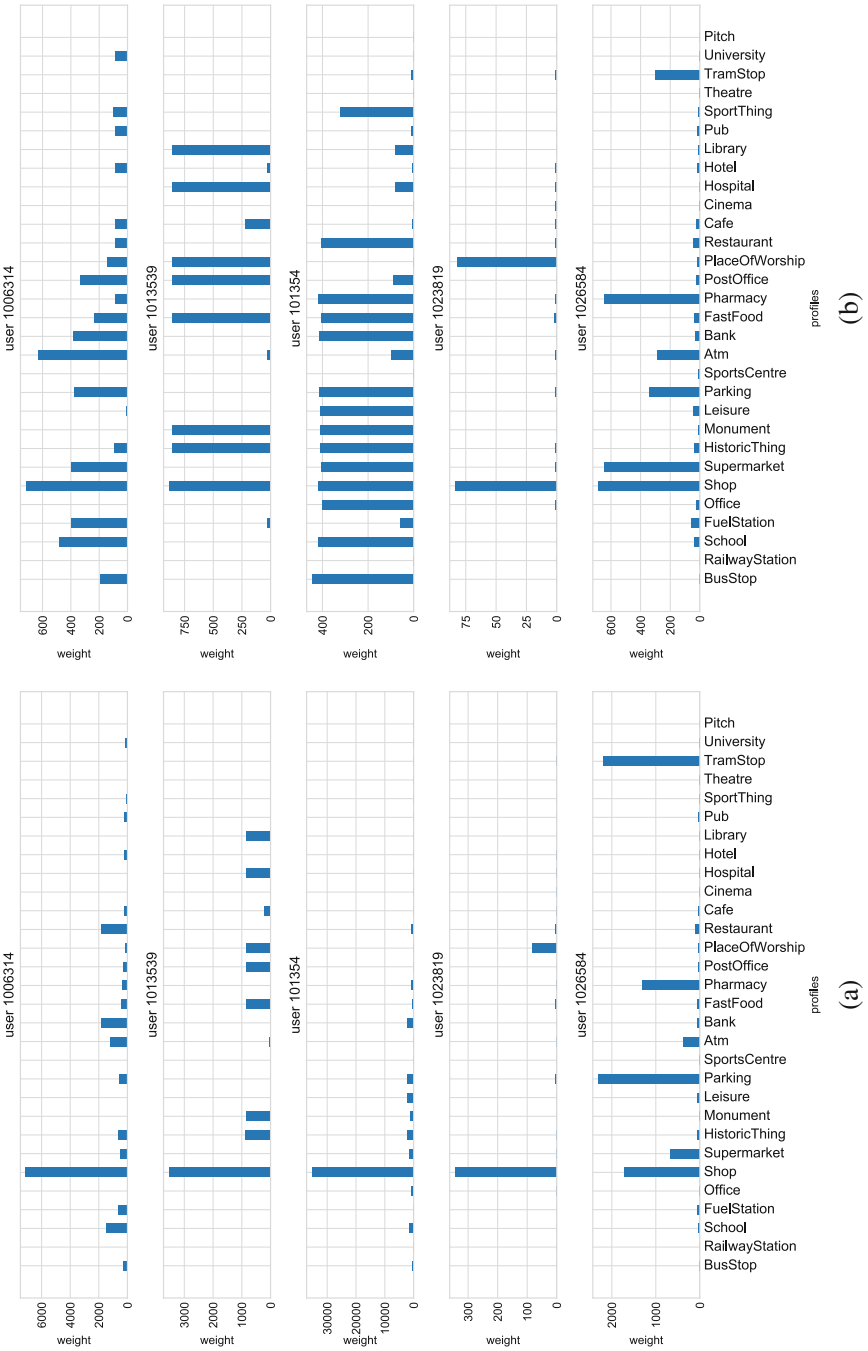



**Fig. 8.27** Profile of users calculated on `nodes_filtered` dataset (no tfidf transformation). (**a**) Term frequencies. (**b**) Binary weights

