FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Open Data Quality

Assessment and Evolution of (Meta-)Data Quality in the Open Data Landscape

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Computational Intelligence

eingereicht von

Sebastian Neumaier
Matrikelnummer 0925308

an der Fakultät für Informatik
der Technischen Universität Wien

Betreuung: Univ.Prof. Dr. Axel Polleres
Zweitbetreuung: Dr. Jürgen Umbrich, WU Wien

Wien, 1. Oktober 2015

Sebastian Neumaier	Axel Polleres

FAKULTÄT
FÜR !NFORMATIK
Faculty of Informatics

Open Data Quality

Assessment and Evolution of (Meta-)Data Quality in the Open Data Landscape

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Computational Intelligence

by

Sebastian Neumaier
Registration Number 0925308

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Univ.Prof. Dr. Axel Polleres
Co-Advisor: Dr. Jürgen Umbrich, WU Wien

Vienna, 1st October, 2015

Sebastian Neumaier

Axel Polleres

Erklärung zur Verfassung der Arbeit

Sebastian Neumaier
Phorusgasse 2, 1040 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 1. Oktober 2015

Sebastian Neumaier

Acknowledgements

First of all, I want to thank my co-advisor Jürgen Umbrich with whom I have been working closely and successfully on this project. I would like to take this opportunity to thank him for the chance to participate in this project, for his guidance and his motivating approach. I very much enjoyed the productive and interesting cooperation over the past year.

Many thanks also go to my advisor Prof. Axel Polleres for reviewing the work, raising issues and improvements and especially for investing a lot of time during the final stage of this thesis.

Thanks to all my friends for their help and support during stressful periods and finally, a big thank you to my parents for supporting me throughout my entire course of study.

Kurzfassung

Die Open-Data-Bewegung erfreut sich wachsender Beliebtheit unter Regierungen und öffentlichen Institutionen, aber auch in der Privatwirtschaft und unter Privatpersonen, und gewinnt so immer mehr Unterstützerinnen und Unterstützer aus all diesen Sektoren. Gleichzeitig melden sich aber auch vermehrt kritische Stimmen zu Wort. Hauptsorge ist die niedrige Metadaten-Qualität in Open Data Portalen, die eine Beeinträchtigung der Suche und der Auffindbarkeit von Ressourcen mit sich bringt.

Bis jetzt konnte diese Sorge jedoch nicht belegt werden, da es noch keinen umfassenden und objektiven Bericht über die wirkliche Qualität von Open Data Portalen gibt. Um so einen Bericht erstellen zu können, wird ein Framework benötigt, welches die Portale über einen längeren Zeitraum hinweg beobachtet und so die Entwicklung und das Wachstum von Open Data abschätzen kann.

Die vorliegende Diplomarbeit hat das Ziel diese Qualitätsprobleme in Open Data Portalen zu untersuchen. Dazu wird ein Monitoring Framework vorgestellt, welches in regelmäßigen Abständen die Metadaten von 126 CKAN Portalen speichert und deren Qualität bewertet. Die Arbeit stellt die dazu notwendigen Qualitätsmetriken vor, diskutiert den Aufbau des Monitoring Frameworks und präsentiert Erkenntnisse und Resultate, die aus dem Monitoring der Portale gewonnen werden konnten. Dazu werden Auswertungen der eingeführten Qualitätsmetriken prästeniert, die auf Qualitätsprobleme in den untersuchten Datenportalen hinweisen. Konkret konnte unter anderem ein schnelles Wachstum von diversen Open Data Portalen und eine hohe Heterogenität bezüglich der Datenformate und Lizenzen beobachtet werden.

Darüberhinaus wird in dieser Arbeit ein Ansatz zur Homogenisierung von Metadaten von unterschiedlichen Datenportalen vorgestellt: Dazu wird ein Mapping vorgestellt, welches die Metadaten von CKAN, Socrata und OpenDataoft Portalen auf ein gemeinsames Schema bringt und damit die Portale vergleichbar und integrierbar macht.

Abstract

While the Open Data movement enjoys great popularity and enthusiasm among governments, public institutions and also increasingly in the private sector, first critical voices start addressing the emerging issue of low quality of metadata and data sources in Open Data portals with the risk of compromising searchability and discoverability of resources.

However, there neither exists a comprehensive and objective report about the actual state and quality of Open Data portals, nor is there a framework to continuously monitor the evolution of these portals. The present thesis tries to fill this gap.

More concretely, in this work we present our efforts to confirm – or refute – various quality issues in Open Data by monitoring and assessing the quality of 126 CKAN data portals. We define our quality metrics, introduce our automated assessment framework and report comprehensive findings by analyzing the data and the evolution of the portals. We confirm the fast evolution of Open Data, pinpoint certain quality issues prevalent across the portals, and include insights about heterogeneity in Open Data such as the diversity of file format descriptions and the licensing of datasets.

Another contribution of this thesis is an approach towards the homogenization of metadata found on different data publishing frameworks: we propose a common mapping for metadata occurring on CKAN, Socrata and OpenDataSoft portal software frameworks in order to improve the comparability and interoperability of portals running these different software frameworks.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

"Data! Data! Data!" he cried impatiently.
"I can't make bricks without clay."

— Sir Arthur Conan Doyle, *The Adventure of the Copper Beeches*

As of today, the Open Data movement enjoys great popularity among governments, public institutions and also – increasingly – industry by promising transparency for citizens, more efficient and effective public services and the chance to outsource innovative use of the published data [JCZ12]. However, first critical voices start addressing – to the public – the emerging issue of low quality of metadata and data sources in data portals, which is a serious risk that could disrupt the success of the Open Data project.[1]

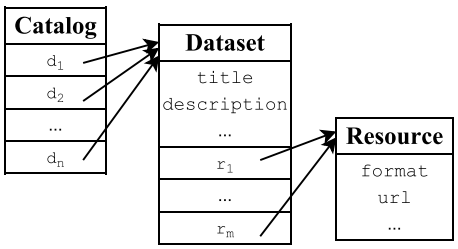

Figure 1.1: Structure of a Data Catalog.

These so called data portals, as depict in Figure 1.1, are catalogs which serves as a single point of access for a number of datasets. Generally speaking, a dataset is just a collection of data. In the context of data portals, a dataset aggregates a group of data files (referred to as resources) which are available for access or download in one or more formats, e.g., CSV, PDF and Microsoft Excel. The resources themselves either can be hosted on the corresponding portal or can be referenced from external data sources. Additionally, a dataset holds metadata (i.e., basic information) of these resources, e.g. authorship, provenance or licensing information. Data portals of this kind usually offer an API to access and retrieve the datasets and (if available) their resources.

[1] http://www.business2community.com/big-data/open-data-risk-poor-data-quality-0101053, last accessed 2015-07-21

As already mentioned, there is the identified risk of low (meta-)data quality in data portals. This risk directly impairs the discovery and consumption of a dataset in a single portal and across portals. On the one hand, missing metadata affects the search and discovery services to locate relevant and related datasets for particular consumer needs. On the other hand, incorrect descriptions of the datasets pose several challenges for their processing and integration with other datasets.

Examples of the risk introduced by low metadata quality in data portals are that incomplete or incorrect metadata significantly influences data discovery services of the portals but also the efficient (re-)use of the data by consumers (e.g., because of missing license or contact information or incorrect data source descriptions). In addition to incomplete or incorrect metadata, the potential heterogeneity across different aspects of Open Data portals poses several challenges if one wants to discover and combine related datasets. For instance, different licenses, data formats and availability of metadata keys substantially impede the integration of various data sources and therefore decrease the profitableness of Open Data.

Complementary to the metadata quality issues is the data quality problem. This is a diverse topic on its own and covers problems across several steps in the publishing process, such as the use of non-standard units, formats or unknown aggregations in the census process of the information or format inconsistencies in the transformation process into an open format. However, addressing this problem is out of scope of this thesis and can be considered as one of the major upcoming challenges.

1.1 Motivation

While the Open Data movement is currently taking the next step and is moving towards opening up data from the private sector,[2] we observe the risk of isolated "data silos" due to missing data integration and decreasing data quality within the catalogs. Assuming that there is the arising issue of low quality of metadata and data sources in data portals, we can recapitulate the aforementioned concerns which are able to impair the Open Data movement: A compromised searchability and discovery of data in data catalogs, an impeded integration of different datasets and data sources, as well as an reduced usability of resources.

Furthermore, different types of heterogeneity affect the interoperability of different data sources. This includes the heterogeneity of metadata, the heterogeneity of data formats, as well as the heterogeneity of licenses and taxonomies in data catalogs.

1.2 Problem Statement

The introduced quality issues are common to any search and data integration scenario and, to the best of our knowledge, there neither exist comprehensive, quantitative and

[2] `http://blogs.worldbank.org/voices/next-frontier-open-data-open-private-sector`, last accessed 2015-08-16

objective reports about the actual quality of Open Data portals, nor is there a framework to continuously monitor the evolution of Open Data in portals.

In an effort to confirm – or refute – the stated quality concerns, this thesis aims to critically examine, report and possibly improve the current status of Open Data. In the scope of this thesis, achieving this objective involves, (i) an extensive review of the literature of quality assessment methodologies, resulting in (ii) a set of objective quality measures which are suitable to detect and assess the issues in an automated way. Further, (iii) a framework is needed to automatically retrieve the datasets from the portals and calculate the proposed metrics. Due to the high heterogeneity of metadata from different data publishing frameworks, we mainly focus our monitoring and quality assessment on portals using the CKAN software. Eventually, the framework is utilized to (iv) provide comprehensive reports of the current status and actual quality of CKAN Open Data portals. Additionally, in order to tackle the above mentioned heterogeneity issue, (v) we investigate different metadata homogenization approaches. Therefore, we discuss a homogenization of metadata from different software frameworks.

Hypothesis. This thesis wants to support and contribute to a rational and well-founded spreading of the Open Data movement by addressing the following questions:

Is Open Data what it claims to be? A large-scale status report of data published under the term "Open Data" allows us to evaluate if the current state of Open Data follows the theoretical definition.

Is there a quality issue in Open Data? A framework for automated quality assessment based on a set of objective quality metrics enables us to discover, point out and measure quality and heterogeneity issues in data portals.

How is Open Data evolving? A continuous monitoring of data catalogs allows us to estimate the development and growth-rate of Open Data.

Do improvement initiatives affect (meta-)data quality on data catalogs? An automated monitoring framework is able to detect and measure improvement methods and initiatives. Further, it is able to report the impact and distribution rate of metadata homogenization efforts.

Approach. Below, we introduce the research approach to verify these hypotheses, consisting of the following individual sub-goals:

G1: As a first goal, we introduce the necessary underlying terms and concepts. We introduce the *Open Definition*, name sources of Open Data and provide an overview of publishing software, formats and licensing models in the Open Data landscape.

G2: In order to provide a profound theoretical overview of data quality assessment methodologies, we review literature regarding general assessment methods and methods tailored to the Open Data setting.

G3: Based on the reviewed literature, we introduce a set of objective quality metrics. We formally define the measures and describe their implementation, so that the metrics' calculation can be done in an automated way.

G4: We develop a monitoring framework which regularly monitors the data catalogs and analyzes their datasets. Our framework computes our defined quality metrics for the various dimensions and reports the observed results.

G5: In order to deal with the different management systems and their heterogeneous metadata, we investigate on various homogenization approaches in the literature. As a result, we propose a homogenization approach compliant with the W3C recommendation DCAT.

G6: Finally, we report findings based on monitoring a corpus of over 200 data portals. This includes software specific analysis and evolution reports (dealing with CKAN portals) and a detailed report of Austrian Open Data portals.

1.3 Contributions & Structure of this Work

The remainder of this thesis is structured as follows:

In **chapter 2**, as a means of achieving sub-goal ***G1***, we will introduce core concepts and terms such as "openness", "Open Data" and "Open Government Data". Further, we will investigate on different sources of Open Data and the current Open Data ecosystem.

Chapter 3 introduces the current metadata schemas on data portals and presents standardization and homogenization approaches (partially aligned with ***G5***). In order to tackle ***G2***, it continues with an overview of state-of-the-art quality metrics and quality assessment methods with regard to Open Data.

In **chapter 4**, we initially discuss in detail a set of selected CKAN specific quality dimensions and metrics, stated in ***G3***. Then, following sub-goal ***G4***, we introduce our automated quality assessment framework which is intended to monitor Open Data portals and assess the previously introduced quality metrics.

In **chapter 5**, we complete sub-goal ***G5*** by proposing a homogenization of different metadata schemas using the W3C's DCAT vocabulary. Furthermore, we discuss the adaption of the metrics defined in chapter 4 to the proposed general metadata model.

We monitor and assess the quality of various active Open Data portals and in **chapter 6** we report comprehensive findings by analyzing the data and the evolution of the portals and therefore achieve sub-goal ***G6***.

Finally, we will conclude in **chapter 7** with a critical discussion of the presented work and an outlook on future work.

1.3.1 Impacts & Publications

In the following, we list the contributions in form of projects and publications which were made in the course of this thesis.

Projects

- **Open Data Portal Watch** (`http://data.wu.ac.at/portalwatch/`) is a web-based platform which monitors Open Data portals in an automatic manner. It displays quantitative insights, reports and quality assessment results based on the quality metrics discussed in this thesis. This project is funded by "Jubiläumsfond der Stadt Wien".

- **data.wu.ac.at** is an Open Data portal hosted by the Institute for Information Business at the Vienna University of Economics and Business. The portal is based on the open source data management system CKAN. We host data about courses and lectures, rooms at the university campus or the university's library collection. The WU's portal is the first university Open Data portal in Austria and serves as a showcase for the Open Data Portal Watch project.

Conferences and Workshops

- 08/2015. Jürgen Umbrich, Sebastian Neumaier and Axel Polleres, Quality assessment & evolution of Open Data portals. In *OBD 2015: The International Conference on Open and Big Data*, Rome, Italy (Aug 2015), `https://ai.wu.ac.at/~polleres/publications/umbr-etal-2015OBD.pdf`.
  Received Best Paper Award.

- 03/2015. Jürgen Umbrich, Sebastian Neumaier and Axel Polleres, Towards assessing the quality evolution of Open Data portals. In *ODQ2015: Open Data Quality: from Theory to Practice Workshop, Network of Excellence in Internet Science*, Munich, Germany (Mar 2015), `http://polleres.net/publications/umbretal-2015ODQ`. pdf.

Technical Reports

- 03/2015. Sebastian Neumaier, Jürgen Umbrich, Quality metrics of the Open Data Portal watch project.

CHAPTER 2

Preliminaries

Herein, we introduce underlying concepts and terms used throughout the rest of the thesis, relating to the Open Data movement. In particular, we present current attempts of defining "openness" of data, we describe different sources of Open Data and differentiate Open Data from related other "Open" movements. In addition, we provide an overview of the present Open Data landscape, including the currently used software frameworks for publishing Open Data, common file formats, a short introduction into the Linked Data model RDF and the licensing of data.

2.1 What is Open Data?

The Open (Knowledge) Definition,[1] developed in 2005, is an attempt to define the meaning of "open" in the explicit context of the terms "Open Data" and "open content". It sums up the Open Data principle in the statement:

> "Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness)."

Here one can highlight the most important terms appearing in any attempt to define openness: **freely used**, **modified**, and **shared** by **anyone** for **any purpose**.

In order to clarify the application of "open" to a concrete data instance, the Open Definition defines the term "open work" as an item or piece of knowledge which must satisfy the following three requirements:

1. Firstly, the work must be **freely accessible** as a whole (preferably downloadable via the Internet). Additional licensing information must also accompany the work.

[1] `http://opendefinition.org/`, last accessed 2015-09-08

2. Further, the work should be provided in a **machine-readable** and **open format**. Machine-readable means that an automated and structured processing of the work is possible. An open format is a format with a freely available specification without restrictions upon its use.

   For instance, the PDF format is an open format with a freely available specification, but cannot be considered as machine-readable and therefore is no proper format for publishing Open Data.

3. Eventually, a essential requirement is the use of an **open licence** for publishing any piece of work. A licence is considered as open if it satisfies a set of conditions and permissions to preserve free usability.

In this definition, the term "work" is used to denote "the item or piece of knowledge being transferred".[2] On the basis of this very general formulation, "open work" can easily be substituted by any concrete data file in a specific data format. Therefore, according to the Open Definition, any published datasets and resources under the term "Open Data" and "open content" should comply with the introduced open work requirements.

In the further course of this chapter we want to look into the three essential points mentioned above. Regarding the free accessibility of data we discuss the current Open Data ecosystem and present popular Open Data publishing frameworks, so-called Open Data portals in section 2.2. We discuss the second point, the appropriate use of file formats for Open Data, in section 2.3. This includes the discussion of machine-readability, proprietary and openness of formats and a classification of popular file format. Eventually, for clarification of the third point in the above listing, we discuss open licences and related legal issues in section 2.4. We address the legal background of the Open Data movement, the need for complete and correct licensing of datasets and arising legal issues due to the integration of diverse licences.

In the following, we begin this chapter with a discussion of the relation to "Open Source" and list typical sources and publishers of Open Data. Then we provide a short terminology distinction of related "Data" movements.

From Open Source to Open Data? The term "Open Data" springs from the same roots as "Open Source" or "Open Access". Although, there are similarities regarding the free availability of source code and data, there are some key differences further developed in [LN14]. While the commercial and economic potential for generating value of open source has been proven in recent years, the value of Open Data can so far only be estimated and may not be immediately evident for data publisher as it is for open sources. However, an identified key similarity between Open Data and the established open source movement is, that both are promising and fast growing phenomenona facing new impediments and challenges.

[2] `http://opendefinition.org/od/`, last accessed 2015-09-08

There is already a remarkable ecosystem publishing numerous resources, even though the current Open Data movement faces a range of impediments [ZJC+12], including the (meta-)data quality concerns tackled in this work. Next, we will take a closer look at the main data sources of Open Data.

2.1.1 Typical Open Data Publishers

By now, we can identify many areas in which Open Data is used and valuable. Different groups of people and organizations can benefit from opening up data of different sources. Having said that, it is impossible to predict how, when and where value can be created in the future. Innovations enabled by freely available data can come from any unlikely place. In the following we will list the main domains and publishers of current open data and discuss the underlying motivation.

Government

The most prominent data source is Open Government Data (see section 2.1.2). In order to increase and improve transparency and democratic control, governments started to publish their documents and proceedings. This allows the public to control their government. Another aspect of opening up governmental data is an increased participation by the citizens.

Private Sector

While there are lots of success stories for openness of government data, the Open Data movement in the private sector is in the early stages of development. In fact, there is an huge potential economical value and a enormous amount of data, which could be utilized.[3] By opening up their data, companies can become more accountable and generate value.[4]

Sharing data across an industry can lead to greater transparency about products and services. For instance, by publishing data about pricing, quality or reliability of products, customers are able to aggregate this data. They may spent more money on products if there is a transparent production chain.

Another aspect of Open Data in the private sector is to gain new insight in consumer thinking. Consumer-facing companies can find out what customers value most by allowing consumers to collaborate. For instance, by rating products, or recommending additional products.

A possible outcome by sharing data with customers is a more trust-based relationships. Further, companies can acquire valuable information and feedback. However, opening up company data can involve risks which have to be identified and managed. E.g., a fully transparent view on a product or a service can easily exploit any present quality issues.

[3] http://blogs.worldbank.org/voices/next-frontier-open-data-open-private-sector, last accessed 2015-08-08

[4] http://www.ft.com/cms/s/0/ba7ae9be-9354-11e3-b07c-00144feab7de.html#axzz3iJag0455, last accessed 2015-08-09

Open Data in Science

Open Data in science is an intensively discussed topic. Currently, there are many barriers to access and (re-)use scientific data [Mol11]. For instance, there are restrictions on the usage applied by the authors, publishers or providers. Furthermore, data may be inaccessible in scientific publications, e.g., data within tables in PDF documents.

According to a 2009 report [SB+08] there are multiple reasons for researchers to keep their data closed. Firstly, there is a fear of exploitation. Scientists may not be willing to open up data if they feel they could extract any publications out of the data. Secondly, according to the report, researchers do not see an incentive to make their acquired data available. There is no direct career reward or benefit recognizable.

On the other hand, it is easily recognizable that the Open Data movement comprises a range of benefits for the scientific community. Freely available (scientific) data increases the level of transparency and reproducibility and hence increases the efficiency of the scientific process. An illustrative example is the science of Astronomy, where according to [Nor07] the increasing availability of data leads the field in a "Golden Age". In recent years astronomers benefit from data centers around the world (e.g., NASA's NED database[5]) offering their observation data and scientific contributions.

The next challenge for the scientific community will be the adoption of Open Data principles in parallel to the conventional formal publications of scientific work [MR08].

Beside the three aforementioned established sources for Open Data, namely governmental data, data from companies and data from science, we can identify another recently arising data origin. The "Open Culture Data" movement aims to open up data in the cultural sector, mainly driven by the *OpenGLAM* initiative.[6] GLAMs (i.e., galleries, libraries, archives and museums) have the possibility of publishing a variety of high quality datasets. For instance, the Dutch Rijksmuseum offers an API[7] to make its collection (consisting of textual content, high resolution images and corresponding metadata) freely available.

2.1.2 Related Data Movements

"Data" is frequently combined with a range of terms describing different concepts and movements. On the one hand we can identify an overlap between different Data movements and on the other hand we can see specializations and subset-relations of particular concepts.

[5] `http://ned.ipac.caltech.edu/`, last accessed 2015-09-10
[6] `http://openglam.org`, last accessed 2015-09-18
[7] `http://rijksmuseum.github.io`, last accessed 2015-09-18

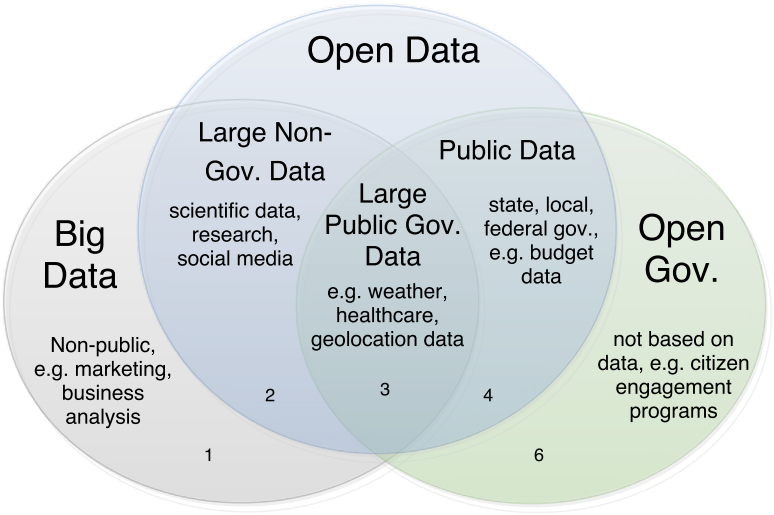

Figure 2.1: Big vs. Open vs. Government Data.[9]

In recent years the broad term "Big Data" emerged and draw a lot of attention as well as criticism.[8] Generally speaking, Big Data covers the processing and analysis of datasets which are so big or so complex such that they require special handling, i.e. the data cannot be handled with traditional tools and traditional processing infrastructure.

The Venn diagram in Figure 2.1 distinguishes and separates the term "Big Data" from the Open Government (section 2.1.2) and Open Data concepts and highlights the overlapping areas. The first part of the diagram covers large datasets which are not publicly available. For instance, this includes data collected by social networks about their customers. The second part stands for freely available, but non-governmental Big Data. This includes large datasets coming from the private sector as well as from science. As mentioned in subsection 2.1.1 more and more researchers from fields dealing with enormous amounts of data are sharing their work and observations, e.g., in genomics or astronomy. Part four covers governmental datasets of moderate size. This part can hold very useful data for citizens; for example, budget reports, or data about local services. Eventually, perhaps the most interesting datasets are covered by part three. For instance, this includes large-scale geolocation datasets, or statistical data acquired over a longer period of time.

Open Government Data

In [vLG10] Open Government Data (OGD) is defined as datasets, selected by a governmental institution, which are structured and provided in a machine-readable format

[8] http://www.ft.com/intl/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz2yQ2QQfQX, last accessed 2015-08-23

[9] Graph derived from http://www.mcgrawhillprofessionalbusinessblog.com/2014/02/18/an-infographic-big-data-is-big-open-data-is-revolutionary/

(see section 2.3 for machine-readability). These datasets are published in catalogs such that they can be browsed, searched, filtered, monitored and processed. In detail, theses datasets consist of statistics, spatial data, plans, maps, environmental and weather data, materials of parliaments, agencies and ministries, budgetary data, laws, ordinances, statutes, judicial decisions and other publications. Some exemplary applications, mashups and services based on open administrative data, can be found in the web-based portals `data.gv.at`[10] of the Austrian Government, `data.gov.uk`[11] of the British Government and the DataSF App Showcase[12] of the city of San Francisco.

The OGD Initiative. On his first day in office, in January 2009, US-President Obama signed the *Open Government Directive* [Ors09]. This directive is built on the principles of *transparency*, *collaboration* and *participation*. In this context, transparency means that governmental information should be treated as a national asset and the public should be able to use this information to hold the government accountable. Participation means the improvement of government decision-making processes by "tapping into the citizenry's collective expertise through proactive engagement."[BMS11] The term collaboration aims at the inclusion of universities, nonprofit organizations, businesses and the public to better execute the government tasks. In order to accomplish this ambitious goals, the directive required all US agencies to take the following steps:

Online publishing of Government information:
: This step includes the release of at least three "high-value" data sets via `Data.gov`, the creation of an Open Government webpage (www.[acency].gov/open) and the use of Open Data formats (i.e., platform independent, machine readable and available without restrictions).

Improve the quality of government information:
: The agencies must follow certain quality guidelines and must regularly report on their progress towards information quality improvement.

Create a "culture of open government":
: This step of the directive includes the integration of public participation and collaboration into the agencies activities.

Create an enabling policy framework:
: The agencies are urged to identify barriers to open government and propose policies and guidance to clarify this issues.

In retrospect, the OGD initiative can be identified as one of the main driving forces behind the current success of Open Government Data.

[10] `https://www.data.gv.at/anwendungen/`, last accessed 2015-08-23
[11] `http://data.gov.uk/apps`, last accessed 2015-08-23
[12] `https://data.sfgov.org/showcase`, last accessed 2015-08-23

Linked Open Data

Linked Open Data (LOD) is Open Data published in a structured data format which allows to interlink the data. It makes use of standard web technologies as Uniform Resource Identifiers (URIs) and the Resource Description Framework (RDF) (cf. section 2.3.3) to connect and link various data sets.

In [BL06] Berners-Lee defines Linked Data (LD) by the use of these four rules: (i) using URIs as names for things, (ii) using HTTP URIs, so that these URIs can refer to those things, (iii) providing useful information at the URI's destination (including the use of standards, e.g. RDF) and (iv) including links to other URIs. Further he states that Linked Open Data "is Linked Data which is released under an open licence, which does not impede its reuse for free".

★	Available on the web, in whatever format, but with an **open licence** (see section 2.4).
★★	Available as **machine-readable** structured data (e.g., Excel file instead of image scan of a table); see machine-readable formats in subsection 2.3.1.
★★★	As above, plus: **non-proprietary** format (e.g. CSV instead of Excel), see proprietary formats in subsection 2.3.2.
★★★★	All the above, plus: Use **open standards** from W3C (RDF and SPARQL) to identify things (see open standards and RDF in section 2.3.3).
★★★★★	Additionally, **link your data** to other people's data to provide context.

Table 2.1: Tim Berners-Lee's Open Data 5 Star rating

Table 2.1 holds the Open Data 5 Star rating proposed by Berners-Lee in [BL06]. The purpose of this rating is to encourage data publisher to provide their data as linked data, in such a manner that it is reusable and integrable by other linked data sources. Above all and in order to be considered as "open", it has to be accompanied by an open licence. Further, a higher rating requires the use of machine-readable and non-proprietary formats, the use of open standards and the linkage of the published data to other data sources.

Linked Open Government Data and the Linked Open Data Cloud. A popular LOD project is the *Linked Open Data Cloud*[13] (see Figure 2.2). It is a visualization of various Linked Open Data sources. The nodes in the cloud represent datasets and the arcs indicate existing links.

On the basis of Open Government Data and LOD, one can also identify Linked Open Government Data (LOGD). In the LOD cloud the LOGD data sources are colored in turquoise and positioned on the top left (see Figure 2.2 on page 14). Interestingly, LOGD covers almost a quarter of all LOD in the LOD cloud. For instance, this includes data from `data.gov.uk`.

[13] `http://lod-cloud.net/`, last accessed 2015-07-23

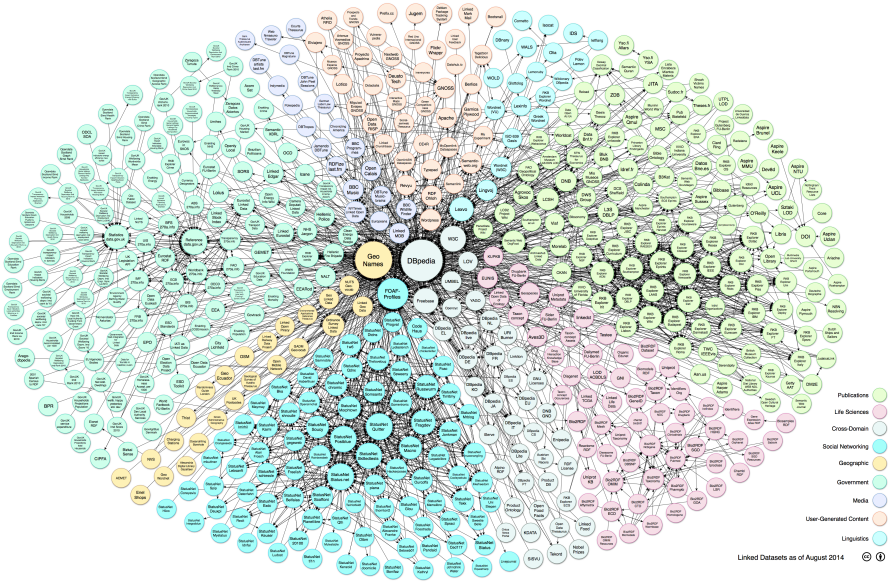

Figure 2.2: The Linking Open Data cloud diagram (from `http://lod-cloud.net/`, state: 2014-08-30)

2.2 Accessing Open Data

In this section we discuss the typical ecosystem for publishing and consumption of Open Data. Therefore, we introduce current data publishing solutions, frequently encountered file formats and legal issues concerning sharing and dissemination of Open Data.

2.2.1 Open Data Portals

Open Data Catalogs are websites which serve as a single point of access to distribute Open Data resources. Usually, this data catalogues offer an API which allows on the one hand the publisher to upload and update the content and on the other hand the user to automatically retrieve resources and their associated meta-data. This kind of catalogues are also referred to as Open Data Portals.

Generally speaking, a *data portal* is a website that hosts a collection of datasets and provides browse, search and filter functionalities over these datasets. Typically, such a dataset holds a single or multiple data files (referred to as resources) which are available for access or download in one or more formats. The resources themselves then either are hosted on the corresponding data portal or link to external data sources. Additionally, a dataset holds metadata (i.e., basic information) of these resources, e.g. a title, a format description and licensing information.

When taking a closer look into the landscape of Open Data catalogs one can observe several large data portals which serve mainly as harvesting portals. This means that these portals continuously collect datasets from small, local portals and mirror the harvested

resources.

Beside various tailored and specific data catalogs there exist two main data portal software products, namely CKAN and Socrata. Furthermore, there are a couple small, rarely used software products. Out of these, we take a closer look at the portal instances of the OpenDataSoft software.

CKAN

The Comprehensive Knowledge Archive Network (CKAN) is an open-source data portal platform developed by the Open Knowledge Foundation. The development of the project started in 2006 with the first release in 2007. The Open Knowledge Foundation is a nonprofit organization that aims to promote the openness of all forms of knowledge. It states the following goals in its original mission statement from 2004 [Win13]:

1. *To promote freedom of access, creation and dissemination of knowledge.*
2. *We develop, support and promote projects, communities and tools that foster and facilitate the creation, access to and dissemination of knowledge.*
3. *We campaign against restrictions both legal and non-legal on the creation, access to and dissemination of knowledge.*
4. *We seek to act as an intermediary between funding institutions and projects that work in areas related to the creation and diffusion of knowledge, particularly those with a strong technological aspect.*

The software is increasingly popular among cities, governments and private data provider worldwide including government portal of countries in Europe, South and North American and the Middle East. Some main features of CKAN is the ability to integrate community extensions, connections to CMS such as WordPress, the ability to link datasets and to integrate several existing CKAN instances into a single one (acting as a meta portal of portals).

Socrata

Socrata is a company founded in 2007 that offers a range of database and discovery services for governmental data. One of the products is an Open Data portal that aims to provide a solution for non-technical users. It serves as a cloud software, i.e. it hosts the data on their own server. The portal provides access to datasets via API. All Socrata products are closed and proprietary software. There are a number of states and cities utilizing Socrata, e.g. Washington[14] or the state of New York.[15] While CKAN instances are widespread around the world, the Socrata software can be found mainly in the US.

[14] `https://data.wa.gov/`, last accessed 2015-08-23
[15] `https://data.ny.gov/`, last accessed 2015-08-23

OpenDataSoft

The France-based company OpenDataSoft[16] provides a commercial data portal software similar to the Socrata portal. OpenDataSoft, headquartered in Paris, was founded in 2011 and it's customers are mainly Open Data portal instances of French cities. Customers include the city of Brussels[17] and the French National Railway Company.

2.3 Open vs. Closed Formats

Herein, we discuss suitable and unsuitable Open Data file formats with respect to *open standards* as defined by the W3C. As already mentioned in Tim Berners-Lee's Open Data 5 Star rating in section 2.1, Open Data resources shall be published using open standards. Following the recommendations of the World Wide Web Consortium in [Wor07], an open standard must meet the following requirements:

- ***Transparency:*** *Due process is public, and all technical discussions, meeting minutes, are archived and referenceable in decision making.*
- ***Relevance:*** *New standardization is started upon due analysis of the market needs, including requirements phase, e.g. accessibility, multi-linguism.*
- ***Openness:*** *Anybody can participate, and everybody does: industry, individual, public, government bodies, academia, on a worldwide scale.*
- ***Impartiality and Consensus:*** *Guaranteed fairness by the process and the neutral hosting of the W3C organization, with equal weight for each participant.*
- ***Availability:*** *Free access to the standard text, both during development and at final stage, translations, and clear IPR rules for implementation, allowing open source development in the case of Internet/Web technologies.*
- ***Maintenance:*** *Ongoing process for testing, errata, revision, permanent access.*

Beside dealing with open standards which specify formats, hereinafter referred to as **Open Formats**, we will look more deeply into different classifications of formats and their importance for Open Data publications. In this respect one can distinguish machine-readable formats, proprietary and non-proprietary formats, and formats with a freely available specification.

2.3.1 Machine Readable Format

According to the US Office of Management and Budget Machine Readable Formatss are defined as follows:

[16] `https://www.opendatasoft.com/company/`, last accessed 2015-08-23
[17] `http://opendata.brussels.be/`, last accessed 2015-08-23

> "The Format in a standard computer language (not English text) that can be read automatically by a web browser or computer system. (e.g., XML). Traditional word processing documents, hypertext markup language (HTML) and portable document format (PDF) files are easily read by humans but typically are difficult for machines to interpret. Other formats such as extensible markup language (XML), (JSON), or spreadsheets with header columns that can be exported as comma separated values (CSV) are machine readable formats. It is possible to make traditional word processing documents and other formats machine readable but the documents must include enhanced structural elements."[US 15]

It is worth noting that the Open Knowledge Foundation defines machine readable formats in its Open Data Handbook [Ope12] simply as structured data which can be automatically read and processed by a computer. Additionally, it states that "non-digital material (for example printed or hand-written documents) is by its non-digital nature not machine-readable".

2.3.2 Proprietary & Open Formats

A proprietary format is a file format which is usually controlled by a company or organization. Due to usage restrictions, proprietary formats possess a high exclusivity on its usage. The exclusivity of proprietary formats violates various requirements of an open standard as introduced above and furthermore affects the interoperability and portability of data files.

The restriction can be either a format whose specification is not released, also referred to as *closed proprietary*, or a format whose specification is in fact published but underlies restrictions (through patents or licences), called *open proprietary* format. Examples for popular closed proprietary formats are the RAR archive file format and PSD, the Adobe Photoshop file format.

Controversial proprietary formats are for example Microsoft's DOC format and the PDF file format. Microsoft published in 2006 an "Open Specification Promise",[18] promising not to assert its patents against implementations for a certain set of specifications (including DOC). However, this promise is not a licence and does not grant any rights. Regarding Adobe's PDF format there are some parts of the specification that are defined only by Adobe and therefore remain proprietary.[19]

In contrast, an Open Format (as introduced above) is not proprietary in both meanings, i.e., its specification is freely available and the format is not restricted through licences and is free to be used by everyone.

[18] `https://msdn.microsoft.com/de-at/openspecifications/dn646765`, last accessed 2015-09-08

[19] `http://www.planetpdf.com/enterprise/article.asp?ContentID=Is_PDF_an_open_standard`, last accessed 2015-09-08

2.3.3 Popular File Formats

In the following we discuss a set of popular file formats. These include formats that are suitable for publishing Open Data (including CSV, JSON and RDF) but also formats that are not fully appropriate for providing Open Data due to machine readability issues, however, widely used (e.g., PDF).

TXT. A TXT file is a text file holding unstructured and unformatted plain text. According to the Unicode standard [The12], plain text "is a pure sequence of character codes". In contrast, rich (i.e., structured and formatted) text "is any text representation consisting of plain text plus added information". The HTML, XML, or JSON formats are examples for rich text relying on underlying plain text.

CSV. A comma-separated values (CSV) file stores tabular data in plain text form using a separator symbol (also called delimiter) to separate the cells of the table. The records of the tabular data (i.e., the rows) are separated by new lines in the CSV format. Furthermore, CSV files provide the option to escape the content of cell entries (i.e., allowing the use of the separator characters within an entry) by surrounding the content of a cell by quotation marks (").

There exist no general standard for CSV files, but the Internet Engineering Task Force (IETF) provides a de facto standard in [Sha05]. This document specifies a syntax with `text/csv` as the registered media-type. Sometimes CSV is also referred to as *character-separated values*, because there are variations on the separator character. While IETF syntax definition names the comma (",") as the default delimiter, one can easily find CSV files separated by a tab or by a semicolon.

Apparently, the variations on the separator characters can cause problems regarding the automated integration and conversion of CSV files. By default, the actual delimiter is not obvious by the file extension or the media-type.

PDF. The Portable Document Format (PDF) is a widely used file format developed by Adobe in 1993[20] using ".pdf" as file extension and `application/pdf` as media-type. Until 2008, when Adobe released PDF as an open standard [Int08], it was a proprietary format, controlled by the company. As already mentioned in subsection 2.3.2, there are still some proprietary technologies which are not standardized and defined only by Adobe. The openness of the format, therefore, is debatable.

The format itself is used to capture textual and visual contents in an immutable way with a fixed layout and font. The main advantages of using PDF are the uniform and locked presentation and the platform independence regarding the viewing software. A disadvantage is the limited possibility to extract content out of a PDF file. While the text in PDF documents is accessible and searchable, structured content (e.g., relational data described in tables) is difficult to extract.

[20] `http://www.adobe.com/devnet/pdf/pdf_reference.html`, last accessed 2015-09-09

JSON. The JavaScript Object Notation (JSON) format [Bra14] is a so-called semi-structured file format, i.e., it consists of data that is not organized in a complex manner and has no fixed schema (as for example data in relational databases), but contains associated structural information. The internet media type for JSON, specified in [Bra14], is `application/json` and the common filename extension is ".json".

Initially, the JSON format was mainly intended to transmit data between servers and web applications, supported by web services and APIs. When dealing with Open Data, one can find the JSON format also in other applications. On the one hand, in many cases it is used for publishing data, i.e., as a file format for actual resources on Open Data portals. Since raw tabular data can easily be transformed into semi-structured and tree-based formats like JSON,[21] many data providers offer their data also in JSON. On the other hand, the JSON format is the de facto standard for retrieving metadata from Open Data portals. All data frameworks presented in subsection 2.2.1 offer their datasets as JSON instances.

RDF. RDF, W3C recommendation since 2004 [KC], is a metadata model language for describing resources on the web. RDF is designed to describe relations between resources such that they can be read and understood by machines.

In the context of RDF, each resource on the web is identified by a unique pointer called URI. A URI can identify web pages, resources or documents, as well as real world things:

```
<http://www.w3.org/ns/dcat#Dataset>
<http://dbpedia.org/resource/Building>
```

RDF itself consists of statements in the form of *subject*, *predicate*, *object* triples:

The **subject** of an RDF statement is the described resource, identified by an URI. This URI is not necessarily an existing Uniform Resource Locator (URL) and is used to identify a physically non-existing or existing resource.

The **predicate** of a triple is a property used to describe the subject. This resource is taken from an RDF vocabulary.

The **object** is the value of the predicate. This can be either another resource or a literal containing a string, integer, date, etc.

The following example makes use of a shorter notation (see Turtle specification[BBL08] for further information). At the beginning of the RDF document there is a prefix declaration to improve the readability of the syntax:[22]

```
@prefix dct: <http://purl.org/dc/terms/>
@prefix dcat: <http://www.w3.org/ns/dcat#>
```

[21] For instance, see Converter Tools on `https://project-open-data.cio.gov/`, last accessed 2015-09-09

[22] The property `a` is a common shortcut for `rdf:type`.

```
<https://open.whitehouse.gov/dataset/...> a dcat:Dataset
<https://open...> dct:title "The White House ..."
```

RDF triples can be displayed as graphs where the subjects and objects are nodes and the predicates are directed edges. In RDF graphs it is common practice to represent resources in ovals and literals in rectangles. The graph in Figure 2.3 makes use of so-called blank nodes. These blank nodes are resources which are currently not identified by a URI. It can be used to formulate incomplete or missing information.

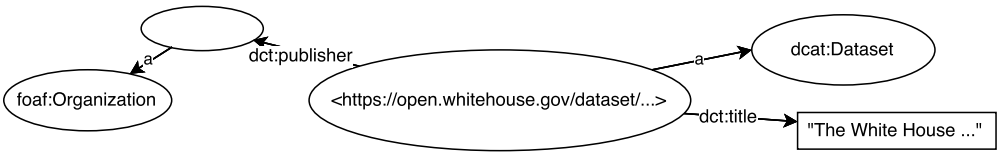

Figure 2.3: RDF Graph of a dataset, including blank node.

There exist several formats to serialize RDF data. Most prominent is RDF/XML, introduced in the course of the W3C specification of the RDF data model. More recent, in 2014, W3C released the first recommendation for JSON-LD [SKL14]. JSON-LD is an extension for the JSON format. Conventional JSON parser and databases can be used; users of JSON-LD, which are mainly interested in conventional JSON, are not required to understand RDF and do not have to use the Linked Data additions.

2.3.4 Classification of Relevant Formats

Table 2.2 lists a number of frequently observed file formats. The list is intended as a survey of common file formats in the Open Data landscape and classifies them into machine readability, openness and free availability of the format's specification. The classifications in this list are far from complete and one can easily see that the openness and machine readability is discussable for some formats. For instance, it is unclear if the Microsoft Excel format can be considered as machine readable because it is encoded in binary data.

Positively noticeable, for all of the formats in the list the format specification is available. There are some very prominent and common formats in this list which can not be considered as machine-readable (e.g., the PDF format). Furthermore, there are commonly used proprietary formats which cannot be considered as open (e.g., Microsoft Excel files).

Beside the classification into machine-readability and openness, the formats in this table can be grouped by the structure of the underlying data. E.g., there are formats dealing with unstructured text (e.g., TXT) as well as structured text (Microsoft Word or Open Document files), or formats to represent tabular data (e.g., CSV, XLS), formats representing tree-based (e.g., XML, JSON) and graph-based structures (RDF). On the other hand, the formats can be categorized by the data's domain. For instance, a

Extension	Full Name	MR[a]	SA[b]	O[c]
txt	Text file		✗	✗
csv	Comma Separated Value	✗	✗	✗
html	Hypertext Markup Language		✗	✗
xml	Extensible Markup Language	✗	✗	✗
rdf	Resource Description Framework	✗	✗	✗
json	JavaScript Object Notation	✗	✗	✗
pdf	Portable Document Format		✗	✗
jpg, jpeg	JPEG image format		✗	✗
png	Portable Network Graphics		✗	✗
gif	Graphics Interchange Format		✗	
doc, docx	Microsoft Word		✗	
xls, xlsx	Microsoft Excel	✗	✗	
rtf	Microsoft Rich Text Format	✗	✗	
odt, ods, ...	Open Document Formats	✗	✗	✗
gml	Geography Markup Language	✗	✗	✗
kml	Keyhole Markup Language	✗	✗	✗
gpx	GPS Exchange Format	✗	✗	✗

Table 2.2: Relevant data formats

[a]Machine readable (see definition of *Machine Readability* in subsection 2.3.1)
[b]Specification freely available
[c]Open (see definition of *open standards* in section 2.3)

prominent domain on Open Data portals are geolocation-based file formats (e.g., GML, KML).

2.4 Licensing of Data

An important subject in the context of the Open Data movement is the correct licensing of released data. In principle, a licence is defined as a statement which legally allows doing something which would be not allowed otherwise. For example, it can grant permissions to use, reuse, own or distribute something.