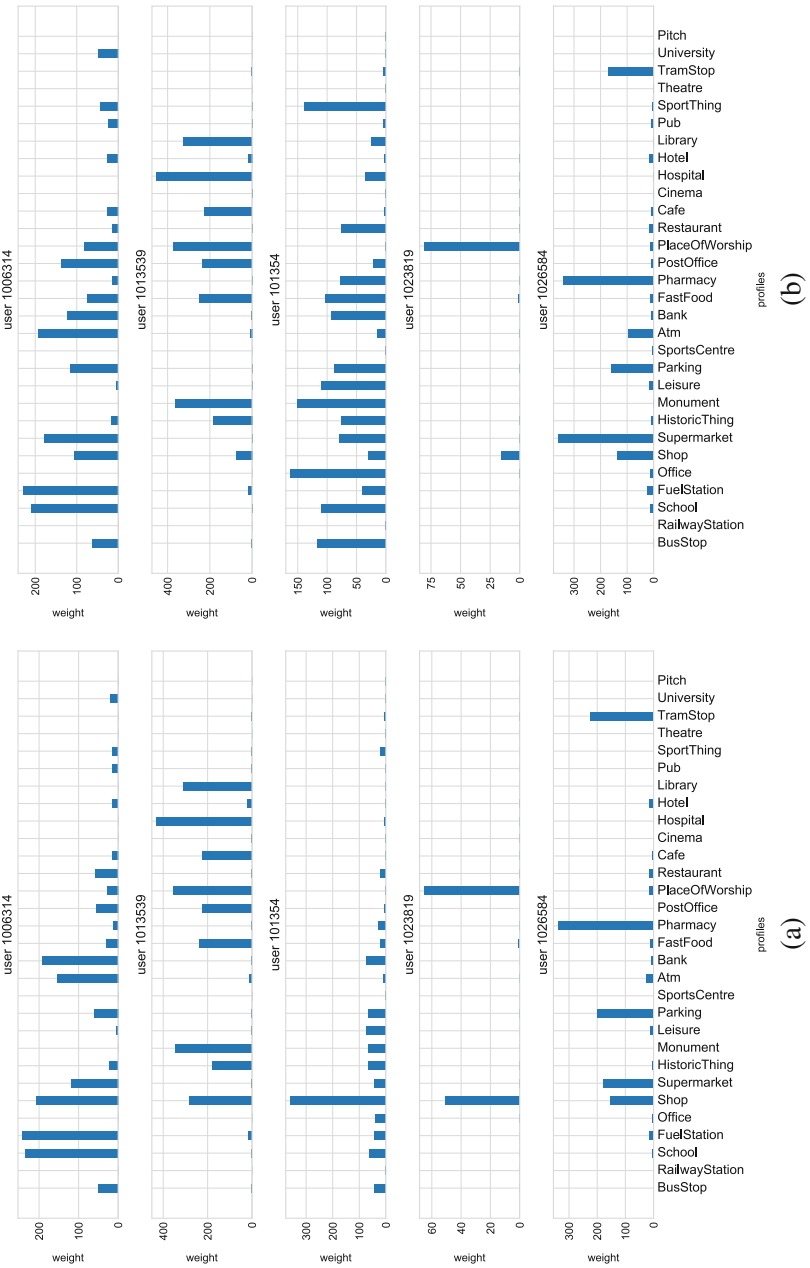



**Fig. 8.28** Profile of users calculated on `nodes_filtered` dataset with tfidf weighting. (**a**) Term frequencies. (**b**) Binary weights