With the development and emergence of open source software there was a need for suitable licences to publish the source code such that the consumer's and publisher's needs are fulfilled. This means on the one hand that providing open source should be attractive for businesses but on the other hand that consumer should be allowed to reuse the source code in their one software project. The open source concept is very similar to the later emerging Open Data movement, so the licensing models are based on the idea of open source licences.

2.4.1 Open Licensing

In the Open Definition[23] a licence is considered as open if it satisfies a set of *required permissions*. Additionally, the definition lists a set of *acceptable conditions* that can be included in the licence. Herein, we list the permissions and conditions as listed in the Open Definition. A detailed explanation of all terms used in the definition would go beyond the scope of this thesis and therefore we refer to the Open Definition for a full and in-depth discussion.

Required Permissions:
Use, Redistribution, Modification, Separation, Compilation, Non-discrimination, Propagation, Application to Any Purpose, No Charge

Acceptable Conditions:
Attribution, Integrity, Share-alike, Notice, Source, Technical Restriction Prohibition, Non-aggression

Further, according to the Open Definition, the acceptable conditions include two kinds of restrictions allowed in any open licence: firstly, the requirement to give *attribution* to the source of the content and secondly, the requirement to publish any derived content under the same licence, called *share-alike*. This possible conditions lead to three levels of licensing:

- A licence which has no restrictions at all is called **public domain** licence.
- **Attribution** licences state that re-users must give attribution to the source of the content.
- **Attribution & share-alike** licences state that re-users must give attribution and have to share any derived content under the same licence.

The Open Data Institute is a private non-profit company acting as a catalyst for developing Open Data by providing public training courses. In its guide to open data licensing[24] it recommends open licences for *creative content* and open licences for *databases*.

Open licences for Creative Contents

For publishing creative content (e.g., text, photographs, slides) the Open Data Institute recommends the use of Creative Commons (CC) licences. CC is an US nonprofit organisation developing copyright-licences known as CC licences. CC was founded in 2001 and released several licences which allow the creators of works to specify which rights they reserve. Theses licences are free of charge to the public. The organisation aims to provide an easy-to-use way of making work available for others to share and to build upon legally.

[23] `http://opendefinition.org/`, last accessed 2015-09-09

[24] `https://theodi.org/guides/publishers-guide-open-data-licensing`, last accessed 2015-09-10

Table 2.3 lists three CC licences worth considering for opening up creative content.[24] The CC0 licence enables owners of copyright to "waive those interests in their works and thereby place them as completely as possible in the public domain" [Com09]. In contrast, the CC-by and CC-by-sa allow content owners to place attribution and share-alike restriction respectively on the content.

Creative Commons licence	**Level of licence**
CC0	Public Domain
CC-by	Attribution
CC-by-sa	Attribution & Share-Alike

Table 2.3: Creative Commons Licences for open content.

A possible drawback of Open Data under CC licences is the usage of attribution within almost all CC licences (except for the CC0 licence). In particular this can affect the integration and combination of multiple data sources which requires attribution (see Legal Issues and Problems below).

Open Licences for Databases

Beside using CC licences for data as well as for content, there is a set specific database licences suitable to license Open Data. The licences listed in Table 2.4 are database licences developed by Open Data Commons,[25] which is an Open Knowledge Foundation project.

Creative Commons licence	**Level of licence**
PDDL	Public Domain
ODC-by	Attribution
ODbL	Attribution & Share-Alike

Table 2.4: Creative Commons Licences for open content.

By looking into the definitions of the listed Open Data Commons licences, one can observe a high conformity with the above mentioned CC licences. The PDDL and CC0 licences both aim to place the content in the public domain, ODC-by and CC-by mainly focus on the usage of attribution and likewise ODbL and CC-by-sa share in principle the same requirements, i.e., attribution and share-alike.

Other Open licences

While there is a range of governments which use and recommend the use of open CC licences (e.g., the Austrian government), there are some governments which developed

[25] http://opendatacommons.org/, last accessed 2015-09-10

their own licensing model. For instance, there is the "Datenlizenz Deutschland"[26] by the German government providing licences for placing content in the public domain (`dl-de/zero-2-0`) and applying attribution requirements (`dl-de/by-2-0`). Another example is the "Open Government Licence"[27] developed by the UK government.

2.4.2 Legal Issues and Problems

Compatibility of licences

Licensing compatibility is an issue, which initially arised in the context of open sources software where it is defined as the "characteristic of two (or more) licences according to which the codes distributed under these licences may be put together in order to create a bigger distributable software" [Lau08]. This definition can directly be adapted to Open Data licensing. In Open Data, compatibility issues may occur when combining and integrating differently licensed data.

Therefore, we can further specify the incompatibility of licences using [Lau08]:

> "Incompatibility is due to contradictory obligations provided in the different licences under which two codes to be merged are distributed."

The main source of compatibility issues in Open Data is due to share-alike requirements (see Open Licensing above). I.e., one combines a set of resources where each of the resources requires the derived content to be placed under the same licence. This raises the problem of which of the parent licences to use.

Attribution Chains of licences

Apparently, the consecutive application of attribution licences for derived datasets is limited and impractical. A full attribution list for content which has been remixed consecutively multiple times can be hard to maintain and can grow relatively quickly, on the assumption the work integrates various sources.

For instance, assuming a combined work remixes sources for three generation, where each parent source integrates four attribution requirements; then the final content is bound to 20 different attribution statements.

Privacy Issues

A 2011 report ordered by the UK government [O'H11] concludes that "privacy and transparency are compatible as long as the former is carefully protected and considered at every stage". This includes the protection of "personal data", i.e., "data which relates to a living individual who can be identified from those data", or from those data and other available information.

[26] `https://www.govdata.de/lizenzen`, last accessed 2015-09-10

[27] `http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/`, last accessed 2015-09-10

CHAPTER 3

Background

In the following chapter we will define the necessary terminology and background to get the reader familiar with *metadata* and *data quality* in the context of Open Data. When differentiating the terms "data" and "information", intuitively, one could describe information as processed data. However, in the literature these terms are often used synonymously (as seen in [PLW02]) and therefore, this thesis will use the terms data and information interchangeably.

Metadata. According to [DHSW02], metadata (literally "data about data") can occur in three ways: (i) embedded metadata, which is "contained within the markup of the resource itself", (ii) associated metadata, which is "maintained in files tightly coupled to the resources they describe" and (iii) third-party metadata, which is "maintained in a separate repository by an organization that may or may not have direct control over or access to the content of the resource".

Regarding Open Data portals, we mainly deal with metadata of type (iii), but also type (ii): in section 3.1 we initially take a look at CKAN metadata, which is third-party metadata, maintained on CKAN portals, describing external resources. Then we look at metadata on Socrata and OpenDataSoft portals. These portals hold associated metadata, tightly coupled to the resources, which are hosted on the respective portals. Eventually, we discuss metadata homogenization approaches and introduce the W3C recommendation DCAT.

Data Quality. The ISO definition of quality is "degree to which a set of inherent characteristics fulfils requirement" [Int92]. When applying this definition to data in a certain domain, one can research into the set of most suitable characteristics (i.e., distinguished features or metrics) to assess the quality.

In section 3.2 we review different Open Data quality assessment approaches, which are related to the effort presented in this thesis. We list the used quality metrics and describe the applied assessment methods found in the literature.

3.1 Metadata on Data Catalogs

In [ATS15] Assaf et al. list the following necessary metadata information to successful integrate and (re-)use data from Open Data catalogs:

- **Access** information is required to get the actual resource of a dataset. I.e., a dataset is useless if there is no retrievable resource URL to download the actual content of the resource (or no other option to access the content).

- A dataset should include **licence** information in human- and machine-readable form in order to clarify terms of (re-)use.

- **Provenance** information within metadata is required to clarify where the data origins from, who created/aggregated it, etc. (in case of unclear licensing or versioning information).

  Generally speaking, provenance information is metadata that describes "[...] people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing" [MG13]. In the context of the Web and Linked Data, there exists PROV [GMe13], an RDF vocabulary defined by the W3C in order to provide an interoperable way of specifying provenance information.

Further, Assaf et al. [ATS15] identified the following benefits of rich metadata:

Data discovery, exploration and reuse:
Associated metadata is the main influence on the discoverability of published raw data. For instance, if the actual data is not indexable or too big to investigate, the user is dependent on informative and rich metadata.

Organization and identification:
A well organized categorization of data (e.g., use of domain specific categories and keywords) enables a better handling of huge data portals in terms of usability and searchability of the actual resources.

Archiving and preservation:
Available provenance information can be useful to track the origin, and therefore the characteristics and specifications of published raw data. In particular, this can be necessary to restore archived data (e.g., data available in outdated formats).

In the following, we investigate the structure of metadata published on data portals using CKAN, Socrata and OpenDataSoft software. We give examples of associated metadata, highlight relevant and important metadata fields and eventually, we recognize lacking metadata information.

3.1.1 Metadata Schemas on Open Data Portals

There exists three prominent software frameworks for publishing Open Data, (i) the commercial Socrata Open Data portal and (ii) the open source framework CKAN, developed by the Open Knowledge Foundation. Furthermore, there is the recent data publishing platform (iii) OpenDataSoft, deployed mainly in French Open Data catalogs. These portal frameworks provide ecosystems to describe, publish and consume datasets, i.e., metadata descriptions along with pointers to data resources. Portal software frameworks typically consist of a content management system, some query and search features as well as APIs to allow agents to interact with the platform.

Since parts of this work focus on monitoring CKAN portals, we initially discuss the metadata schema of CKAN and provide formal definitions used in the remainder of this work.

CKAN Portal

The central entities in a CKAN portal are *datasets* which contain general metadata to describe important contextual information about the dataset such as the publisher, used license, the data format or its encoding, and a link to the actual data resources. This metadata is organized as key-value pairs, where the key contains the label of a property describing the respective dataset, and its value contains the corresponding description or numerical value of this property. In order to avoid confusion in the further course of this thesis, we will use the term *dataset* synonymously for the metadata description of the actual dataset.

Figure 3.1 shows an excerpt of a metadata description for a CKAN dataset (d) in the JSON format. The value of some metadata key (as defined by the JSON specification [Bra14]) can be a number, a string, a boolean value (`true` or `false`), an array (in square brackets), an object (in curly brackets) or `Null`. In order to better classify missing information, we treat empty strings and `Null`-values equal. We distinguish between three categories of metadata keys in a CKAN portal:

core keys: a set of predefined keys which are generic and by default available in any CKAN portal, such as the `license_id` key in Figure 3.1.

extra keys: a set of arbitrary additional metadata keys to describe a datasets defined by the portal provider. These keys are listed under the `extras` key (cf. `schema_language` in Figure 3.1)

resource keys: a mix between some default keys and additional keys defined by the portal provider to describe the particular resources (e.g., a data file or also an API). For instance, while the default keys `format` and `url` are present in any resource description, the rarely occurring `size` key can be considered as an additional key. Each resource is described under the `resources` key.

In Figure 4.1 (section 4.2 on page 49) we present a full dataset description of an example CKAN dataset found on an Austrian portal. In this example we can see a set of *core keys*,

```
d: {
  "license_id": "cc-by",
  "author":"National ...",
  ...
  "extras":{
     "schema_language": "ger",
     ...
     "k_m^e": value(k_m^e),
  },
  "resources": [
    {
      "format":"CSV",
      "url": r_1,
      ...
      "k_k^r": value(k_k^r),
    },{ "format":"RDF", ··· }
  ],

  ...
  "k_n^c": value(k_n^c)
}
```

Figure 3.1: High level structure of the metadata for a CKAN dataset.

which are the top-level keys in the JSON document (e.g., `author`, `name` and `notes`). Under the `resources` key we can find a list of resource descriptions, which consists of a single resource in this example. The keys in this resource description (e.g., `format`) are the *resource keys*. Under the `extra` key in the JSON document we can see a single additional *extra key*: "`Sprache des Metadatensatzes`".

Definition 1 (Formal definition of a CKAN portal) *Formally, let* p *be a CKAN portal which hosts* m *datasets* $\mathcal{D}(p)=\{d_1, d_2, \ldots, d_m\}$ *and* n *resources* $\mathcal{R}(p)=\{r_1, r_2, \ldots, r_n\}$. *Let the function* $\text{res}(d) \in \mathcal{R}$ *denote all resources described by a dataset* d.

Next, let $\text{keys}(p) \subseteq \mathcal{K} = \mathcal{K}^{\mathcal{C}} \cup \mathcal{K}^{\mathcal{E}} \cup \mathcal{K}^{\mathcal{R}}$ *return the set of used metadata keys for a portal* p, *with* $\mathcal{K}^{\mathcal{C}} = \{k_1^c, k_2^c, \cdots, k_n^c\}$ *be the set of core metadata keys,* $\mathcal{K}^{\mathcal{E}}$ *the set of extra metadata keys and* $\mathcal{K}^{\mathcal{R}}$ *the set of metadata keys for resources (cf. Figure 3.1).*

Eventually, let $\text{keys}(.) \subseteq \mathcal{K}$ *be the set of keys used in a portal* p, *dataset* d *or resource* r *and* $\text{keys}(.|\mathcal{K}^*) \subseteq \mathcal{K}^*$ *be the keys belonging to a specific key set* ($\mathcal{K}^*$) *used in a dataset* d.

Table 3.1 presents the necessary overview of our notation used in the remainder of this work.

$\mathcal{K} = \mathcal{K}^{\mathcal{C}} \cup \mathcal{K}^{\mathcal{E}} \cup \mathcal{K}^{\mathcal{R}}$	Set of all available meta keys
$\mathcal{K}^{\mathcal{C}}$	Set of all available **core** meta keys
$\mathcal{K}^{\mathcal{E}}$	Set of all available **extra** meta keys
$\mathcal{K}^{\mathcal{R}}$	Set of all available **resource** meta keys
keys(.) $\subseteq \mathcal{K}$	All unique keys used in a portal p, dataset d or resource r
keys(.$\|\mathcal{K}^*$) $\subseteq \mathcal{K}^*$	All unique keys belonging to a certain set $\mathcal{K}^*$ used in a portal p, dataset d or resource r

Table 3.1: CKAN metadata key sets.

Analysis. In principle, CKAN encourages a very rich and detailed description of a dataset. On the one hand, the use of extra keys allows portal provider to introduce additional keys in their metadata. On the other hand, these extra keys involve the risk of increased heterogeneity across portals. Further, CKAN is non-restrictive regarding the metadata values. For instance, the value under the `format` key can contain any user specified information (e.g., "CSV", "csv", ".csv" or "text/csv").

Socrata Portal

In contrast to CKAN, in Socrata data portals there are no references to external resources. All data is stored in the system's internal database and can be exported and downloaded in several data formats. Therefore, in Socrata there is no file format specification within the metadata.

The internal stored data is represented in different *view types* (e.g., tables, charts, maps, calendars or forms). A view roughly corresponds to a CKAN dataset consisting of only one resource. By far the most common dataset type in Socrata portals are tables; thus, we only consider tabular views in our metadata analysis. Socrata's table views provide interactive functionalities (e.g., filter of values, sorting of columns or simple visualizations) and can be exported to CSV, JSON or XML files. In the remainder of the work, we will refer to the metadata of a tabular view as a (Socrata) *dataset*.

Similar to CKAN datasets, Socrata provide metadata for views as key-value pairs, where the values are either strings (e.g., for the `name` key in Figure 3.2) or nested descriptions providing additional metadata for a key (e.g., under the key `owner`). Additionally, a Socrata dataset contains some information about the columns of the corresponding table under the `columns` key (see Figure 3.2). Figure 3.1 is an example metadata description for a Socrata dataset in the JSON format.

Additionally to viewing the full metadata for a dataset, Socrata provides the option to output RDF metadata, partially (cf. Analysis below) conforming to the DCAT schema (see section 3.1.2). An example of this data can be found in Figure 3.3.

Analysis. Interestingly, within the JSON data of a Socrata dataset, there is no download URL for the actual resource specified. To get this URL, one has to utilize the `id` on the

```
{
  "id": "n5m4-mism",
  "name": "The White House - Nominations & Appointments",
  "createdAt": 1243376954,
  "downloadCount": 20302,
  "owner":[
    {
      "id": "bhnt-uir2",
      "displayName": "whitehouse",
    }
  ],
  "columns":[
     {
       "name": "Name",
       "dataTypeName": "text",
       "position": 1,
       ...
     },
     ...
  ],
  ...
}
```


Figure 3.2: Example metadata for a Socrata view.

Socrata data-export API. Furthermore, the metadata provides hardly any provenance information. The metadata contains an `owner` key, but no further fields containing contact information.

Regarding the DCAT output of a dataset, one can observe that the RDF data is not fully conforming to the standardized DCAT model. For instance, in Figure 3.3 Socrata emits properties using the `ods` namespace (`ods:created`, `ods:last_modified`) instead of using the corresponding specified `dcat` predicates (`dcat:issued`, `dcat:modified`).

OpenDataSoft Portal

Very similar to Socrata, in OpenDataSoft portals there is no distinction between a resource and a dataset. As the aforementioned portals, OpenDataSoft provides metadata in JSON format (Figure 3.4). The key `metas` contains a series of key-value pairs, where some of the keys are aligned with the predicates in the DCAT schema.

As in Socrata, OpenDataSoft provides the option to export the tabular resources of a dataset to CSV, XLS or JSON format. Additionally, an OpenDataSoft dataset can support "geo" features (indicated in the `features` metadata field). Theses datasets support exporting functionality to geolocation-based file formats (e.g., KML).

```
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix ods: <http://open-data-standards.github.com#> .
@prefix dct: <http://purl.org/dc/terms/> .

[] a dcat:Dataset ;
    ods:created "2009-05-26" ;
    ods:last_modified "2014-07-07" ;
    dct:identifier "n5m4-mism" ;
    dct:issued "2009-05-26" ;
    dct:title "The White House - ..." ;
    dcat:accessURL <https://opendata.socrata.com/api/...> ;
    dcat:distribution [ a dcat:Download ;
            dct:format [ a dct:format ;
                    rdfs:label "xml" ;
                    rdf:value "application/xml" ] ;
            dcat:accessURL <https://opendata.socrata...> ],
        [ a dcat:Download ;
            dct:format [ a dct:format ;
                    rdfs:label "csv" ;
                    rdf:value "text/csv" ] ;
            dcat:accessURL <https://opendata.socrata...> ] .
```

Figure 3.3: DCAT output for a Socrata view.

3.1.2 Metadata Homogenization Approaches

Next, we provide an overview over various attempts to standardize and homogenize metadata in data catalogs. This literature review includes approaches supported by strong business interests (Schema.org[1]), as well as approaches driven by the W3C, making use of the already existing Linked Data environment (DCAT [ME14]).

Dublin Core Metadata Initiative

Dublin Core (DC) is a metadata standard that has been specified by the Dublin Core Metadata Initiative (DCMI) [WKLW98]. It contains elements for describing resources that are used primarily for cataloging, archiving and indexing of documents (e.g., in archives, libraries). There is a distinction between two different levels of metadata in DC: simple, standardized DC and extended DC (DCMI Metadata Terms). The former consists of 15 core elements (extension of originally 13 elements). This includes terms like “publisher”, “title”, “format” or “description”. The elements can be used optional or multiple times; depending on the relevance to the particular resource. E.g., a publication

[1] `https://schema.org`, last accessed 2015-09-26

```
{
  "datasetid":"killings-by-law-enforcement-officers",
  "metas":{
    "publisher":"Wikipedia Contributors",
    "domain":"public",
    "language":"en",
    "license":"CC BY-SA",
    "records_count":1718,
    "title":"Killings by law enforcement officers in the USA",
    "modified":"2014-12-10T13:51:32+00:00",
    "visibility":"domain",
    "theme":"Justice, Safety, Police, Crime",
    "references":"http://en.wikipedia.org/...",
    "keyword":[
      "killings",
      "law enforcement officers",
      "USA"
    ],
    "description":"Lists of people killed by ..."
  },
  "has_records":true,
  "features":[ "geo", "analyze", "timeserie" ],
  "attachments":[],
  "fields":[ ... ]
}
```


Figure 3.4: Example of an OpenDataSoft dataset.

may contain multiple authors. The second one, "DCMI Metadata Terms", further specify these core elements and provide a more comprehensive set of description elements.

Schema.org

Schema.org[2] is a markup language developed in collaboration of the search engines Google, Bing, Yahoo and the Russian engine Yandex. The language can be used to markup, identify and structure content on web pages, such that they can be easily indexed by search engines. Since all of the main search engines in the world have agreed on Schema.org as an uniform standard, it is widely used and has become a de-facto standard in recent years.

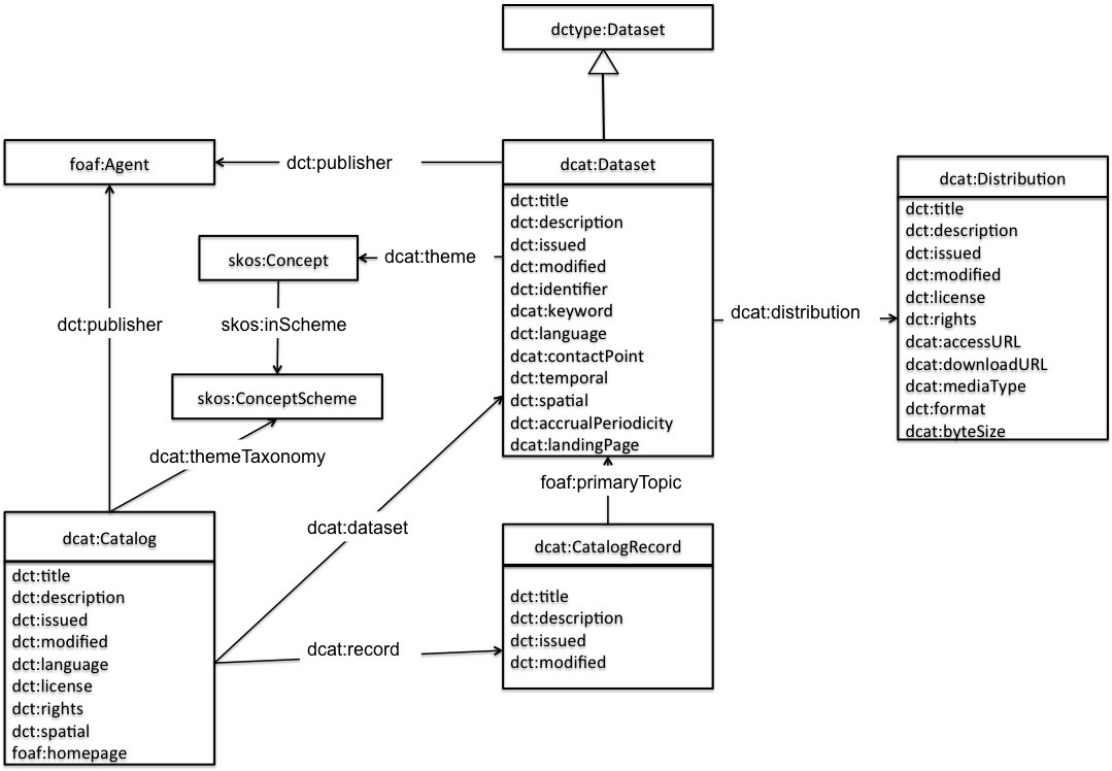

Figure 3.5: The DCAT model [ME14]

DCAT

The Data Catalog Vocabulary (DCAT) [ME14] is a W3C metadata recommendation for publishing data on the Web. DCAT is defined in RDF and, inter alia, reuses the Dublin Core Metadata (`dct`) [WKLW98] vocabulary.

The DCAT model (see Figure 3.5) includes four main classes: `dcat:Catalog`, `dcat:CatalogRecord`, `dcat:Dataset` and `dcat:Distribution`. The definition of a `dcat:Catalog` corresponds to the previously in subsection 2.2.1 introduced concept of Open Data portals, i.e., it is a web-based data catalog holding a collection of metadata about datasets. Most interestingly in the context of Open Data portals are the RDF resources `dcat:Dataset` and `dcat:Distribution`. On the one hand, a `dcat:Dataset` is a metadata instance which can hold one or multiple distributions, a publisher, and a set of properties describing the dataset. A `dcat:Distribution` instance, on the other hand, provides the actual references to the data (in `dcat:accessURL` or `dcat:downloadURL`). Further, it contains properties to describe license information (`dct:license`), format (`dct:format`) and media-type (`dct:mediaType`) descriptions and general descriptive information (e.g, `dct:title` and `dcat:byteSize`).

[2] `https://schema.org/docs/full.html`, last accessed 2015-09-21

The dcat:CatalogRecord class is an optional class and can be used if there is a distinction between metadata about a dataset and metadata about the dataset's entry in the corresponding data portal. For example, a dcat:CatalogRecord instance can hold a dct:modified property, containing the date when the metadata description of the dataset was last changed (and not the actual dataset).

DCAT-AP

The DCAT application profile for data portals in Europe (DCAT-AP)[3] is a currently running project by the European Commission which aims towards the integration of datasets from different European data portals.

In its description it states that

> "*[...] the availability of the information in a machine-readable format as well as a thin layer of commonly agreed metadata could facilitate data cross-reference and interoperability and therefore considerably enhance its value for reuse.*"[3]

In its current version (May 2015) it extends the existing DCAT schema by a set of additional properties. DCAT-AP allows to specify the version and the period of time of a dataset. Further, it classifies certain predicates as "optional", "recommended" or "mandatory". For example, in the DCAT-AP profile it is mandatory for a dcat:Distribution to hold a dcat:accessURL.

At the moment there is only a single portal supporting DCAT-AP metadata, namely the Swedish opengov.se. Moreover, the Pan-European Data Portal (publicdata.eu), which currently harvests 28 other European data portals, intends to use DCAT-AP for describing datasets.

HDL

In 2015 Assaf et al.[ATS15] proposed HDL, an harmonized dataset model. At the moment HDL is mainly based on a set of frequent CKAN keys. On this basis, the authors define mappings from other metadata schemas, including Socrata, DCAT and Schema.org. For example, HDL maps the Socrata key description and DCAT information dcat:Dataset→dct:description[4] to the CKAN key notes.

3.2 Related Work

Data quality assessment (QA) and improvement methodologies are widely used in various research areas such as in relational databases, data warehouses, information or process

[3] https://joinup.ec.europa.eu/asset/dcat_application_profile/description, last accessed 2015-09-21

[4] The → notation indicates that there is a triple with a subject of type dcat:Dataset, the dcat:description property and the corresponding mapped value as object.

management systems [SLW97, JV97], but also to assess the quality of Linked Open Data [ZRM+14]. Over times, different application and research areas established various measures and techniques to assess the quality of data and services and for keeping up with the increasing complexity of the tasks [ZMLW14]. Batini et al.[BCFM09] published in 2009 a detailed and systematic description of methodologies to assess and improve data quality. Generally, the different methodologies involve various phases starting from the definition of quality metrics, the measurement, an analysis phase and possible improvements with small differences how feedback loops are integrated.

Various efforts already exist to study different aspects of Open Data portals. Some projects deal with the quality of Open Government Data (see subsection 3.2.1) and aim to assess and compare the state of Open Data across different countries. Recent projects try to assess the progress and spreading of Open Data in the private sector (see subsection 3.2.2). Further, there are projects which propose various metrics to evaluate the (meta-)data quality within data catalogues (see subsection 3.2.3). The following literature review focuses on highlighting the quality metrics, and the analysis phase in each of the examined projects.

3.2.1 Government Data

OpenData Barometer. The Open Data Barometer project [Wor15] assesses the readiness of countries to exploit their governmental Open Data efforts and measures the achieved impact based on expert judgements. The results of the report are based on experts from 86 countries. These experts are asked to carry out a number of detailed questions about the situation in their country. They are researchers, giving their response in a 0 - 10 scale and provide citations and justification for their scores. Further, the authors of the report investigate for the availability of different kind of data within a country (e.g., budget data, public transport timetable data or mapping data). The overall evaluation then is categorized in three factors: *Readiness*, *implementation* and *impacts*. The authors define *readiness* as

> "*[. . .] not measuring readiness to start an open government data initiative, but rather readiness to secure positive outcomes from such an initiative. As such, we include measures relating to the existence of open data, and a range of interventions that support engagement with and re-use of open data.*"[5]

The *implementation* measure focuses on the technical parts of publishing Open Data. The report includes, inter alia, if specific data is available, if it is openly licensed or if the data format is machine-readable. The *impacts* are measured by carrying out questions regarding the impact of Open Data on politics, social factors and the economy of a country. An *impacts* question to the experts is for example: "To what extent has open data had a noticeable impact on environmental sustainability in the country?"

[5] `http://www.opendatabarometer.org/report/about/method.html`, last accessed 2015-07-20

In contrast to the approach presented in this thesis, all measures are manually evaluated by country experts. This kind of report is very costly in terms of collaborators and depends highly on the knowledge and the correctness of the judgement of these experts. Further, an expert's judgement is normally based on an excerpt of all Open Data of a country, since the amount of all published data is often too large and wide-spreaded to fully evaluate it.

Global Open Data Index. Similarly to the Open Data Barometer project, the Global Open Data Index[6] assesses and ranks the state of Open Government Data in various countries. It is run by the Open Knowledge Foundation and its evaluation methods are based on a voluntary basis. The score is calculated by looking at the *availability*, *quality* and *openness* of key datasets (e.g., election results or government budget). Volunteer contributors have to answer question regarding the technical and the legal status of openness in a country. For instance, the questionnaire includes the technical question "*Is the data machine-readable?*" and the legal question "*Is the data openly licensed?*". The questions are weighted differently and result in a total score for a country.

Open Data Monitor. In addition, the Open Data Monitor project[7] was recently released and aims to provide a general overview about various data portals. At the current state the project inspects 7 different quality aspects: Open licenses, machine-readable formats, completeness of metadata, non-proprietary formats, total distribution size, unique publishers and software platform. On the project website there is no information on how the scores for the different quality metrics are defined (e.g., the openness of licenses or the completeness of metadata) and if the calculation is manually by user evaluation or automatically done. Further, the platform deals with different metadata sources by mapping the collected data to an internal common schema. This mapping is based on static key-value pairs, where a given metadata field is harmonized if there is a predefined key for this field. For example, if there are mappings *location* $\rightarrow$ *location* and *position* $\rightarrow$ *location* then the fields *location* and *position* can be harmonized but not the field *Location*.

Similarly to our quality assessment approach, the Open Data Monitor project selects a set of metrics for assessing an overall Open Data quality of a portal or country respectively. The project, however, lacks concrete information on how these metrics are computed and it uses metrics where it is unclear in what sense these can be considered as quality aspects (e.g., the number of unique publishers and the used software platform).

3.2.2 Private Sector

Open Company Data Index. The Open Company Data Index[8] is a project supported by the World Bank Institute and can be seen as a counterpart to the Global Open Data

[6] `http://index.okfn.org`, last accessed 2015-09-21
[7] `http://www.opendatamonitor.eu/frontend/web/index.php`, last accessed 2015-07-21
[8] `http://registries.opencorporates.com/`, last accessed 2015-08-23

Index project. This report aims to register open corporate data and provides an aggregated quality assessment per country. The QA score is based on several factors, including the licensing and machine-readability of the data. Furthermore, the report takes the availability of specific datasets into account, e.g., information on significant shareholdings and annual accounts for each company.

Again in contrast to the approach presented in this thesis, we can identify that the quality assessment in the Open Company Data Index report is based on a manual evaluation and depends very much on the expert's judgement.

3.2.3 Quantitative Quality Metrics

OPQUAST. More related to the actual data quality assessment is the OPQUAST project[9] which provides a checklist for Open Data publishing, including questions related to quality aspects. The checklist is very extensive and the question reaches from general questions about the data catalog (e.g., "*The concept of Open Data is explained*") to in-detail questions about specific metadata keys and available meta-information (e.g., "*It is possible to obtain information regarding the level of trust accorded to the data*"). The 72 questions are categorized in 12 categories: Animation, API, Applications, Metadata, Format, Identification, License, Linkeddata, Naming, Transparency, Usability and Privacy.

Kučera et al. (2013). In relation to data quality assessment in Open Government Data catalogues, such as `data.gov` or `data.gv.at`, recent work by Kučera et al. [KCN13] discusses quality dimensions and requirements of such catalogues. Table 3.2 (taken from [KCN13]) lists the proposed quality dimensions *Accuracy*, *Completeness*, *Consistency* and *Timeliness*. Unfortunately, the article is short of detail in some respects. For instance, the accuracy dimension requires that "All information in a catalog record should correspond to the data described by the record." However, there is no further explanation on how to quantify this dimension and on how to assess the accuracy of "all information" in a record. Similarly, the consistency dimension is defined by utilizing a set of semantic rules, but the article lacks explanation and concrete examples of such semantic rules.

Further, the authors propose a set of general improvement techniques, namely data-driven and process-driven techniques. While process-driven methods aim to eliminate the cause of quality issues by modifying the data creation process, data-driven methods directly modify the data itself (e.g., correction of invalid data or normalization).

Braunschweig et al. (2012). An earlier survey in 2012 analyzed 50 Open Data portals wrt. standardization, discoverability and machine-readability of data [BETL12]. The authors propose a set of general (*Global View*) and complex features (*Detailed Analysis*). *Global View* features include measures as the existence of standardized metadata attributes, the existence of an API, respectively the API granularity, and the existence of standardized domain categories. The *Detailed Analysis* looks into the

[9] `http://checklists.opquast.com/en/opendata`, last accessed 2015-07-21

Quality dimension	Definition	Requirements
Accuracy	Extend to which a catalog record correctly describes the data.	All information in a catalog record should correspond to the data described by the record.
Completeness	Portion of the filled in mandatory attributes of a catalog record.	All mandatory attributes of the record should be filled in.
Consistency	Conformance of a catalog record to the set of semantic rules.	There should be no contradiction or discrepancy between the facts in the catalog attributes.
Timeliness	Extend to which a catalog record is up to date.	All information in the catalog record should be up-to-date.

Table 3.2: OGD catalog quality dimensions and requirements, taken from [KCN13]

machine-readability of datasets, the existence and number of tags, or the existence and length of the dataset description.

Unfortunately, computation formulae and details are largely missing in the article and it is unclear how exactly the features can be computed and the findings where derived.

Reiche et al. (2014). Most closely related to the efforts in this thesis is the work of Reiche et al. [RHS14] which also identified the need for an automatic quality assessment and monitoring framework to better understand quality issues in Open Data portals and to study the impact of improvement methods over time. The authors developed a prototype of such a framework which is unfortunately now offline.[10]

Although this paper influenced the metrics and framework presented in this thesis, we extended the work of Reiche et al. in terms of useful quality metrics in the context of Open Data (e.g., by adding an openness metric), in terms of the extent of monitored data portals and in terms of a continuous monitoring of these portals.

Ochoa et al. (2009). In [OD09] the authors discuss a set of quality metrics for metadata in digital repositories. The paper includes a detailed description, definition and evaluation of the metrics on the ARIADNE Learning Object repository[DFC+01]. Inter alia, the authors defined the *conformance*, *consistency* and *coherence* metrics in their work:

Conformance: The conformance to expectations measures the amount of useful information contained in metadata. The proposed method is based on calculation of the entropy value of a metadata instance. Datasets with non-common words or unique descriptions are more easily discoverable within a data catalog.

[10] `http://metadata-census.com`, last accessed 2015-03-06

Consistency: This metric measures to which degree the metadata matches a given metadata standard or metadata schema definition. The authors define three ways of consistency breaking: i) missing mandatory fields or not defined fields included in the metadata, ii) categorical fields contain non-domain values, iii) the combination of different categorical fields within a dataset is not allowed in the standard.

Coherence: The coherence of metadata fields is the degree to which theses fields describe the same object in a similar way. For example, if the values in the fields "title" and "description" in a dataset contain similar words then the coherence value is high.

Based on our analysis of related projects and relevant literature we can conclude that there is no common consensus on the set of used quality metrics and likewise there is currently no comprehensive effort to monitor such metrics across portals in a continuous fashion. However, there are quality aspects occurring in almost all of the reviewed efforts (e.g., the use of open licenses and the machine-readability of formats). Moreover, we can see that current quality assessment in Open Data is mainly based on manual evaluation or expert's rating. Therefore, we identified the need for an automatic and objective quality assessment and monitoring framework.

CHAPTER 4

CKAN specific Quality Assessment Framework

This chapter contains the theoretical foundation and a detailed description of our implementation of a quality assessment framework, tailored to the CKAN software.

In section 4.1 we propose a set of objective quality metrics, provide a formalization of these measures and go into some of the implementation details. In order to illustrate the calculation of the quality measures, we provide an evaluation of an example data record in section 4.2. In section 4.3 we discuss a set of quality and data profiling measures not yet considered in our assessment framework. I.e., we provide a short outlook for possible future research directions of our framework. Eventually, in section 4.4 we describe the current implementation of our monitoring framework in detail.

4.1 Quality Metrics

In this section, we discuss in detail six quality dimensions and metrics, based on the formal definition of a CKAN portal (see section 3.1.1). The metrics are *retrievability*, *usage*, *completeness*, *accuracy*, *openness* and *contactability* (cf. Table 4.1 for a short description). These metrics are based on an exhaustive literature review (section 3.2), partially aligned with existing ones [RHS14] and extended by the openness and contactability dimension.

The metrics we defined herein all show the following characteristics:

- The calculation of the metrics can be done in an automated fashion.
- To improve the comparability within the set of metrics, the metrics are normalized to values in the range between 0 and 1.
- Most of the metrics (except for the *retrievability* metric) work over the values of metadata instances, which contain either text or numbers.

	Dimension	Description
Q_r	Retrievability	The extent to which dataset information and the listed resources in a portal can be retrieved by a human or software agent.
Q_u	Usage	The extent to which metadata keys are used to describe a dataset.
Q_c	Completeness	The extent to which the used metadata keys contain information.
Q_a	Accuracy	The extent to which a selected set of specific metadata keys accurately describe the resource.
Q_o	Openness	The extent to which the license and available formats are suitable to classify a dataset as open.
Q_i	Contactability	The extent to which the data publisher provide useful contact information.

Table 4.1: Quality metrics together with their informal description.

In the further course of this section we provide the necessary details and formulae, and go into some implementation details.

4.1.1 Retrievability

Our first quality metric concerns *retrievability* of dataset information from a portal and for the actual resources on that portal. The metric measures if a legal or software agent can retrieve the content of a portal and its resources without any errors or access restrictions.

Definition 2 (Retrievability) *The degree to which the description of a dataset and the content of a resource can be retrieved based on an* `HTTP GET` *operation. Let the function* $\text{status}(d)$ *(or* $\text{status}(r)$*, respectively) return the Hypertext Transfer Protocol (HTTP) response status code of a* `HTTP GET` *request for a particular dataset* d *or resource* r*, both identified by a URL. Further, let* $\text{ret}(d) = 1$ *if* $\text{status}(d)$ *equals 200, otherwise 0, analogously for* $\text{ret}(r)$*. The aggregated average retrievability for the datasets, denoted* $Q_r^{\mathcal{D}}$*, and the for resources, denoted* $Q_r^{\mathcal{R}}$*, of a portal is defined as:*

$$Q_r^{\mathcal{D}}(p) = \frac{\sum_{x \in \mathcal{D}(p)} \text{ret}(x)}{|\mathcal{D}(p)|} \tag{4.1}$$

$$Q_r^{\mathcal{R}}(p) = \frac{\sum_{x \in \mathcal{R}(p)} \text{ret}(x)}{|\mathcal{R}(p)|} \tag{4.2}$$

Note, one would expect a retrievability on a dataset level to be very close to '1' since CKAN portals are generally open and have a central access point. However, we experienced

that some portals return a `403 Forbidden` status code on an `HTTP GET` request and therefore introduced this measure as such. Additionally, since resources are typically maintained by their providers and hosted on an external server, we expect to have smaller retrievability values due to outdated links or also server errors for retrieving the actual linked resources.

4.1.2 Usage

Our second quality metric is the availability or *usage* of metadata keys across the datasets of a portal. We use this measure since we observed that not all portals make all metadata keys available to the data publishers or because keys can be left out if publishers use the CKAN API. While this usage metric is a rather weak quality measure, it can be used either as a weight for other quality formula or as a filter; e.g., one can compute a certain metric by considering only the keys which are used in all datasets. We define the usage quality metric as follows:

Definition 3 (Usage) *The* usage *defines the degree to which the available metadata keys are used in the datasets of a given portal. We define the usage of a dataset d as:*

$$\text{usage}(d) = \frac{|\,\text{keys}(d|\mathcal{K}^{\mathcal{C}} \cup \mathcal{K}^{\mathcal{E}})| + \sum_{r \in d} |\,\text{keys}(r)|}{|\,\text{keys}(p|\mathcal{K}^{\mathcal{C}} \cup \mathcal{K}^{\mathcal{E}})| + (|\,\text{keys}(p|\mathcal{K}^{\mathcal{R}})| * |\,\text{res}(d)|)} \tag{4.3}$$

and for a portal p as the average of the key usage per dataset:

$$\text{Q}_{\text{u}}(p) = \frac{\sum_{d \in \mathcal{D}(p)} \text{usage}(d)}{|\mathcal{D}(p)|} \tag{4.4}$$

In addition, and to get a more fine-grained analysis, we also define the usage of a key k in a portal p over all datasets as follows:

$$\text{usage}(k, p) = \frac{\sum_{d \in \mathcal{D}(p)} \text{usage}(k, d)}{|\mathcal{D}(p)|} \tag{4.5}$$

$$\text{usage}(k, d) = \begin{cases} 1 & \text{if } k \in \text{keys}(d) \\ 0 \end{cases} \tag{4.6}$$

This definitions also allows us to investigate the usage for a certain subset of keys.

4.1.3 Completeness

The completeness of the metadata description is a widely used and important measure to provide an indication of how much meta information is available for a given dataset.

Definition 4 (Completeness) *The completeness of a portal is the degree to which the available metadata keys to describe a dataset have non empty values.*

Slightly reformulating the metric in [RHS14], we define the completeness function for a key k and a dataset d as $\text{compl}(k, d)$*, returning 1 if $k \in \mathcal{K}^{\mathcal{C}} \cup \mathcal{K}^{\mathcal{E}}$ and if the value of key*

k in $d \neq Null$*. If the key is a resource key (*$k \in \mathcal{K}^{\mathcal{R}}$*), the function returns the average completeness of k over all resources in d , otherwise it returns 0. More formally speaking:*

$$\text{compl}(k,d) = \begin{cases} 1 & \textit{if } k \in \mathcal{K}^{\mathcal{C}} \cup \mathcal{K}^{\mathcal{E}},\ \textit{and } \text{value}(k,d) \neq Null \\ \frac{\sum_{r \in res(d)} \text{compl}(k,r)}{|res(d)|} & \textit{if } k \in \mathcal{K}^{\mathcal{R}} \\ 0 & \textit{oterhwise} \end{cases} \tag{4.7}$$

The completeness of a resource in the above formula is computed the following:

$$\text{compl}(k,r) = 1 \textit{ iff } \text{value}(k,r) \neq Null$$

The completeness of a dataset, respectively portal, is then calculated as follows:

$$\text{compl}(d) = \frac{\sum_{k \in \text{keys}(d)} \text{compl}(k,d)}{|\,\text{keys}(d)|} \tag{4.8}$$

$$\text{Q}_\text{c}(p) = \frac{\sum_{d \in \mathcal{D}(p)} \text{compl}(d)}{|\mathcal{D}(p)|} \tag{4.9}$$

4.1.4 Accuracy

The accuracy metric reflects the degree of how accurate the available metadata values describe the actual data. Here one can distinguish between a syntactic and semantic accuracy [BCFM09]. Syntactic accuracy is defined as the closeness of a value to the corresponding definition domain, e.g., does the value of an author email correspond to the email format, or do date values conform to a particular formats. In the context of Open Data portals, syntactic accuracy on the one hand would mean for example to check if the value of the metadata key *size* is of type integer. On the other hand, semantic accuracy compares the value with its real-world value; e.g, comparing the content size value with the real content size [RHS14].

Most commonly, one defines different distance functions for different keys and also limits the set of keys which are used to compute the overall accuracy for a dataset. In the following definition the limited key-set is denoted by $\mathcal{K}^*$. In general, we define the accuracy of a metadata key for a portal as follows:

Definition 5 (Accuracy) *The* accuracy *is the degree of closeness between metadata values and their actual values. In general, let* $\text{accr}(k,r)$ *be the normalized distance function for a certain key and a resource returning a value between* 0 *and* 1*, whereby* $\text{accr}(k,r)$ *returns* 1 *if* $\text{value}(k,r)$ *(i.e., the value of the metadata key k) accurately describes the actual resource and* 0 *if* $\text{value}(k,r)$ *is inaccurate. In particular, this is also the case if* $\text{value}(k,r) = Null$*. Further, let*

$$\text{accr}(k,d) = \frac{\sum_{r \in \text{res}(d)} \text{accr}(k,r)}{|\,\text{res}(d)|} \tag{4.10}$$

be the average accuracy for key k in a dataset d over all dataset resources. In order to aggregate the accuracy of a dataset over a predefined set of keys, let

$$\mathrm{accr}(d) = \sum_{k \in \mathrm{keys}(d|\mathcal{K}^*)} \mathrm{accr}(k,d) / |\,\mathrm{keys}(d|\mathcal{K}^*)| \tag{4.11}$$

be the accuracy for a dataset for the defined set of keys (i.e., $\mathcal{K}^* \subseteq \mathcal{K}$).