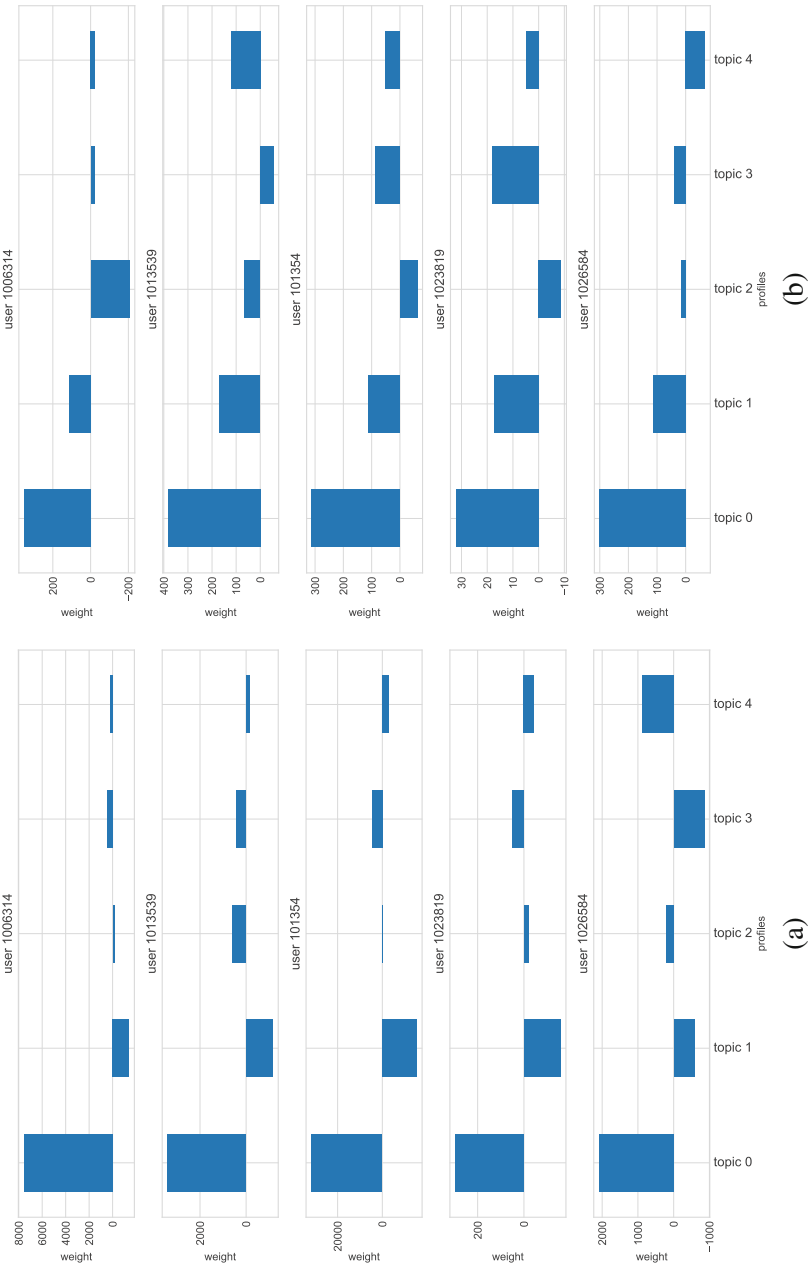



**Fig. 8.29** Profile of users calculated on `nodes_filtered` dataset with LSI model. (**a**) Raw counts. (**b**) tfidf weigths