In the following, we define both the key-specific accuracy of a portal p with a given key k, as well as the accuracy of a portal p with the accumulated accuracy over the set of predefined and relevant keys $\mathcal{K}^*$:

$$\mathrm{Q_a}(k,p) = \frac{\sum_{d \in \mathcal{D}(p)} \mathrm{accr}(k,d)}{|\mathcal{D}(p)|}, \text{ with } k \in \mathcal{K}^* \tag{4.12}$$

$$\mathrm{Q_a}(p) = \frac{\sum_{d \in \mathcal{D}(p)} \mathrm{accr}(d)}{|\mathcal{D}(p)|} \tag{4.13}$$

As of now, we only compute the semantic accuracy for different keys, which is ideally measured by inspecting the content of all resource files. However, retrieving the actual data can be very resource consuming in terms of bandwidth and disk space. As such, we decided to perform a HTTP `HEAD` lookup for all resources and store the response header. We automatically compute the accuracy values using these header information for the following keys:

format (*file format accuracy*):

In algorithm 4.1 we present the pseudo code of our algorithm which is used to compute the format accuracy if a file-extension and/or HTTP-header format information is available. The functions `get_format` and `guess_extensions` in line 8 and 13 respectively, make use of Python's *mimetypes*[1] package. `get_format` normalizes the extension and `guess_extensions` maps the mime-types to a set of common file extensions (e.g., it maps "application/vnd.ms-excel" to ".xls", ".xlsx", ".xlsm", ...).

Initially, our algorithm checks in line 4 if there is any format description available in the metadata description of the resource.[2] In order to check the accuracy of the specified format value for a given resource, we normalize the specified metadata value using the `get_format` function (e.g., mapping ".csv" to "csv") and we compare the normalized value to the file extension of the resource if available in line 8. In addition, we take the format specification in the HTTP `content-type` header field into account (see line 13 of algorithm 4.1). The function `guess_extensions` returns a list of possible extensions of the corresponding resource mime-type. If one of the possible extensions matches with the format in the metadata description of the resource then we increase the score of the outcome value.

[1] `https://docs.python.org/2/library/mimetypes.html`, last accessed 2015-09-22

[2] Please remember that, according to our definition, *Null* can be both, either an empty string or a Null value.

Algorithm 4.1: Code fragment for calculating the format accuracy value.

```
1  def format_accuracy(meta_data, resource):
2    score = 0
3    count = 0
4    if meta_data.format is not Null:
5      // check file extensions
6      if resource.extension is not Null:
7        count = count + 1
8        if get_format(resource.extension) = meta_data.format:
9          score = score + 1
10     // check mime type
11     if resource.header.mime_type is not Null:
12       count = count + 1
13       for ext in guess_extensions(resource.header.mime_type):
14         if get_format(ext) = meta_data.format:
15           score = score + 1
16           break
17   return score/count
```

mime-type (*mime-type accuracy*):
Similarly to the file format accuracy described above, we use the value of the HTTP `content-type` header field. We simply compare the specified resource metadata mime-type value with the value of the content-type header field and return 1 if the values match and 0 otherwise.

size (*content size accuracy*):
The accuracy of the content size can be computed based on the information specified in the HTTP `Content-Length` header field. At the moment, our algorithm tries to match the size values of the HTTP header field and the metadata description of the resource by taking the following steps: Initially, we parse the metadata value and extract only the numerical content; e.g., if there is the value "1024 byte" we extract 1024. Then we (i) try to directly match the values, (ii) convert the metadata value to kilobyte and (iii) to megabyte. In each case we allow a range of 10% of the HTTP-header content-length for matching the values.

The reason why we selected these three measures is that we can determine the distance in an automatic way by inspecting the HTTP-header fields of the resource. Secondly, we believe that the accuracy assessment of those keys is important information for the data consumers and also for the value of a data portal. Furthermore, there is only a limited set of keys which can be assessed automatically. For instance, it is impossible to automatically calculate the semantic accuracy of the `description` or the `author` metadata information.

In future work, we plan to extend the accuracy calculation to other metadata fields, e.g., assessing the accuracy of the `language` field by using language detection on the actual resource.

4.1.5 Openness

Our next metric focuses on measuring how "open" the data of a portal is. One criterion for openness is the used license for a dataset, e.g., if this respective license allows to share and reuse the data. Another criterion is the actual data format, which should be based on an open standard.

Definition 6 (Openness) *The openness of a portal is the degree to which datasets provide a confirmed open license and if the resources are available in an open data format. Let* $\text{open}(d)$ *be a user defined function that determines the openness of a dataset based on the license (subscript* l*) or based on the available formats for the resources of a dataset (subscript* f*). The average openness of a portal is computed as follows:*

$$\mathrm{Q}_{\mathrm{o}}^{\mathrm{l}}(p) = \frac{\sum_{d \in \mathcal{D}(p)} \text{open}_l(d)}{|\mathcal{D}(p)|} \tag{4.14}$$

$$\mathrm{Q}_{\mathrm{o}}^{\mathrm{f}}(p) = \frac{\sum_{d \in \mathcal{D}(p)} \text{open}_f(d)}{|\mathcal{D}(p)|} \tag{4.15}$$

License Openness

We confirm the license openness per dataset by evaluating if the specified license is conform with the list provided by the Open Definition[3]. This list contains details about 109 different licenses including their typical id, url, title and an assessment if they are considered as "open" or not. The license information of a dataset in CKAN can be described with three different CKAN keys, namely `license_id`, `license_title` and `license_url`.

In our framework we implemented and extended the license matching heuristic proposed by Blaim [Bla14]. The algorithm tries to match a dataset license to one of the defined ones in the list by performing the following steps. Firstly, we try to perform the match using the `license_id` value, if available. If this check fails we use next the `license_title`, which is matched either against the id or title in the Open Definition license list.[4] If this check also fails, we use as a fall back solution the `license_url` value for the match. Once a match was successful we decide on the openness based on the assessment of the Open Definition.

Please note, that as such, our metric reports only on the confirmed licenses. It might be that the non-confirmed licenses are also adhering to the Open Definition. Further, we

[3] `http://licenses.opendefinition.org/licenses/groups/all.json`, last accessed 2015-09-21

[4] We perform the additional id match because we observed that in several cases the datasets contain the license id in the license title field.

currently do not pay attention to any occurring inconsistencies, i.e., the corresponding metadata fields describe different licenses. At the moment, we use the first matching license for our openness assessment, as described above.

Format Openness

The format openness metric has to consider that a dataset can have various resources with different formats and we mark a dataset as open as soon as one resource of the dataset has an open format. Regarding the openness of a format we currently use the following fixed set of file formats:

$$\{csv, html, latex, dvi, postscript, json, rdf, xml,$$
$$txt, ical, rss, geojson, ods, ttf, otf, svg, gif, png\}$$

Please note that we require the formats to adhere to the "open standards" as introduced in section 2.3. For instance, we excluded the *xls* format (Microsoft Excel) since there exists not yet a clear agreement if it should be considered as open.

Again, we only measure the confirmed open formats and might miss other formats that can be considered as open but that are not included in our list. However, we can easily adapt this by including new formats or licenses as required and identified. In future work, we plan to adapt this metric in order to take the "machine-readability" of formats into account.

4.1.6 Contactability

Another important issue concerning datasets in Open Data portals is the contactability of their creates/maintainers, that is, if information are available to a user to contact the data provider. This metric is not commonly used, however, it is important that data consumers are able to contact the data providers in case of problems or questions. A CKAN portal has four default metadata keys that allow a data publisher to provide contact details: the author and maintainer email field. One way to measure the contactability is to check if the contact information fields contain values, or fields are syntactically valid email addresses.

Definition 7 (Contactability) *The degree of which the datasets of a portal provide a value, an email address or HTTP URL to contact the data publisher. To provide less restrictive contactability results, we define the* $\mathrm{Q_i}$ *metric, indicating that the dataset has some kind of contact information by adapting the completeness metric for a particular set of keys. Let* $\mathrm{cont}(d)$ *return* 1 *if and only if* $\mathrm{compl}(k, d) = 1$ *for one of the following CKAN metadata fields:* `maintainer`, `maintainer_email`, `author` *and* `author_email`.

$$\mathrm{Q_i}(p) = \frac{\sum_{d \in \mathcal{D}(p)} \mathrm{cont}(d)}{|\mathcal{D}(p)|} \tag{4.16}$$

Further, let $\text{cont}_\text{e}(d)$ *be a verification function that returns "1" if a dataset d has an email address in the corresponding metadata field and "0" otherwise, respectively, let* $\text{cont}_\text{u}(d)$ *be the function to denote if a dataset has a maintainer or author http address:*

$$\text{cont}_\text{e}(d) = 1 \textit{ iff } \text{value}(\texttt{maintainer\_email}, d) \textit{ or } \text{value}(\texttt{author\_email}, d) \textit{ is a valid email address (checked with regex)} \tag{4.17}$$

$$\text{cont}_\text{u}(d) = 1 \textit{ iff } \text{value}(\texttt{maintainer}, d) \textit{ or } \text{value}(\texttt{author}, d) \textit{ is a valid URL (checked with regex)} \tag{4.18}$$

Then the average email- and URL-contactability of a portal is computed as follows:

$$\text{Q}_\text{i}^\text{e}(p) = \frac{\sum_{d \in \mathcal{D}(p)} \text{cont}_\text{e}(d)}{|\mathcal{D}(p)|} \tag{4.19}$$

$$\text{Q}_\text{i}^\text{u}(p) = \frac{\sum_{d \in \mathcal{D}(p)} \text{cont}_\text{u}(d)}{|\mathcal{D}(p)|} \tag{4.20}$$

4.2 Example Evaluation

In this section we show the evaluation of the introduced metrics using a selected CKAN dataset (Figure 4.1). While the original dataset contains additional keys and information, we reduced this example[5] to better demonstrate the metrics' calculation.

Usage. Computing the usage of a single key is simply done by checking its existence within the dataset (formalized in Equation 4.6), e.g.:

$$\text{usage}(\text{"}\textit{organization}\text{"}, d) = 1$$
$$\text{usage}(\text{"}\textit{language}\text{"}, d) = 0$$

Further, to compute the degree to which certain metadata keys are available for a portal we use Equation 4.6. The underlying portal `data.graz.gv.at` currently holds 151 datasets. Assuming that the key *language* is in 15 datasets available, the aggregated usage for this key in the portal would result in $\text{usage}(\text{"}\textit{language}\text{"}, p) = 15/151 \approx 0.1$.

The evaluation of the usage over all keys for the given dataset (Equation 4.3) requires the information of how many unique keys occur in the underlying portal. For demonstration purposes, here we assume that there are 20 unique core keys, 5 extra keys and 5 resource keys on the portal. Eventually, the usage for the dataset is computed as follows:

$$\text{usage}(d) = \frac{|\,\text{keys}(d|\mathcal{K}^\mathcal{C} \cup \mathcal{K}^\mathcal{E})| + \sum_{r \in d} |\,\text{keys}(r)|}{|\,\text{keys}(p|\mathcal{K}^\mathcal{C} \cup \mathcal{K}^\mathcal{E})| + (|\,\text{keys}(p|\mathcal{K}^\mathcal{R})| * |\,\text{res}(d)|)} = \frac{(13+1)+4}{(20+5)+5*1} = 0.6$$

[5] We removed some irrelevant keys, e.g., `notes_rendered` and `num_tags`.

```
d: {
    "license_title": "Creative Commons Namensnennung",
    "maintainer": "Stadtvermessung Graz",
    "maintainer_email": "stadtvermessung@stadt.graz.at",
    "author": "",
    "author_email": "stadtvermessung@stadt.graz.at",
    "resources": [
        {
            "size": "6698",
            "format": "CSV",
            "mimetype": "",
            "url": "http://data.graz.gv.at/.../Bibliothek.csv"
        }
    ],
    "tags": [
        "bibliothek",
        "geodaten",
        "graz",
        "kultur",
        "poi"
    ],
    "license_id": "CC-BY-3.0",
    "organization": null,
    "name": "bibliotheken",
    "notes": "Standorte der städtischen Bibliotheken...",
    "owner_org": null,
    "extras": {
        "Sprache des Metadatensatzes": "ger/deu Deutsch"
     },
    "license_url": "http://creativecommons.org/.../by/3.0/at/",
    "title": "Bibliotheken"
}
```

Figure 4.1: A CKAN dataset found on the Austrian portal `data.graz.gv.at`.

Completeness. Computing the completeness of certain keys of dataset d (respectively resource r) in Figure 4.1 using Equation 4.7 would for example result in the following values:

$$\begin{aligned} &\text{compl}(\text{"}\mathit{title}\text{"}, d) = 1 \\ &\text{compl}(\text{"}\mathit{mimetype}\text{"}, r) = 0 \\ &\text{compl}(\text{"}\mathit{owner\_org}\text{"}, d) = 0 \end{aligned}$$

Remember, that we consider an empty string as a *Null*-value.

In order to compute the aggregated completeness over all keys for the datasets using Equation 4.8 we sum up the completeness values and divide by the total number of keys:

$$\text{compl}(d) = \frac{14}{18} \approx 0.78$$

Accuracy. Before calculating the accuracy of the resource in the CKAN dataset in Figure 4.1 we have to perform an HTTP-header lookup on the given URL resulting in the following output:

```
HTTP/1.1 200 OK
Date: Sun, 16 Aug 2015 12:22:48 GMT
Server: Apache/2.2.22 (Ubuntu)
Last-Modified: Tue, 12 Jun 2012 10:42:58 GMT
ETag: "63cee-1a2a-4c24421449628"
Accept-Ranges: bytes
Content-Length: 6698
Content-Type: text/csv
```

Comparing the dataset description with the resource URI lookup's header, we observe that out of the three relevant keys (*format*, *mime-type*, *size*), due to the availability of only two keys, we are able to compute the file format and content size accuracy. The distance functions, computed by comparing the header information to the dataset values, for the keys respectively evaluate to 1 (accr("$format$", r) and accr("$size$", r)). Since the dataset holds only one resource, applying Equation 4.10 results in accr("$format$", d) = 1 and accr("$size$", d) = 1 for the dataset.

However, by using the HTTP header of a resource URL we cannot discover and check the actual content size of a resource. We fully rely on the correctness of the information provided by the underlying server. In order to actually compute the true accuracy of a resource we would have to download the file (cf. further work in section 7.1).

Openness. Regarding the license openness of the exemplary dataset, we have to inspect the *license_id*, *license_title* and *license_url*. Since we are able to match the value of *license_id* ("CC-BY-3.0") with the opendefintion license list and `opendefinition.org` considers the given license as open and suitable for Open Data, the license openness metric (see Equation 4.14) evaluates to $\text{open}_l(d) = 1$ for the dataset d.

Similarly, the format openness is obtained by inspecting the value of the *format* key of the resource, "CSV". Since we consider the CSV format as an open format, the format openness metric (see Equation 4.15) for the dataset likewise evaluates to $\text{open}_f(d) = 1$.

Contactability. The dataset in Figure 4.1 provides all possible metadata fields which are used to provide contact and provenance information, namely `author`, `maintainer`, `author_email` and `maintainer_email`. To compute the Q_i metric (Equation 4.16) of the example dataset, i.e., the availability of any contact information, we inspect the

completeness values of these keys. In this case this results in $\text{cont}(d) = 1$, since some of the keys are non-empty.

In order to compute the Q_i^e and Q_i^u metrics (Equation 4.19 and Equation 4.20) we have to perform syntactic checks on the values of the relevant fields. We detect the existence of valid email addresses, but no valid URLs, which leads to the following results:

$$\text{cont}_e(d) = 1$$
$$\text{cont}_u(d) = 0$$

4.3 Additional possible Metrics not yet considered

In the following, we discuss a set of quality and data profiling measures not yet considered in our assessment framework. This section is intended to provide a short outlook of possible future research directions of this framework. We currently investigate on how to integrate these ideas in the near future in our monitoring system.

4.3.1 Timeliness

The timeliness is in general a measure of how sufficiently up-to-date a dataset is for a certain task (e.g., live timetables or current election results). However, it is hard to automatically understand the time dimension from the content of a dataset, e.g., to distinguish between historical data vs. real-time data.

There are already existing efforts towards capturing and archiving the changes on the Web of Data [UMP15] and related Linked Data efforts dealing with the observation of changes within the RDF content [KAU+13].

4.3.2 Information richness/uniqueness

Another frequently used metric [RHS14, OD09] is the information richness of the metadata description typically measured by how much unique information is provided compared to all other datasets. Nevertheless, portal owners want in certain cases a low uniqueness value for certain metadata keys. For instance, all datasets should be published under the same license. Also, a common pattern[6] is to publish the same type of data but for different timestamps (grouped by year or month), in which case, the meta information differ only by the time value and the uniqueness value would be low again. These observations need to be carefully considered for the overall value.

4.3.3 Interoperability

In order to better verify and measure the heterogeneity which may be present in the open data landscape, we are currently working on a new *interoperability* metric. This

[6] `https://www.data.gv.at/katalog/dataset/ogdlinz_geburten`, last accessed 2015-09-23. This dataset consists of multiple resources where each resource corresponds to a specific year.

metric is intended to be able to estimate the state of heterogeneity based on different factors. A possible factor affecting the interoperability of two different datasets is the compatibility of licenses. Note, that the task of assessing the compatibility of licenses is a non-trivial and diverse one on its own [HK08].

Other possible sources of interoperability are the integration of different file formats or mutually shared tags or categorizations of datasets. While it is relatively easy to assess shared keywords or classifications, the interoperability of formats depends on various factors. Here we can identify three levels of format interoperability.

First, we may encounter format incompatibilities within the same file format description. For instance, since there is no official standard for the CSV file format, CSV files often differ in terms of field separator, line separator and character encoding. Recent efforts by the W3C[7] aim towards a better interoperability of CSV files by providing attached metadata which exactly describe the CSV tables located on the web.

Secondly, it is easy to see that two formats can be integrated if they have the same underlying data structure, e.g., both datasets describe tabular-based or tree-based resources. These would be measures for interoperability on a purely syntactic level.

Thirdly, one could inspect the file content in order to assess the level of integrability, e.g., by considering the header information of tabular resources. Using this information would allow a semantic interoperability measure. Further, one can define the level of integration of different file formats by taking the openness and proprietary of formats into account.

4.3.4 Accuracy by Inspection

A main future target for our monitoring system is the accuracy assessment of a resource description in a dataset based on the inspection of the actual linked data file. At the moment, we use the resource URL's HTTP lookup header for calculating the accuracy of a resource description, and therefore we fully rely on the information provided by the server. The inspection of the actual data file would allow us, for example, to find out the real file size, and to assess the correctness of the `last-modified` metadata information. However, this accuracy assessment by resource inspection requires the download and storage of the actual resources which can be very time and space consuming.

4.4 Automated Quality Assessment Framework

We developed a monitoring framework, termed "*Open Data Portal Watch*", to continuously assess the quality of open data portals similar to [RHS14]. Currently the monitoring framework retrieves datasets from CKAN (126), Socrata (102) and OpenDataSoft (11) portals. Please note that at the moment the quality assessment step, however, only includes CKAN portals. The aim of future work is a full integration of Socrata and OpenDataSoft portals into the quality assessment process (see Further Work in chapter 7).

[7] `http://www.w3.org/2013/csvw/wiki/Main_Page`, last accessed 2015-09-24

The components of our framework are depicted in Figure 4.2. The *fetch component* periodically retrieves the dataset information of a given portal and stores the meta data in a document store. The stored information is analyzed by the *quality assessment component*. This component computes our defined quality metrics for the various dimensions introduced in section 4.1. A publicly available *dashboard component*[8] presents and reports our findings.

The fetching and analysis code is implemented in Python and all data are stored in a MongoDB instance. We also make all snapshots of collected raw meta data for all monitored portals publicly available to motivate and engage other researchers in analysing it.[9]

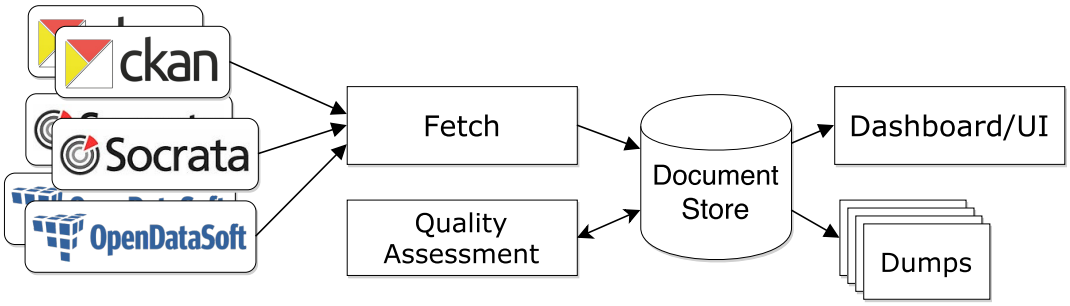

Figure 4.2: The *Open Data Portal Watch* components

4.4.1 Fetching component

The fetching component of our framework runs periodically – once a week – and downloads all datasets found on the portals in our system. Additionally, it performs in parallel HTTP header requests on the resources specified in the datasets. A list of all observed portals, along with the date when we added the portal to the system, can be found online on our dashboard component.

Dataset fetching. The basic idea of downloading and storing all datasets of all portals is represented in algorithm 4.2. An execution of the fetching process is associated with the starting time, referred to as *snapshot*. Initially, we request a list of all datasets hosted by a portal. Next, we download each dataset description and store the retrieved metadata in our database. For some portals (depending on the software and the API version) we are able to retrieve multiple datasets at once and thereby speed up the fetching process. In order to avoid getting locked out by the server due to an excessive amount of HTTP requests, we wait for a short amount of time before the next query after retrieving a dataset (cf. web crawling politeness [NH02]). In our actual implementation

[8] http://data.wu.ac.at/portalwatch/
[9] http://data.wu.ac.at/portalwatch/data

we parallelized the outer loop of the algorithm, i.e., the fetching of all datasets of a specific portal.

Beside adding the raw retrieved metadata to the document store, we store a number of additional information during the fetching process:

- The current date associated with the fetching process. It serves as a key to associate the retrieved datasets with a specific snapshot.
- The HTTP status code for the portal's main entry URI (e.g., `200` if the portal is accessible).
- The HTTP status code per dataset URI. We need this results to calculate our previously introduced retrievability metric.
- Any occurring connection or API errors.

Algorithm 4.2: Pseudocode for fetching metadata from a set of portals.

```
1  def fetch(portals, analyzers):
2    // a download is associated with the current date
3    snapshot = date.today()
4    for portal in portals:
5      // retrieve a list of all dataset on portal
6      datasets_list = get_datasets(portal.api)
7      for id in datasets_list:
8        dataset = get(portal.api, id)
9        // store the retrieved dataset in a DB
10       store_in_db(dataset, snapshot)
11       // analyze the dataset and store the results in the DB
12       for analyzer in analyzers:
13         analyzer.analyze(dataset)
14       store_in_db(analyzers.result, snapshot)
15       // in order to avoid getting locked out by the server
16       // we wait about one second before the next query
17       sleep()
```

For CKAN, we currently use the API version 3 if it is available. Since this version is not supported by all portals we also use API version 2 as a backup variant.[10] The advantage of version 3 is that it provides to option to retrieve multiple dataset at once and therefore it avoids unnecessary requests to the servers.