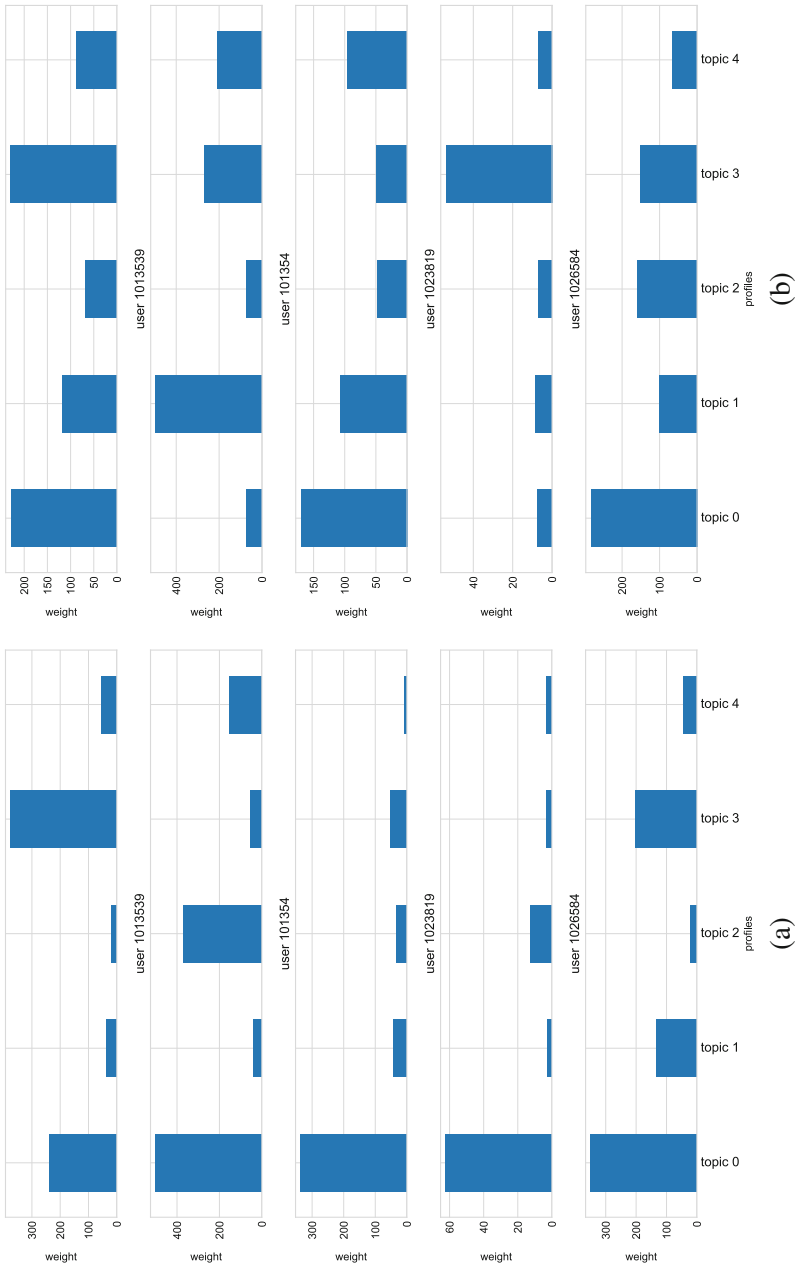



**Fig. 8.30** Profile of users calculated on `nodes_filtered` dataset with LDA model. (**a**) Raw counts. (**b**) tfidf weigths

## 8.6 Summary

Throughout this section, we described different variants for modeling BTS and user profiles. Of the variants presented at the beginning, both nodes and ways provide valuable data. Term frequency is preferred over binary indexing, as the latter is to some extent reducing available information. It is advisable to apply TF-IDF transformation as it reduces the overall impact of frequent terms on profiles, and it also alleviates the differences in binary and weighted schemes. Datasets with predefined sets of categories are preferred for better interpretation of results. Full term space should not be used unless a space reduction technique, such as LDA is applied. Of the two models, LDA is preferred over LSI—mostly for a more intuitive interpretation of the results. Moreover, the results seem to be better although learning takes much more time. It is important to allow TF-IDF preprocessing for LSI and LDA. We have not observed significant differences for user comparison between hand-picked categories and full-term space reduced with LDA.

An interesting direction of the research seems to be the profiling of residents. It is especially important for geo-marketing purposes. For such an analysis, additional information can be utilized—home location of mobile users, which itself can be determined from the user behavior reflected in CDR.

## References

Abel, F., Hauff, C., Houben, G. J., & Tao, K. (2012). Leveraging user modeling on the social web with linked data. In M. Brambilla, T. Tokuda, R. Tolksdorf (Eds.), *Web Engineering: 12th International Conference ICWE 2012, Berlin, Germany July 23–27, 2012. Proceedings* (pp. 378–385). Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-642-31753-8. https://doi.org/10.1007/978-3-642-31753-8_31 (page 217)

Abramowicz, W. (2008). *Filtrowanie informacji* [Information Filtering]. Poznań: Wydawnictwo Akademii Ekonomicznej. ISBN: 9788374173155. (page 225)

ACIL. (2011). *The value of spatial information for Tasmania*. Department of Premier and Cabinet. (pages 216, 217)

Center for Open Data Enterprise. (2016). *Open data impact map*. (page 217)

Cunningham, P. (2007). *Dimension reduction* Technical Report UCD-CSI-2007-7. University College Dublin. (pages 227, 231)

de Montjoye, Y., Smoreda, Z., Trinquart, R., Ziemlicki, C., & Blondel, V. D. (2014). D4D-senegal: The second mobile phone data for development challenge. CoRR arXiv: abs/1407.4885 [cs.CY]. (page 219)

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., Blondel, V. D., & Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences, 111*(45), 15888–15893. https://doi.org/10.1073/pnas.1408439111. Eprint: http://www.pnas.org/content/111/45/15888.full.pdf (page 219)

DuVander, A. (2010). *5 years ago today the web mashup was born*. https://www.programmableweb.com/news/5-years-ago-today-web-mashup-was-born/2010/04/08 (visited on 2017-12-08). (page 215)

European Commission (2004). *Proposal for a decision of the European Parliament and of the Council establishing a multiannual Community programme to make digital content in Europe more accessible usable and exploitable. SEC(2004)169, COM(2004) 96*. http://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:52004PC0096. (pages 216, 217)

Fornefeld, M., Boele-Keimer, G., Recher, S., & Fanning, M. (2008). *Assessment of the re-use of public sector information (PSI) in the geographical information, meteorological information and legal information sectors final report* MICUS Management Consulting GmbH. (page 216)

Grant, A., Razdan, R., Shang, T. (2014). Coordinates for change: How GIS technology and geospatial analytics can improve city services. In *Innovation in local government. Open data and information technology* (pp. 32–43). McKinsey&Company. (page 215)

Hart, G., & Dolbear, C. (2013). *Linked data: A geographic perspective* (p. 290). CRC Press. ISBN: 9781439869956. (page 217)

Jahani, E., Sundsøy, P., Bjelland, J., Bengtsson, L., Pentland, A., & de Montjoye, Y. A. (2017). Improving official statistics in emerging markets using machine learning and mobile phone data. *EPJ Data Science, 6*(1), 3. ISSN: 2193–1127. https://doi.org/10.114/epjds/s13688-017-0099-3 (page 220)

Koski, H. (2011). Does marginal cost pricing of public sector information spur firm growth? *Keskusteluaiheita Discussion Papers* 1260. (page 216)

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute. (page 215)

Miritello, G., Moro, E., & Lara, R. (2011). Dynamical strength of social ties in information spreading. *Physical Review E, 83*(4), 045102. https://doi.org/10.1103/PhysRevE.83.045102 (page 219)

Naef, E., Muelbert, P., Raza, S., & Frederick, R. (2014a). *Mobile data for development*. Cartesian and Bill & Melinda Gates Foundation. (pages 218, 219)

Naef, E., Muelbert, P., Raza, S., Frederick, R., Kendall, J., & Gupta, N. (2014b). *Using mobile data for development*. Cartesian and Bill & Melinda Gates Foundation. (page 219)

OECD. (2013). Exploring data-driven innovation as a new source of growth: Mapping the policy issues raised by "Big Data. *OECD Digital Economy Papers, 222*, 1–44. ISSN: 2071-6826. https://doi.org/10.1787/5k47zw3fcp43-en (page 216)

Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Martino, M., Claxton, R., & Strogatz, S. H. (2010). Redrawing the map of Great Britain from a network of human interactions. *PLOS ONE, 5*(12), 1–6. https://doi.org/10.1371/journal.pone.0014248 (page 219)

Toole, J. L., de Montjoye, Y. A., González, M. C., Pentland, A. (2015). Modeling and understanding intrinsic characteristics of human mobility. In B. Gonçalves, N. Perra (Eds.), *Social phenomena: From data analysis to models* (pp. 15–35). Cham: Springer International Publishing. ISBN: 978-3-319-14011-7. https://doi.org/10.1007/978-3-319-14011-7_2 (page 219)

Węcel, K. (2015). Linked geodata for profiling of telco users. *Studia Ekonomiczne. [Economic studies], 234*, 199–213. ISSN: 2083-8611 (pages 225, 228)

Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., & Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. *Science, 338*(6104), 267–270. ISSN: 0036-8075. https://doi.org/10.1126/science.1223467. eprint: http://science.sciencemag.org/content/338/6104/267.full.pdf (page 219)

Wesolowski, A., Qureshi, T., Boni, M. F., Sundsøy, P. R., Johansson, M. A., Rasheed, S. B., Engø-Monsen, K., & Buckee, C. O. (2015). Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proceedings of the National Academy of Sciences, 112*(38), 11887–11892. https://doi.org/10.1073/pnas.1504964112. eprint: http://www.pnas.org/content/112/38/11887.full.pdf (page 219)

# Chapter 9
# Conclusions


The massive growth of data created and consumed by enterprises suggests that data is becoming a new factor of production. Growing investments in data management and analytics partly reflect the increasing economic role of data.