[10]For version details please see the CKAN API documentation: `http://docs.ckan.org/en/ckan-2.4.0/api/index.html`, last accessed 2015-09-24

Resource fetching. In order to assess the status and quality of resources of a dataset we perform HTTP header requests on the dataset's URLs and store the resulting HTTP header fields in the document store. Currently, this process is done for CKAN resources only since OpenDataSoft and Socrata do not support references to external resources (see subsection 3.1.1). We execute this task independently from the dataset fetching to increase the overall running time of the system. Due to potentially many dead URLs we set the waiting time for a HTTP response to one second before giving up on the resource. However, considering the number of resources in our system (see chapter 6) this can still be very time consuming.

4.4.2 Quality Assessment

The calculation of the quality metrics defined in section 4.1 as well as the collection and assessment of general descriptive statistics is done directly after fetching the dataset (see line 12 in algorithm 4.2). The system is organized in various *analyzers*, each assigned to a specific task. These analyzers work on the retrieved datasets (and similarly on the resources) and collect and aggregate information. For instance, there is an analyzer in our system for counting the total number of different format description and another analyzer for calculating the in section 4.1 introduced completeness metric.

4.4.3 Document Store

The data of the assessment framework is stored in a MongoDB[11] instance. The store consists of a collection of portal objects corresponding to all monitored Open Data portals (see Figure 4.3). In principle, this portal objects store the portal's URL, the APIs, a list of all available snapshots in our system and some additional information.

A portal object corresponds to a collection of dataset objects. This dataset objects are unique by their ID (e.g., the CKAN dataset identified) and the snapshot. The snapshot is represented by an UNIX timestamp of the fetching start time. Further, a dataset object holds the identifier of the belonging portal, the HTTP status code and the raw metadata of the dataset (see *content* in *Datasets* in Figure 4.3). Similarly, there is a collection of resources (if the catalog software supports resources), where a resource object holds its URL, the portal identifier, the corresponding dataset identifier and the HTTP header fields and status code.

The resulting quality assessment (described in subsection 4.4.2) for a snapshot of a portal holds then the identifier of the portal, the snapshot, and the assessed quality metrics together with some descriptive statistics of the snapshot.

4.4.4 Dashboard

The UI component displays vital quality metrics for each portal using various views and charts. An example of a "portal-evolution view" for an Austrian portal is depict in Figure 4.4. The top part shows the evolution of the dataset in the portal and the

[11] `https://www.mongodb.org/`, last accessed 2015-09-24

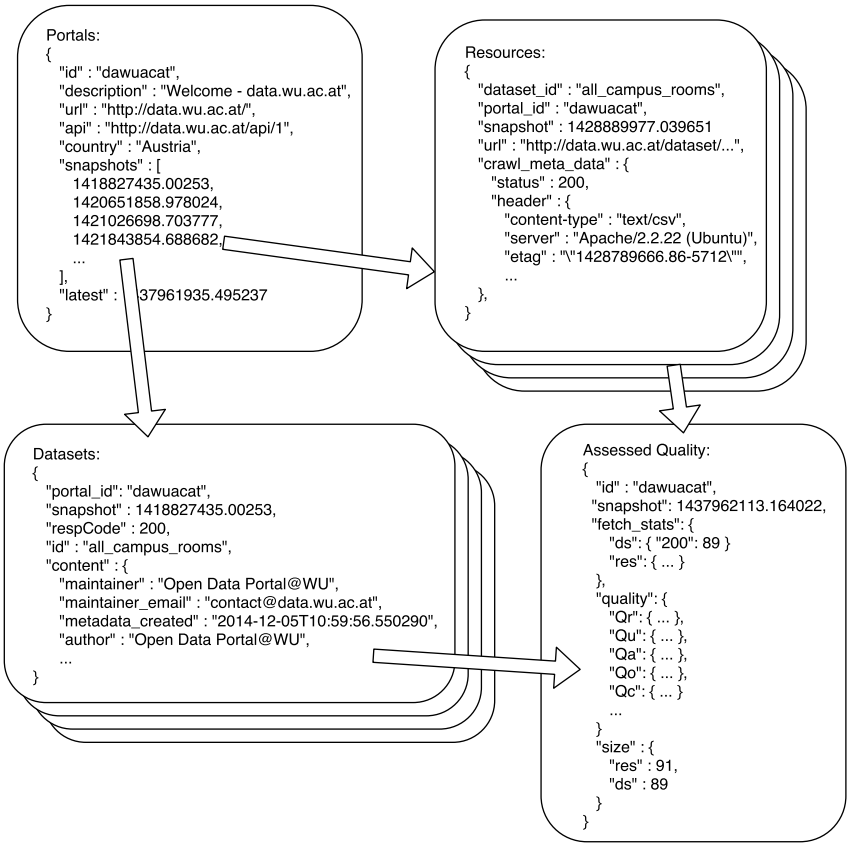

Figure 4.3: Example of a portal snapshot and the corresponding datasets and resources in the document store.

bottom part the values for our quality metrics (each dot is one snapshot). This example shows the usefulness of the monitoring and quality assessment over time since we can clearly see how added or removed datasets influence certain quality metrics.

The frontend dashboard is based on NodeJS and various JavaScript libraries (e.g. the jQuery library for table rendering and interaction and D3.js for the visualizations).

In the next chapter we tackle the issue of highly heterogeneous metadata over different data publishing frameworks. Firstly, we introduce a homogenization approach based on the reviewed metadata schemas in subsection 3.1.1. The homogenization makes use of the metadata specification DCAT (see section 3.1.2). We propose a mapping for the portals introduced in subsection 3.1.1, to achieve a better comparability and homogeneity over the different metadata variants. Secondly, we discuss a possible adaption of the quality measures introduced in this chapter to the homogenized DCAT model.

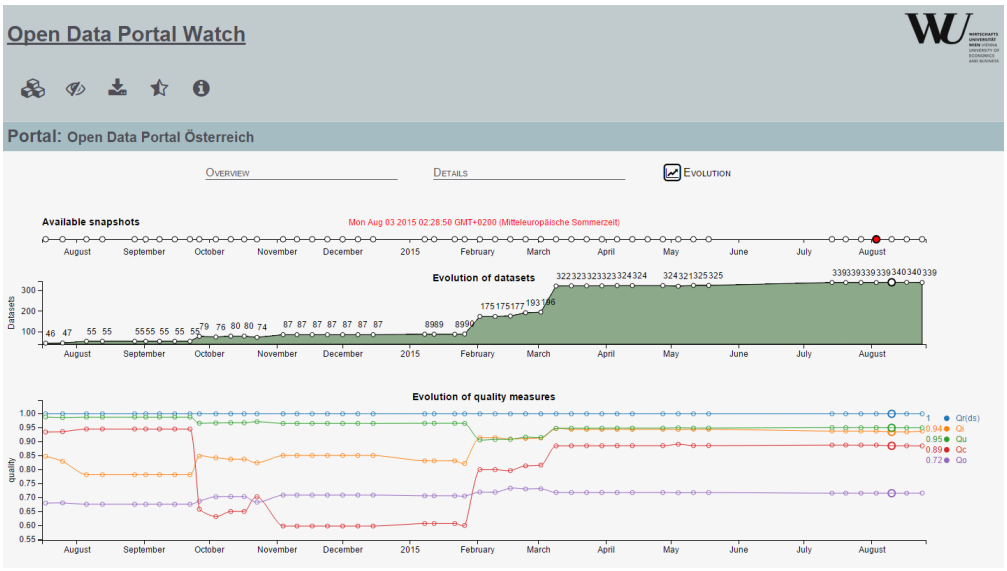

Figure 4.4: Screenshot of an evolution view of a portal.

CHAPTER 5

Towards a general QA Framework

Herein, we examine the potential of generalizing the CKAN specific quality assessment approach introduced in the previous chapter to other Open Data portal software frameworks. In section 3.1 we discussed the structure of different metadata schemas and observed heterogeneity issues amongst metadata from different data publishing frameworks and portals. This observation raises the need of a common schema applicable to a range of metadata sources in order to improve the comparability and integration of data from different portals.

In this chapter we want to contribute to metadata homogenization across portals by proposing a mapping for metadata from different portal frameworks to the common DCAT vocabulary [ME14], cf. section 5.1. Additionally, we discuss the adaption of the quality metrics introduced in the previous chapter to the proposed mapping in section 5.2.

5.1 Homogenized Metadata

Ideally we want to be able to compare and measure the quality of Open Data portals independently from their corresponding publishing software. Therefore, we investigate on homogenization approaches for metadata, suitable for handling different data management systems.

As a first step towards a generalized metadata schema, we propose a manual mapping for metadata schemas observed on CKAN, Socrata and OpenDataSoft portals in subsection 5.1.1. The respective metadata schemas for this frameworks have been introduced in subsection 3.1.1. The proposed mapping is intended as a homogenization of different metadata sources by targeting the W3C's DCAT vocabulary [ME14]: in subsection 5.1.2 we discuss the implementation details of the mapping. In particular, we discuss the adaption and extension of existing DCAT mappings for CKAN and Socrata.

5.1.1 DCAT Mapping

The mapping targets the DCAT vocabulary, introduced in section 3.1.2. Using DCAT brings the following characteristics and advantages:

DCAT follows the Linked Data publishing best practices [BL06].
C.f. section 2.1.2 for the definition.

DCAT is defined in RDF.
Using Linked Data technologies enables better integration of datasets and enables search and discovery of datasets from different sources.

DCAT is a 2014 W3C recommendation.
Originally developed at the Digital Enterprise Research Institute (DERI) in Galway,[1] it was refined and standardized by the World Wide Web Consortium in 2014.

DCAT reuses Dublin Core Metadata vocabulary.
It makes extensive use of the widespread Dublin Core Metadata vocabulary. Furthermore, it uses the `foaf`,[2] `skos`[3] and `vCard`[4] RDF vocabularies.

DCAT profiles enable additional constraints.
Existing DCAT information can be easily extended by using DCAT profiles, which is a specification for data portals that adds additional constraints on the existence and usage of certain properties. For instance, there is the European DCAT-AP profile, introduced in section 3.1.2. DCAT-AP aims towards a deeper integration of data from different sources, aligned to European data portals.

In Figure 5.1 we introduce our mapping for the different metadata keywords. The mapping makes use of the following namespaces, `dcat` is the DCAT vocabulary and `dct` the Dublin Core Metadata vocabulary:

```
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dct:  <http://purl.org/dc/terms/> .
```

The mapping includes metadata keys from Socrata, CKAN and OpenDataSoft mapped to `dcat:`/`dct:` properties. The bold headers in the table indicate a class (i.e. a subject) within the DCAT model; the part after → represents a property. Blank fields within the table indicate, that we were not able to match a corresponding key with the same semantic meaning. Please note, that individual datasets may contain a suitable key, but that we only map default, regularly occurring metadata keys.

For instance, `dcat:Dataset`→`dct:title` denotes that there is an RDF triple

dataset `dct:title` *title*

[1] `http://vocab.deri.ie/dcat`, last accessed 2015-09-18
[2] `http://www.foaf-project.org`, last accessed 2015-09-18
[3] `http://www.w3.org/2009/08/skos-reference/skos.html`, last accessed 2015-09-18
[4] `http://www.w3.org/TR/vcard-rdf/`, last accessed 2015-09-18

Table 5.1: DCAT mapping of different metadata keys.

DCAT	CKAN	Socrata	OpenDataSoft
dcat:Dataset			
→ dct:title	title	name	title
→ dct:description	notes	description	description
→ dct:issued	metadata_created	createdAt	
→ dct:modified	metadata_modified	viewLastModified	modified
→ dct:identifier	id	id	datasetid
→ dcat:keyword	tags	tags	keyword
→ dct:language	language		language
→ dct:publisher	organization	owner	publisher
→ dct:contactPoint	maintainer, author (-email)	tableAuthor	
→ dct:accrualPeriodicity	frequency		
→ dct:landingPage	url		
→ dct:theme		category	theme
dcat:Distribution			
→ dct:title	resources.name		
→ dct:issued	resources.created		
→ dct:modified	resources.last_modified		
→ dct:license	license_{id, title, url}	licenseId	license
→ dcat:accessURL	resources.url	*export URL*[a]	*export URL*[a]
→ dcat:downloadURL	resources.download_url		
→ dct:format	resources.format	*export format*[a]	*export format*[a]
→ dct:mediaType	resources.mimetype	*export mime-type*[a]	*export mime-type*[a]
→ dct:byteSize	resources.size		

[a]Socrata and OpenDataSoft offer data export in various formats via the API

in the resulting mapping, where *dataset* is a `dcat:Dataset` and *title* is the corresponding mapped value (i.e., a RDF literal holding the value of the mapped metadata key).

The proposed mapping of the keys is mainly based on matching names. For instance, considering the mapping of the OpenDataSoft metadata keys, we can see that all mapped keys use the same key-names as the DCAT vocabulary. If the key-names are not matching (as for most of the CKAN keys), we mainly rely on existing mappings, further explained in subsection 5.1.2.

Example. Figure 5.1 displays an application of the proposed DCAT mapping for a previously introduced OpenDataSoft dataset (see Figure 3.4). The DCAT mapping is presented as a graph, where oval nodes represent RDF resources and square nodes represent literals. The DCAT mapping in the given example makes use of blank nodes

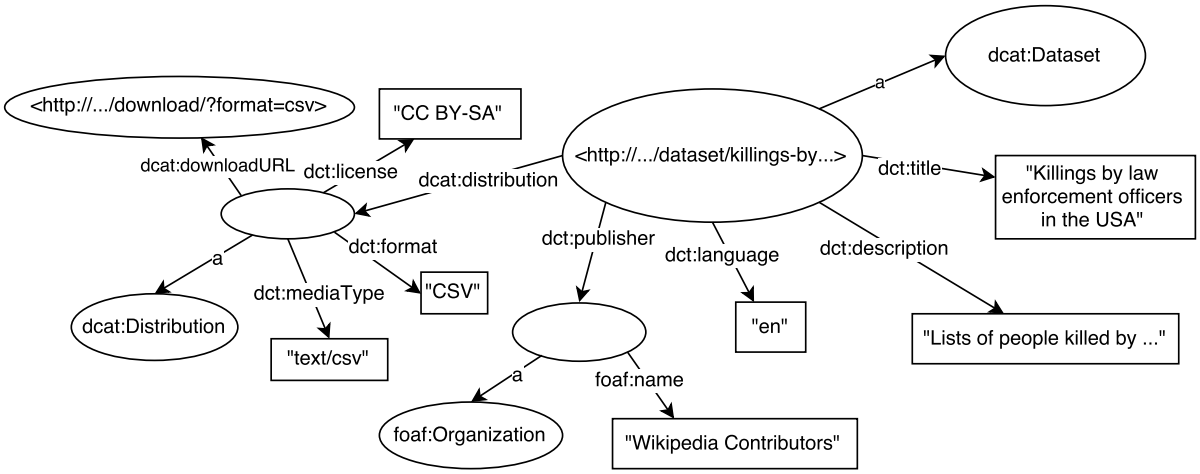

Figure 5.1: DCAT mapping of the dataset in Figure 3.4 in section 3.1.1

(represented as empty oval nodes). For instance, a `dct:publisher` is described using a blank node, which holds a `foaf:name` and is a `foaf:Organization`. Further, note that the `dct:license` in the DCAT model belongs to a distribution, while in the original metadata it is attached to a dataset instance. (This holds likewise for the license keys in Socrata and CKAN portals.)

5.1.2 Adapting existing Mappings

In order to make use of the proposed homogenization within our QA framework (section 4.4) we implemented a mapping algorithm for each of the data management systems covered by Figure 5.1.

Regarding the CKAN software we took a closer look at the source code of the not yet published DCAT extension for CKAN,[5] currently being developed by the Open Knowledge Foundation. We used the existing mapping of datasets mostly "as is",[6] except for the licenses information which is currently not mapped properly: the original mapping in the extension assumes a license key for each resource in a dataset which is not existing in CKAN datasets.

For Socrata portals, we mainly rely on the pre-existing DCAT output (see section 3.1.1). Additionally, we modify the result so that it is conform to the standardized DCAT model. This means, firstly, we replace non-DCAT with standardized DCAT properties in the result if they are synonymous and secondly, we add provenance and authorship information if it is available in the default metadata.

[5] `https://github.com/ckan/ckanext-dcat`, last accessed 2015-08-20. We currently use the code committed on August 13, 2015.

[6]

Regarding the homogenization of OpenDataSoft portals we map the values of the metadata keys as described in Figure 5.1.

5.2 Adapted Quality Metrics

In this section we briefly discuss how we can adapt the previously introduced quality metrics (in section 4.1) to the homogenized DCAT model.

Retrievability. Regarding the retrievability of datasets and resources we can adopt the metric's definition from subsection 4.1.1. In order to evaluate the metric we try to retrieve the URLs specified in `dcat:accessURL` or `dcat:downloadURL`.

Usage. The usage metric as it is defined in subsection 4.1.2 becomes less meaningful because in our homogenization process we only map a selected subset of the available metadata keys to a predefined set of DCAT keys. I.e., during the mapping process information about the overall availability of metadata keys gets lost and the usage of certain keys depends strongly on the specified DCAT mapping. Therefore, the usage results observed by applying this definition are very biased by our proposed mapping.

Completeness. In contrast to the usage metric, the completeness of mapped datasets is meaningful and reasonable. The definition in subsection 4.1.3 can be adopted (up to replacing *resources* by *distributions*) and provides an indication of how much information is available in the underlying dataset with respect to the proposed mapping.

Accuracy. Similarly to the completeness metric, we can adopt the definition of the accuracy metric in subsection 4.1.4 mainly as it is. In order to calculate the specific accuracy for a resource's format, mime-type and size we have to inspect the values of the `dct:format`, `dct:mediaType` and `dct:byteSize` properties.

Openness. Again, we can adopt the definition in subsection 4.1.5 by evaluating the values of the `dct:license` property for the license openness and `dct:format` for the format openness.

Please note, that for some portal software products the format openness is not very meaningful because the products offer various exporting methods for each catalog entry (e.g., Socrata offers a number of formats for exporting resources including CSV, JSON and XML).

Contactability. For adopting the contactability metric in subsection 4.1.6 the DCAT model offers the properties `dct:publisher` and `dct:contactPoint`. Regarding the Q_i metric we have to check if there is any contact information in one of the properties available. Analogous to subsection 4.1.6, the Q_i^e and Q_i^u value is calculated by checking

if the mentioned contact information fields contain values which are syntactically valid email addresses, respectively URLs.

CHAPTER 6

Findings

"You know my methods. Apply them."
— Sir Arthur Conan Doyle, *The Sign of Four*

In this chapter, we present our comprehensive findings about the current landscape of 239 active Open Data portals, including quality assessment results and observed evolutionary patterns. The full list of all current portals is available on the web-interface of our framework.[1]

The portals in this list are based on the CKAN, Socrata and OpenDataSoft software, introduced in subsection 2.2.1 on page 14. At first we will take a look at descriptive results aggregated over all portals. Next, we discuss in detail the quality metrics applied to CKAN portals (introduced in section 4.1 on page 41). We report on heterogeneity observations, evolution patterns and growth-rates of datasets and resources, as well as quality changes over time. Finally, as an example of an in detail analysis using our system, we select the Austrian data portals for discussing monitoring results and metadata conformance.

The results for this chapter are based on snapshots of all data catalogs gathered in the second week of August 2015 (week 33 of the year). The evolution reports are based on snapshots starting in the fourth week of 2015.

6.1 Portals overview

Currently our system holds in total 239 portals, of which are 126 CKAN, 102 Socrata and 11 OpenDataSoft portals. Table 6.1 lists the top and bottom 5 portals with respect to the number of datasets. It is worth noting that 4 out of the top-5 portals are based in North America.

[1] `http://data.wu.ac.at/portalwatch`

Table 6.1: Top and bottom 10 portals, ordered by datasets.

URL	Origin	Software	$\lvert\mathcal{D}\rvert$	$\lvert\mathcal{R}\rvert$	|C.-L.|[a]	%[b]	C.-L.[c]
www.data.gc.ca	Canada	CKAN	244602	1164896	522283	44.8	7814 GB
data.gov	US	CKAN	160049	569180	129003	22.7	967 GB
data.noaa.gov	US	CKAN	64359	619544	127524	20.6	1664 GB
geothermaldata.org	US	CKAN	56179	62406	26206	42	163 GB
publicdata.eu	Europe	CKAN	55459	140278	69325	49.4	189 GB
opendatareno.org	US	CKAN	7	13	3	23.1	2 MB
data.salzburgerland.com	Austria	CKAN	6	34	7	20.6	0.1 MB
www.criminalytics.org	US	Socrata	6	-	-	-	-
datosabiertos.ec	Ecuador	CKAN	3	3	2	66.7	1.8 MB
bistrotdepays.opendatasoft	France	OpenDataSoft	2	-	-	-	-

[a]Number of resources with existing `Content-Length` HTTP header field
[b]Percentage of resources with existing `Content-Length` HTTP header field
[c]Estimated total size of retrievable `Content-Length` HTTP header fields

In addition to the number of datasets and resources, Table 6.1 displays the estimated total size of the resources, based on the number of resources URL's HTTP headers containing the `Content-Length` field. Surprisingly, this header field can be found only in 60% of all HTTP header lookups (cf. section 6.2).

Regarding the CKAN portals, our main source for collecting the set of portals was `dataportals.org`. This service, run by the Open Knowledge Foundation, lists in total 431 Open Data portals, out of which 75 are no longer accessible and 50 are active CKAN portals. Another collection of Open Data portals is provided by the OpenDataMonitor project.[2] This list consists of 217 portals, including 52 CKAN portals. Regarding the Socrata[3] and OpenDataSoft[4] portals, the list of customers can be found on the respective website.

6.1.1 Origin

Table 6.2 displays the origins of the data catalogs in our monitoring system. With 112 portals, around 50% of the observed portals are located in the US and second-most, 14 of the portals are located in UK. Looking into more detail, 90 of the total 112 US portals in the system use Socrata software, whereas only 20 utilize CKAN (however, including the nationwide `data.gov` portal consisting of 160k datasets). Regarding the distribution of the different data management systems one can observe that CKAN is basically uniformly

[2]`http://project.opendatamonitor.eu/`, last accessed 2015-09-24
[3]`https://www.opendatasoft.com/company/customers/`, last accessed 2015-09-24
[4]`https://opendata.socrata.com/dataset/Socrata-Customer-Spotlights/6wk3-4ija`, last accessed 2015-09-24

Table 6.2: Country distribution of the monitored portals.