Throughout the book, we have made a journey from data resources to concrete applications. We first learned the main pillars of the work: open data, linked data, and big data. Open data is crucial as a data supply, especially concerning macroeconomic information. Linked data is currently the main asset for reuse, which is particularly relevant for building (enterprise) knowledge graphs. Big data needs to be tamed not only from a technological perspective but also from a business relevance perspective. The proposed linked-data-based unification model was verified by referring to the 4Vs of big data and demonstrating how the challenges have been addressed.

Our particular attention was devoted to value. We studied it from two perspectives: macroeconomic and microeconomic. The first justified intervention on a government level, and the latter called for actions to be taken by stakeholders within the open data ecosystem. The value of data not only depends on time, as we have demonstrated, and the following phenomena were determined as relevant for providing results and value to the economy:

- **Data linkage**—The same datasets can lead to different information depending on their structure, including their links to other datasets. Therefore, the value can also differ. This is a strong argument for the adoption of linked data.
- **Data analytic capacities**—The value of data depends on the extracted or interpreted meaning. Once more, the same datasets can thus lead to different information depending on the analytic capacities, skills, available techniques, and technologies for data analysis. Here, big data plays the main role, along with hardware and algorithms supporting analytics.

The above can also be summarized in business language. Big data contributes to increasing returns to scale—the accumulation of data can lead to improvements



K. Węcel, *Big, Open and Linked Data*, Business Information Systems,
https://doi.org/10.1007/978-3-031-07147-8_9

in data-driven services (positive feedback). Linked data contributes to increasing returns to scope—contextualization of data is a source for insights that are greater than the sum of isolated parts (super-additive value).

Value not only needs to be created in organizations, it also has to be captured and transferred properly, which is the subject of the study of business models. An interesting contribution here was a systematic literature review to identify business models used in the data domain and on the Web.

At the end of the book, we find an empirical verification of the proposed data unification model, where open and linked geographical data was combined with big data of a telecommunications company in order to profile telco subscribers, allowing innovative business models. Overall, it also added to the conclusion that innovation is a main vehicle for gaining value from the unification of big, open, and linked data.

The leitmotiv of our considerations was the hypothesis that application of linked data can help overcome many problems of today's enterprises, and for its deployment, it is necessary to apply a modern methodological approach. After reading the book, you can see the approaches to value creation and value capture from different types of data. The findings from cognitive, methodological, and empirical studies help us formulate recommendations for the effective building and deployment of solutions combining big and open data with the mediation of linked data. In a business practice, we are able to propose solutions providing methodological foundations for better understanding of business models and their development for different kinds of data assets. Thus, companies can create and improve their competencies and, as a result, build a competitive advantage.

In the **epistemological layer**, we learned of the concept of big and open data unification catalyzed by linked data. The focus was on resources to reuse along with discovery procedures. We studied the advantages of sharing, along with data ownership and privacy issues, thus coming up with theoretical dimensions of value creation. We also gained knowledge of a typology of business models with reference to potential use for linked-data-based applications.

In the **methodological layer**, we learned of a model for linked-data-based unification of data. We became familiarized with the aspects and methods to analyze the value of data both from the microeconomic and macroeconomic point of view. Among the tools, we can count ontologies as ones to be used as an integration layer (among others for disambiguation) and business models for value creation and capture.

In the **empirical layer**, we verified the theoretical model for data unification. First, we studied the potential of leveraging linked data for contemporary modern systems. Second, we applied geographical linked open data to take advantage of big data in support of profiling users in the telecommunications industry.

The scope of linked data should not be restricted to corporate or government circles. An interesting research question is how useful linked data is for the average user, as holistic understanding of the value creation process is required. Further developments should be supported with revenue generated using the technologies in question. Individuals should be aware as both consumers and providers of data. Quality of data will only be improved if people observe that this affects them personally.