Country	Total	CKAN	Socrata	OpenDataSoft
	239	126	102	11
US	112	20	90	2
UK	14	11	3	
Germany	9	9		
Italy	9	7	2	
France	9	2		7
Spain	8	7	1	
Austria	7	7		
Canada	7	5	2	
Australia	6	5	1	
Japan	5	5		
Brazil	5	5		
Ireland	4	3		1
Belgium	4	3		1
Netherlands	4	3	1	
Greece	3	3		
others	33	31	2	

distributed, while Socrata can be found mainly in the US and OpenDataSoft portals are mainly located in France.

Table 6.3: Distribution of number of datasets over all portals.

	<50	$<10^2$	$<5\times10^2$	$<10^3$	$<5\times10^3$	$<10^4$	$<5\times10^4$	$<10^5$	$>10^5$
$\lvert\mathcal{D}\rvert$	58	24	78	27	36	8	3	3	2

6.1.2 Datasets

Looking into the number of published datasets over all portals (Table 6.3) one can observe that the majority of 67% contains less than 500 datasets. Please not that the table cells in Table 6.3 should be interpreted as intervals. For instance, in the third column we can see that 78 portals hold between 100 and 500 datasets. There are 5 portals in the system with more than 50k datasets. The largest two portals are Canada's Open Government[5] data catalog consisting of 245k datasets followed by the U.S. government's portal `data.gov` holding 160k entries.

[5] `http://open.canada.ca/`, last accessed 2015-09-21

6.2 CKAN

Currently, we actively monitor 126 CKAN portals, consisting of 745K datasets describing 3.12 million resources. We found 218 different unique licence IDs, 1516 file format descriptions and ∼173k tags (see Table 6.4). The portals use a total of 3900 different meta data keys, of which 240 keys belong to the core or default keys ($\mathcal{K}^C$), 3634 to the extra keys ($\mathcal{K}^E$) and 366 unique keys are used to describe resources ($\mathcal{K}^R$).

Table 6.4: Basic statistics of 126 CKAN portals

$\lvert\mathcal{D}\rvert$	$\lvert\mathcal{R}\rvert$	$\lvert\mathcal{K}^C\rvert$	$\lvert\mathcal{K}^E\rvert$	$\lvert\mathcal{K}^R\rvert$	Licenses	Formats	Tags
745196	3120119	240	3634	366	218	1516	173586

Regarding the portal sizes, Table 6.5 shows the distribution of the portals based on their datasets and resources. We can clearly see that more than half of the portals have less than 500 datasets or resources and around 25% of all portals are in the range between 10^3 – 10^4 datasets or resources. There is one portal in our system holding over a million resources (the Canadian portal).

Table 6.5: Number of CKAN Open Data Portals with a given size.

	$<10^2$	$<5\times10^2$	$<10^3$	$<10^4$	$<5\times10^4$	$<10^5$	$<10^6$	$>10^6$
$\lvert\mathcal{D}\rvert$	35	42	10	32	2	3	2	0
$\lvert\mathcal{R}\rvert$	18	31	19	38	13	3	3	1

We located 3.12M values for the `url` resource meta key, of which 1.92M are distinct values and 1.91M are syntactically valid URLs. Based on the performed HTTP-header lookups, we found the `Content-Length` field in 1.1M unique resource URLs. Summing up this information results in an estimated content size of around 12.297TB.[6]

6.2.1 Portal overlap

The difference between total resource URL values and unique ones indicates that resources are multiple times described, either in the same datasets, portals or across portals. A closer look reveals that 260k unique resource URL values appear more then once, out of which the majority of 227k resource URLs are described in datasets in different portals and the remaining 33k resource URLs are described in the same portal several times.

Looking into the overlapping portals (i.e. resource URLs occurring in datasets in different portals), we discovered that out of the 227k overlapping URLs the majority of about 143k resource URLs appear in the US Government's data portal[7] and second most

[6]Note, that this number is based on HTTP-headers from 1.1 million resource URLs, i.e., only ~60% of the unique resource URLs.

[7]`http://data.gov`, last accessed 2015-09-08

(∼120k) on the US' NOAA (National Oceanic and Atmospheric Administration) portal. Third place is the Pan European data portal.[8] This portal itself contains 140k resource descriptions in total, with about 109k unique resource URLs. Out of these 109k unique URLs, 60k URLs are overlapping resource URLs. The main aim of the Pan European data portal is to harvest other European portals to provide a single-point of access.

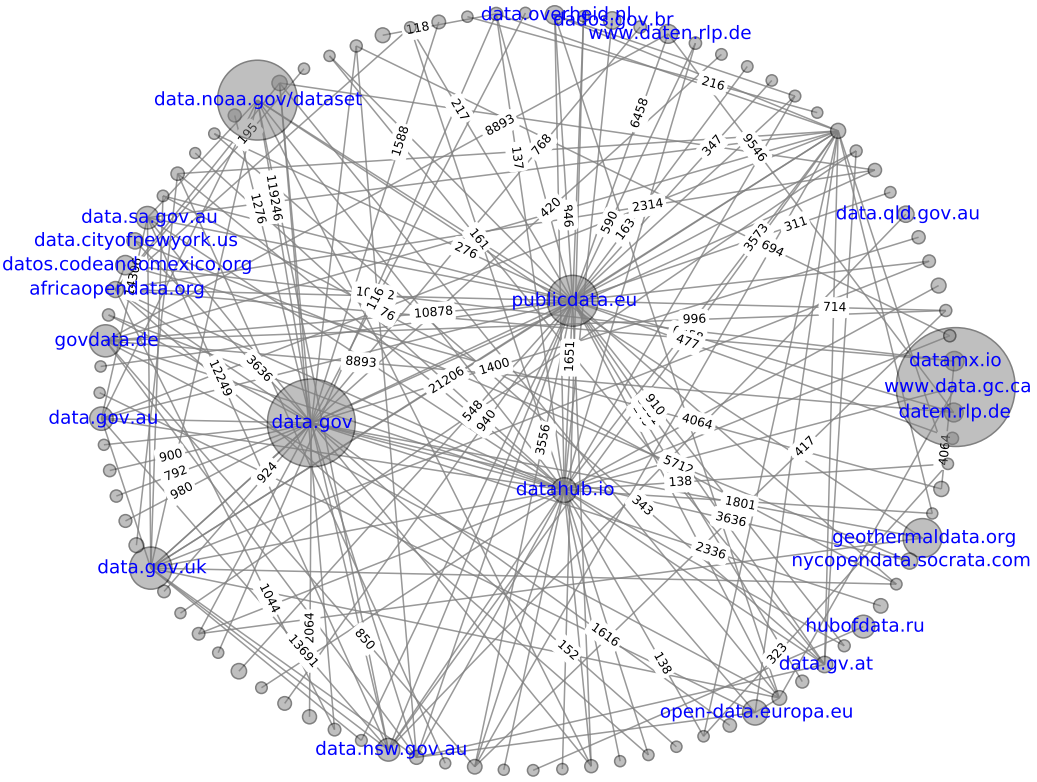

Figure 6.1: Graph of overlapping Resources.

Figure 6.1 provides a graph where the nodes represent a portal within our system and the edges indicate that there exist shared resources. The sizes of the nodes correspond to the number of resources in a portal and the labels of the edges state the number of shared resources.[9] On the one hand, it is recognizable that the nodes for some of the largest portals (`data.gov`, `datahub.io` and `publicdata.eu`) have a very high degree and have many shared resources (indicated by the high numbers on the edge labels). This is an indication that these portals harvest other portals and serve as a superordinated

[8] `http://publicdata.eu/`, last accessed 2015-09-21

[9] In order to preserve readability in the graph we omit isolated nodes and removed labels for edges with less than 100 and portals with less than 5000 resources.

access point. On the other hand, the largest portal in our system, `www.data.gc.ca`, is very isolated and consists of hardly any shared resources.

Further, we can confirm the aforementioned observation that `data.gov` and the US' NOAA portal hold shared resources. But rather than sharing resources among various portals, in fact it is the case that these portals share an edge consisting of 119k resources.

In general, the density of the graph in Figure 6.1 seems to indicate that not only harvesting portals hold shared resources, but rather that data provider tend to pro-actively push their resources to different portals.

6.2.2 Heterogeneity

We observe a large degree of heterogeneity across portals if we look at the overall used extra meta data keys, tags and format values.

A first interesting observation is that out of the 3634 used extra meta data keys a total of 1868 keys appear in only one portal, indicating that extra keys are often portal specific and not much re-use or alignment happens between portals concerning extra keys. We found 1766 keys in more than one portal of which only 261 are used in more than two portals. Only 9 keys are shared in more than 20 portals. Similar observations can be found for the tags used to describe a dataset. Out of the 173586 used tags, 136594 appear in exactly one portal, 36992 in more than one portal and 13085 in more than two portals. There is not a single tag occurring in more than 35 of the 126 CKAN portals.

In this respect it should be noted that tags are in general language specific and therefore, this high heterogeneity is not particularly surprising. In order to gain a deeper insight in the diverse use of tags and, thereby, in the different areas and subjects of Open Data, we plan on integrating and using multilingual semantic networks over the set of tags, such as BabelNet [NP12]. This would allow us a language independent alignment and investigation of tags.

In addition and surprisingly, we discovered 1516 different values to describe the format of a resource. The main reason for this is that there exists no standards for describing the resource formats. For instance, we observed several values for the comma-separate file format such as, *csv*, *comma-separate-values*, *character-separate-values* or *csv-format*, just to name a few.

A similar observation, as already mentioned above, is the diversity of used licence IDs throughout the portals. In total, we discovered 218 different unique licence IDs in the corresponding `license_id` CKAN metadata field. Next, we present for each quality dimension the main interesting findings.

6.2.3 Quality Metrics Evaluation

Retrievability (Q_r)

Table 6.6 shows the results of our dataset and resource retrievability analysis. We grouped the response codes by their first digit; *others* indicate socket or connection timeouts. As expected, nearly all datasets could be retrieved without any errors (99.8%). The 1260

datasets that could not be retrieved responded with a 403 FORBIDDEN HTTP status code, indicating that an access token is required to retrieve the information.

A slightly different picture can be observed if we try to retrieve the content of the actual resources. As mentioned above, out of the 3.12M resource descriptions, ∼1.92M are unique distinct values and 1.91 are valid URLs. We performed lookups on 1.64M URLs, resulting in the response code distribution in Table 6.6. Around 80% of these resources are accessible without any errors or restrictions (resulting in a response code of 2xx). An slightly alarming observation is that 228k described resources (∼14%) point to a non-existing data source and returned a response code of 4xx and 85k resources (∼5%) caused some socket or timeout exception upon the lookup (indicated with others). The number of exceptions should interpreted with caution since the unavailability of the URL might be temporary. In future work we plan to distinct between persistent and temporary errors by considering the evolution of the URL's retrievability.

Table 6.6: Distribution of response codes.

	№	2xx	4xx	5xx	others
$\mathcal{D}$	745196	743929	1260	1	6
$\mathcal{R}$	1635787	1301328	228795	20508	85156

Overall, the retrievability of datasets is very good. Please note, that the resource retrievability of our analysis is based on HTTP-header requests performed on 86% of the unique resource URLs.

Meta Data usage (Q_u) and completeness (Q_c)

Next, we analyze the usage and completeness of meta data keys across all 126 portals. Figure 6.4 on page 72 plots the average usage (Q_u) against the average completeness (Q_c) for each portal for the three different key subsets. Figure 6.2 is a histogram of the Q_c distribution and Figure 6.3 of the Q_u distribution. The distributions also contain the total values over all keys (black bars) for a general overview.

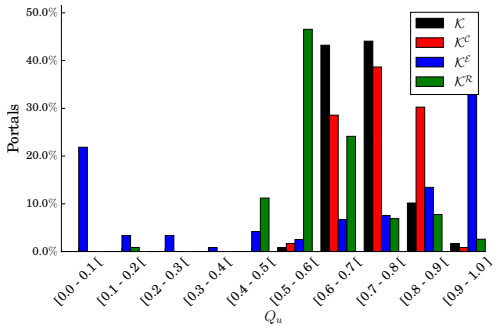

Figure 6.2: Completeness distribution.

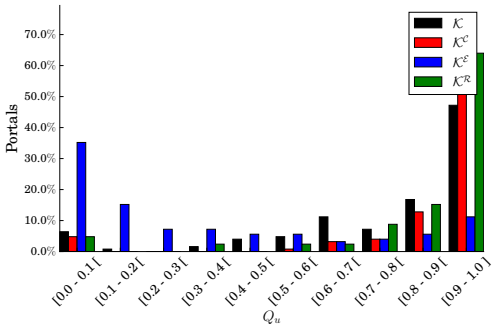

Figure 6.3: Usage distribution.

Looking at the histogram in Figure 6.3, we observe that 60 % of the portals have an average Q_u value per dataset and all keys of more than 0.9. Drilling deeper, we see that nearly all portals have a Q_u value of over 0.9 for the core meta data keys (red bar) and around 70% of the portals have a Q_u value of over 0.9 for the resource meta data keys (green bar). In contrast, the usage value for extra meta data keys is widely spread across the portals with around 70% of the portals having a value of less than 0.4, and about 35% a value of less than 0.1 (see the lower part of the axis and the blue bars).

One explanation for the low usage values of the extra keys might be that for these particular portals the datasets are mainly uploaded by software agents using the CKAN API which does not require that all keys are used and thus are left out. In contrast, a high usage value for portals might be because the datasets are mainly created by using the UI for humans. This UI has a predefined mask using the full set of keys. In addition, the better usage value for the core and resource keys might be because those keys become more standard and documented (e.g. on CKAN documents) and as such, are known to the data publishers and portal specific extra keys might be not well documented or advertised.

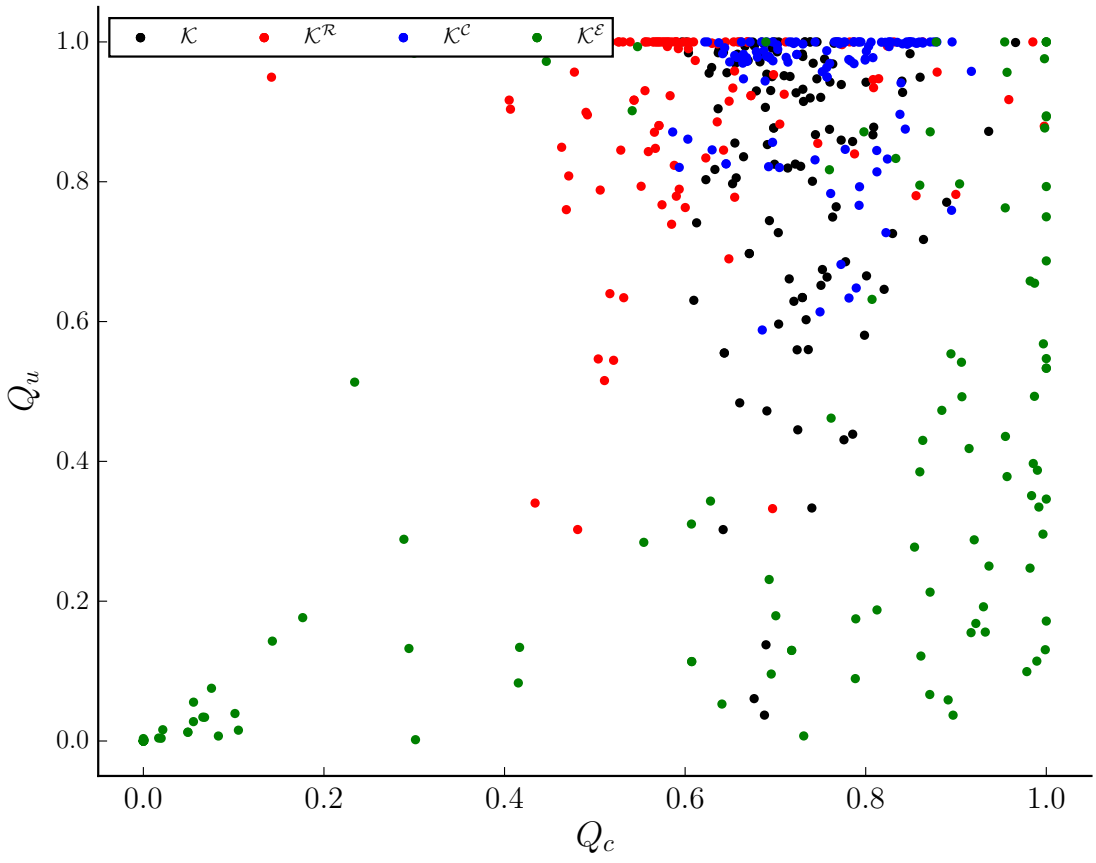

Figure 6.4: Usage and completeness scatter plot.

The histogram in Figure 6.2 shows the respective completeness distribution for the different key sets. Overall, we can observe that the majority of the portals have an average meta data key completeness value (black bar) in the range of 0.6 - 0.8 and only a few portals have value of over 0.9. Inspecting the key subsets, we can see that the overall values are influenced by the set of extra meta data keys which have a completeness value of less than 0.1 for around 22% of the portals. In contrast, over 40% of the portals have a Q_c value of over 0.7 for the core keys. Looking at the key set used to describe resources, we also observe that the majority of the portals provide a specific Q_c value between 0.5 and 0.6.

The scatter plot in Figure 6.4 helps to highlight groups of portals with different average usage and completeness values. For instance, we can see in the bottom left part a group, covering around 35% of the portals, for which the extra keys show very low usage and completeness values. In such a case, a portal owner could rethink the necessity of the extra keys.

Openness (Q_o)

It is crucial for the Open Data movement that published datasets and formats are adhering to the open definition and that everybody is allowed to use, re-use and modify the information which should be provided in an open format. Table 6.7 shows the top-10 used licences per dataset and top-10 used formats per total and unique resources together with their number of portals they appear in. Bold highlighted values indicate that the licence or format is considered by our metric as open. Please note, that we count the number of datasets for the licences and the number of resources for the formats.

Table 6.7: Top-10 formats and licences.

`license_id`	№	%	$\|p\|$	`format`	№	%	$\|p\|$
ca-ogl-lgo	244597	48.6	1	PDF	679440	25.6	89
notspecified	55918	11.1	65	**HTML**	611313	23	70
cc-by	51800	10.3	76	other	182963	6.9	6
uk-ogl	25585	5.1	16	**CSV**	153617	5.8	107
us-pd	21806	4.3	1	originator data format	115434	4.3	1
empty	11858	2.4	23	geotif	95323	3.6	2
dl-de-by-1.0	8862	1.8	4	**XML**	84571	3.2	79
cc0	7254	1.4	1	ZIP	69696	2.6	76
cc-nc	6393	1.3	28	tiff	66045	2.5	10
dl-de-by-2.0	6180	1.2	2	SHAPE	57565	2.2	12
others	63066	12.5		others	542969	20.4	

Formats. The first surprising observation is that ∼25% of of all the resources are published as PDF files. This is remarkable, because strictly speaking, PDF cannot be considered as an open format, and therefore is not suitable for publishing Open Data.

By looking at the top-10 used formats in 6.7, we can see that only 32% of the top-10 formats are covered by open formats. Again, note that for now we do not consider PDF and ZIP files as open.

The occurrence of certain format descriptions within very few portals (e.g., “originator data format”, “geotif”), or even only a single portal, suggests that there are data catalogs which do not stick to conventions or consist of insufficient format specifications.[10]

By considering the plots in Figure 6.5 and Figure 6.6 we can look into detail of the most used formats, providing results for the PDF, CSV and HTML format specification. The x-axis in Figure 6.5 corresponds to a linear subdivision of the number of datasets in a portal, e.g., the largest portal in our system with about 245k is in the range 0.9 - 1.0. The y-axis then gives the distribution of the specified format. For example, more than 80% of the PDF resources can be found in the largest 10% of all portals. As one might expect, the CSV format provides a quite different result. About 90% of all CSV resources are distributed over the smallest 30% of the portals.

The second plot in Figure 6.6 shows the ratio of the format descriptions within the portal. Here the x-axis correspond to the proportion occupied by the specified format. For instance, here we can see that for about 85 to 90% of the portals the ratio of PDF and HTML is less then 10%. Regarding CSV resources we observe that 80% of the portals consist of at least 10% CSV files or more. Using the results of the plots, we can

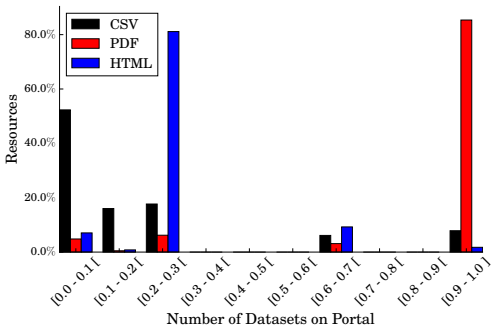

Figure 6.5: Total distribution of specific formats.

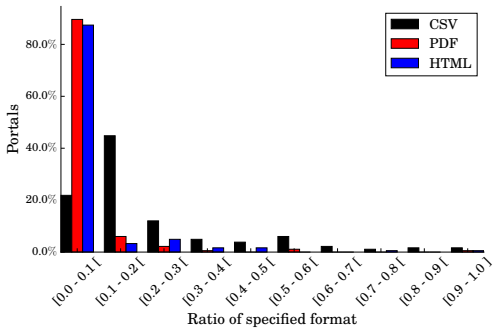

Figure 6.6: Ratio of specific formats in a the portals.

conclude that there are very few but very large portals in our system holding almost all HTML and PDF resources, while the CSV resources are distributed relatively even over most of the portals.

Another interesting observation is that by looking into the 500 most common format descriptions we were able to match the term “csv” (without paying attention to capitalization) 21 times. This indicates that there is no common agreement on how to describe a file format. Even more illustrating, the term “zip” occurred 48 times; e.g., “zip:csv”, “zip (csv)”, “zip+csv”.

[10]Please note that “geotif” in Table 6.7 is not a spelling error.

Table 6.8: Top-10 most used licence IDs grouped by portals.

URL	$\|\mathcal{D}\|$	license_id	№	%
www.data.gc.ca	244602	ca-ogl-lgo	244597	99
data.gov	160049	*empty*	84774	53
data.noaa.gov	64359	*empty*	64286	99
geothermaldata.org	56179	*empty*	56176	99
data.gov	160049	notspecified	48669	30
data.gov	160049	us-pd	21806	14
data.gov.uk	26931	uk-ogl	12225	45
publicdata.eu	55459	uk-ogl	11694	21
data.gov.uk	26931	*empty*	10224	38
publicdata.eu	55459	*empty*	8743	16

Licenses. Regarding the used licence IDs in Table 6.7, we see that the confirmed open licences in the top-10 cover only ∼17% of all datasets. In total, we observed 218 different licence IDs across all CKAN portals. Out of these 218 IDs only 16 IDs occur in the Open Definition list,[11] which we use for our openness evaluation: 9 IDs are approved for licensing Open Data, 2 are rejected and 5 are marked as "not reviewed".

In order to get a better understanding of the distribution and main sources of the most-used licence IDs we display in Table 6.8 the top-10 most used licence IDs grouped by the CKAN portals. Therein we list the total number of datasets on the corresponding portal and the total number of datasets using the concerning licence ID. Additionally, we provide the share of this licence on the corresponding portal. A first interesting observation is that three of the largest portals (www.data.gc.ca, data.noaa.gov, geothermaldata.org) mainly hold the same value for all their datasets. Surprisingly, there is a very high share of *empty* licence ID metadata fields across the portals listed in this table.

In addition, Figure 6.7 shows the distribution of the average Q_o values per portal. From the plot we can see that around 40% of the portals have an average licence openness value of over 0.9 and around 30% of the portals have an format openness value of over 0.9. There is also a group of around 27% of the portals for which we could only confirm an average licence openness per dataset of less than 0.1. The average values for the remaining portal spread more or less equally from 0.1 to 0.9.

Overall, we could not confirm for the majority of the portals that the datasets provide an open licence and their resources are available in open formats. In future work, we will investigate methods to address the unconfirmed licences and formats.

[11] http://licenses.opendefinition.org/licenses/groups/all.json, last accessed 2015-09-21

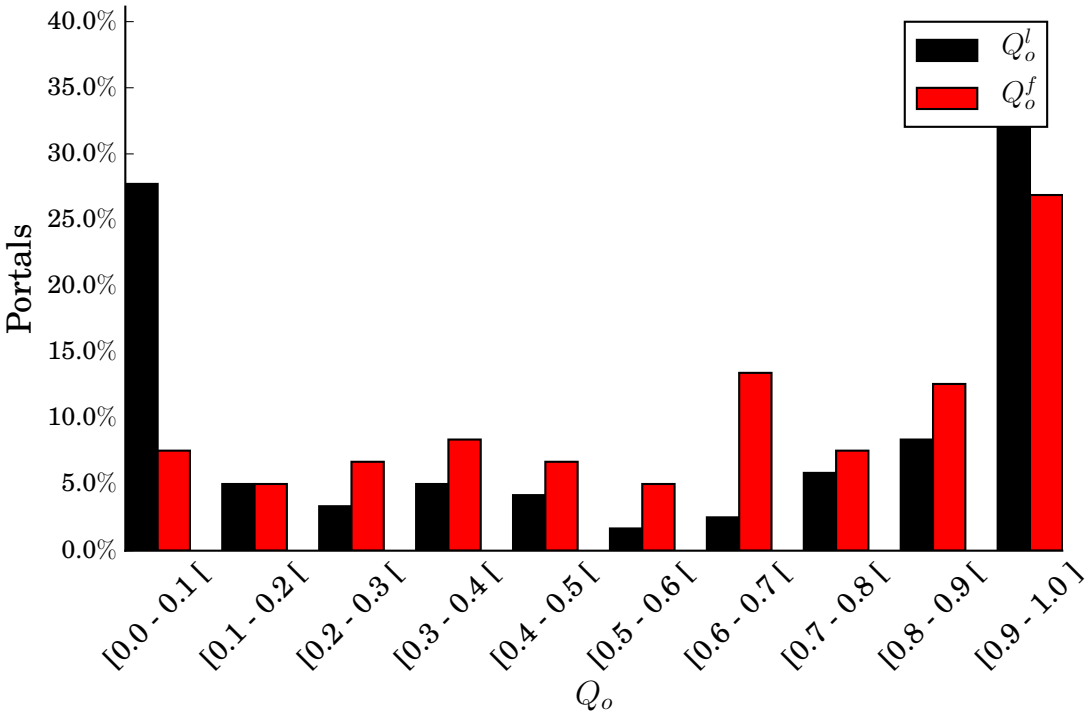

Figure 6.7: Distribution of Q_o metrics.

Contactability (Q_i)

Next, we report on our findings regarding the contactability information provided by the datasets, plotted in Figure 6.8. Firstly, considering the availability of any information for contacts, we can see that around 30% of the portals have an average Q_i value of over 0.9 and 25% of the portals a respective value of less than 0.1.

Regarding the contactability by email, we discovered that a subset consisting of 25% of the portals have an average Q_i^e value of over 0.9 and again about 25% of the portals do not really contain any email information (average Q_i^e value of < 0.1). The remaining 50% of the portals are more or less equally spread over the range of 0.1 – 0.9.

Regarding the appearance of URLs for either the author or maintainer contact values, we observed an average URL contactability over almost all portals of less than 0.1 (with one single portal in the range 0.5 – 0.6, namely `data.overheid.nl`), meaning that there are basically no URLs provided for contacting the publisher or maintainer of a dataset.

Overall, we can conclude that the majority of the portals have a low contactability value which bears the risk that data consumers stop using dataset if they cannot contact the maintainer or author (e.g., regarding the re-use if the licence is not clearly specified or in case of any data related issue).

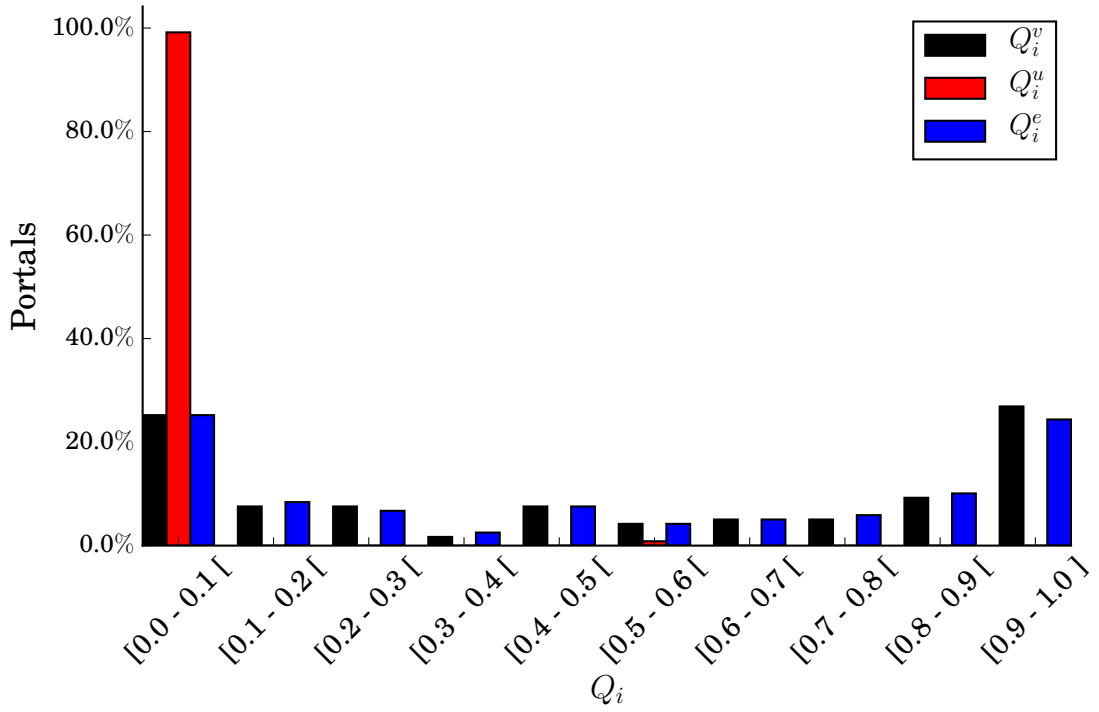

Figure 6.8: Distribution of Q_i metrics.

Accuracy (Q_a)

Our current accuracy analysis is based on header information from the available resource URLs. We performed in total 1.64M HTTP HEAD lookups, of which 1.55M successfully returned HTTP response headers and 1.4M contained the *content-type* field and 1.1M a *content-length* field. Considering datasets for which we have meta data values available and resources with a HTTP GET response header, we compute the format accuracy $Q_a(format, .)$ for 656k datasets over 115 portals, the mime-type accuracy for 252k datasets over 56 portals and the size accuracy for only 27k datasets over 60 portals.

Figure 6.9 shows the Q_a distribution of the average accuracy per portal. We can see that there exists a subset of 35% of the portals for which the meta data description about the mime-type type is not accurate with the header content type information of the resource, if available (see $Q_a(mime\_type)$). Regarding the file format, we observe that the provided formats information in general do not match with the derived file format from either the file extension or the header. One reason might be that there are over 1500 different variations of format descriptions in the datasets which can cause many incorrect format matches.

Overall, we derive two main findings. Firstly, the results reflect only a subset of the datasets and resources since we rely only on header information and not on the actual file content. However, the results show that those header information are not very precise and that there exists a mismatch between the provided meta data and the header

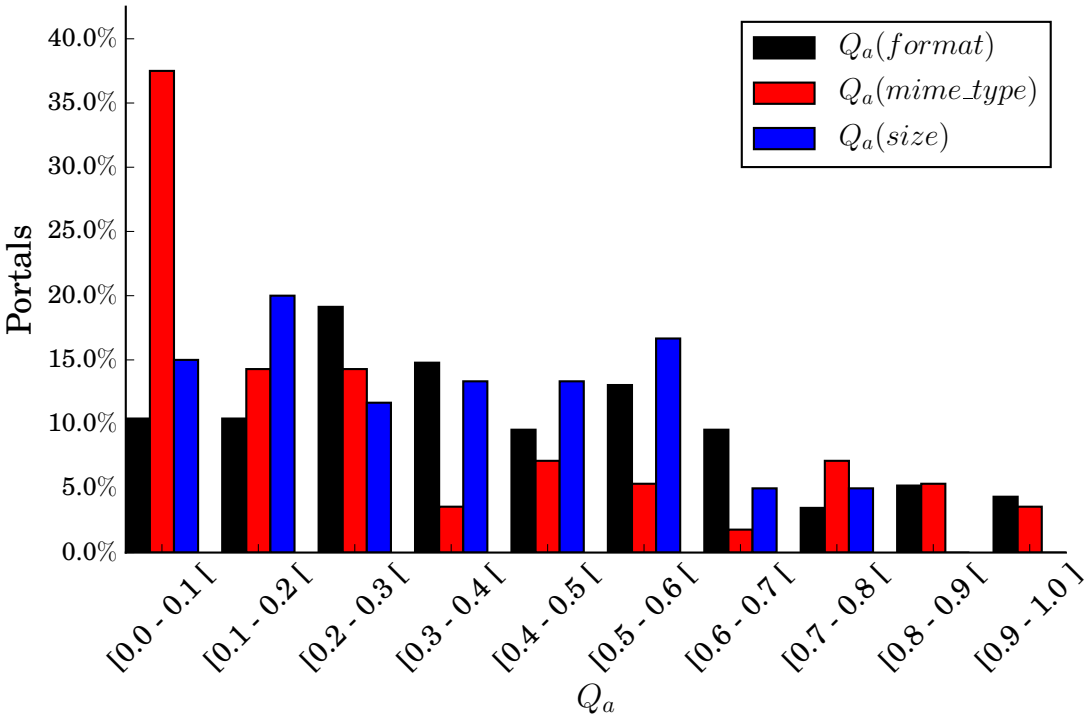

Figure 6.9: Accuracy distribution of the keys *mime_type*, *format* and *size*.

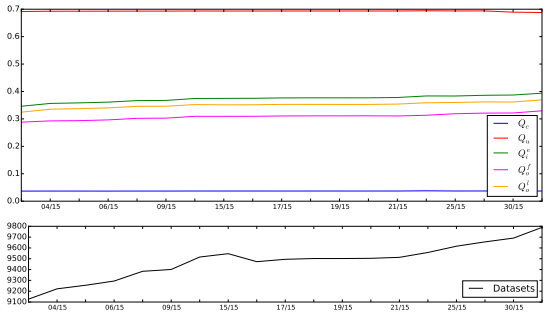

Figure 6.10: Evolution on `datahub.io`.

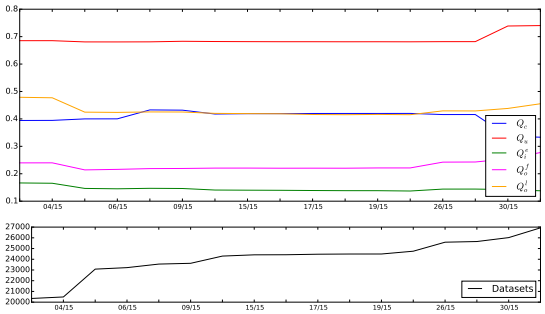

Figure 6.11: Evolution on `data.gov.uk`.

information of the resources. Secondly, to provide highly accurate measures we need to download and monitor the actual content of the resources and also need to improve the handling of the various format descriptions .

Evolution

Eventually, we report evolutionary patterns, using two large portals as an example, namely `data.gov.uk` and `datahub.io`. `Data.gov.uk` is the UK's open government data portal and is the sixth-largest portal in our portal list, consisting of about 27k

datasets. Datahub, on the other hand, is a non-governmental, community-run catalogue maintained by the Open Knowledge Foundation. It consists of about 10k datasets. The reason for selecting only these two portals is that over the last weeks we steadily increased the number of portals in our system and therefore we are not able to report an aggregated evolution of all portals currently in the system.

Figure 6.10 and Figure 6.11 display the evolution of the two selected portals, starting in the fourth week of 2015. The lower plot in the figure shows a mainly continuous increase of datasets on both portals. In the upper line plots we report the evolution of the completeness, accuracy, email-contactability, and format- and licence-openness metrics (see section 4.1).

If we want to conclude a trend on `datahub.io` from this plot, one can say that in general the portal uses more open licences (yellow line), more open formats (pink line) and also shows an increase of contactability information in form of emails (green line). The very low completeness and the very high usage values seem to remain stable over time. Regarding the quality metrics evolution on `data.gov.uk` we observe a rather stable, but also rather poor contactability value as well as a low but slightly increasing format-openness value. Concerning the completeness and usage of metadata keys we can see that a recent addition of datasets has led to a noticeable increase of the usage and a drop of the completeness value on the portal.

6.3 Socrata

Our system monitors at the moment 102 active Socrata portals, consisting of 74429 datasets. Table 6.9 shows the distribution of the portals based on their number of datasets. Interestingly, only one of the portals has more then 20k datasets: `opendata.socrata.com`

Table 6.9: Distribution of number of datasets over all Socrata portals.

	<50	$<10^2$	$<5\times10^2$	$<10^3$	$<5\times10^3$	$<10^4$	$>10^4$
$\lvert\mathcal{D}\rvert$	27	14	32	16	11	1	1

consists of 23k datasets. Further, two thirds of the Socrata portals hold less than 500 datasets.

6.4 OpenDataSoft

Regarding OpenDataSoft portals, we currently monitor 11 portals consisting in total of 1619 datasets. The largest portal, `data.iledefrance.fr`, holds 561 datasets (see Table 6.10).

Table 6.10: Distribution of number of datasets over all OpenDataSoft portals.

	<20	<50	$<10^2$	$<5\times10^2$	$>5\times10^2$
$\lvert\mathcal{D}\rvert$	3	2	1	4	1

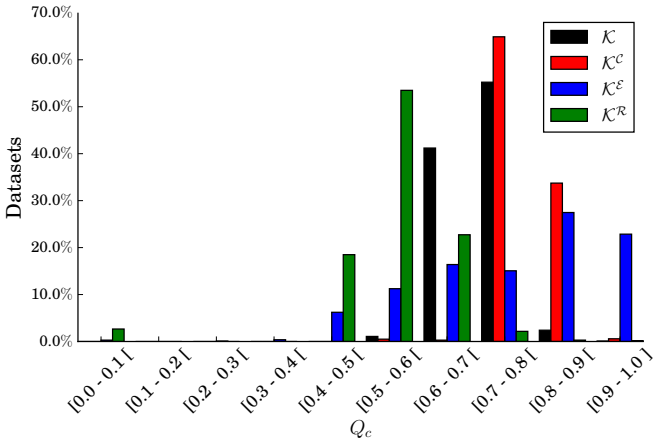

Figure 6.12: Completeness distribution over all datasets on Austrian Portals.

license_id	№	%
cc-by	2024	84.4
CC-BY-3.0	247	10.3
cc-by-sa	109	4.5
others	18	0.8

format	№	%
CSV	2638	32.3
png	541	6.6
TXT	516	6.3
others	4462	54.7

Table 6.11: Top-3 formats and licences on Austrian Portals.

6.5 Austrian Data Catalogs

Currently our system holds 7 Austrian data catalogs. Overall, these portals comprise 2410 datasets, of which 1705 are located on `data.gv.at`. Within this 7 portals we observe 5 different unique licence IDs, 100 different file formats and 2952 different tags. Regarding the licence IDs we identified a high homogeneity: 84% of the datasets use the `cc-by` identifier and in particular it is remarkable that in none of the datasets the licence ID field is empty. In total there are over 8k resource URLs on the portals, of which ∼32% are published as CSV (Table 6.11).

Figure 6.12 displays the aggregated completeness distribution over all 2.4k datasets hosted on the Austrian data portals. It is noticeable that all datasets have a core-key completeness between 0.7 and 0.9, and that there is a reasonable amount of datasets with an extra-key completeness between 0.4 and 0.8. This indicates that there is a possible set of extra keys which is available but not completed.

Regarding the overlap of resources, we can observe that in fact `data.gv.at` serves as a harvesting portal for the provincial level portals in our system. In the graph in Figure 6.13 we can see that `data.gv.at` covers almost all resources of `data.graz.gv.at` and `data.ktn.gv.at` while the non-governmental portals are rather isolated. (The numbers within the nodes indicates the number of total resources on a portal.)

Figure 6.14 display the evolution of the CKAN quality metrics and the number of datasets on `data.gv.at`, ranging from the fourth week to the 33rd week of 2015. In

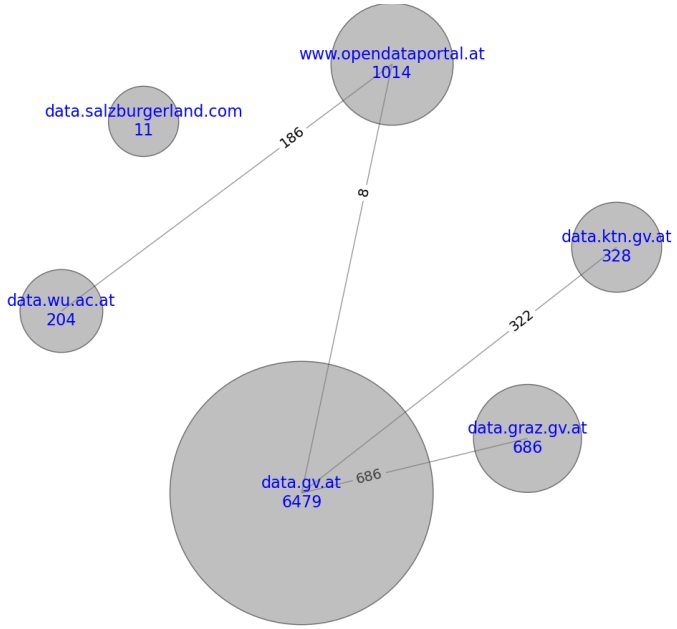

Figure 6.13: Graph of overlapping Resources on Austrian Portals.

this period the number of datasets has been growing for about 15%.

6.5.1 OGD Metadata Conformance

In 2011 the Austrian Federal Chancellery together with the cities of Vienna, Linz, Salzburg and Graz founded the "Cooperation Open Government Data Austria". In the course of this cooperation the stakeholders agreed on a common metadata structure "OGD Metadata" [Aus14], currently on version 2.3 (February 2015). This standard consists of 12 core fields, which are considered as mandatory, and 21 additional optional metadata fields. The specified fields are based on the CKAN metadata structure.

data.gv.at In the following we investigate the usage and completeness of the OGD metadata fields on `data.gv.at`, Austria's governmental data portal.

Regarding the mandatory OGD metadata fields, we observed a completeness value of over 0.8 for 11 and a usage value of over 0.8 for 10 out of the 12 mandatory fields. In detail, we observed very low usage and completeness for the OGD key *license* (< 0.01). A possible explanation for the absence of this key is that CKAN by default provides three other keys for describing a licence (*license_id*, *license_title*, *license_url*), with high completeness and usage values in the Austrian portal.

Regarding the optional OGD metadata fields, we can see a slightly different observation. The completeness of 8 and the usage of 3 out of the 21 optional keys is less than 50%. `data.gv.at` uses a CKAN extension which provides by default a range of extra keys, including the optional OGD fields. This explains the high usage values of the optional

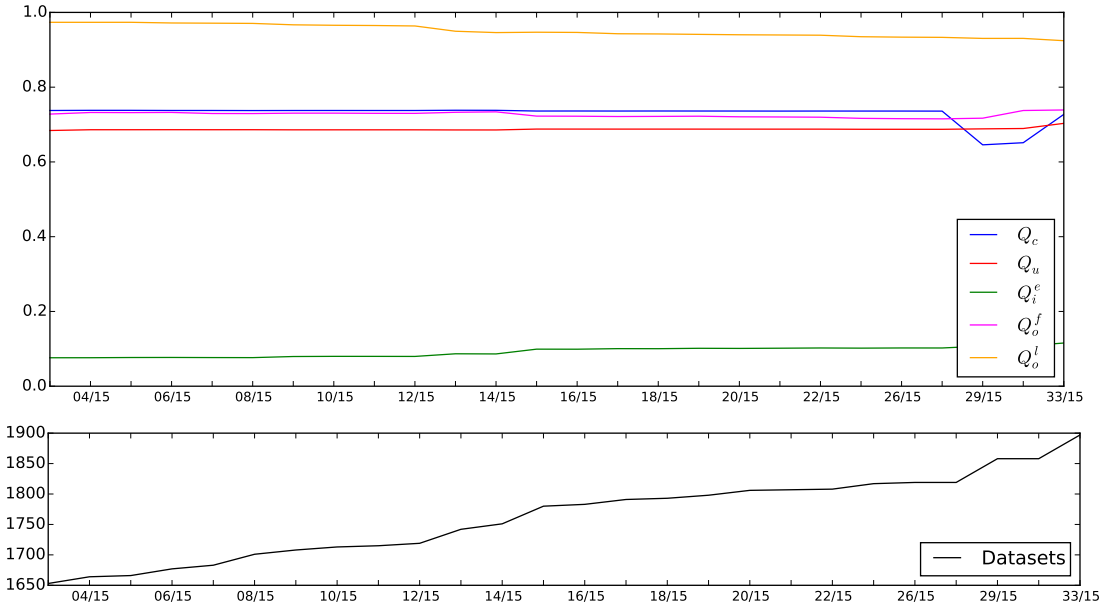

Figure 6.14: Quality Metrics Evolution on `data.gv.at`.

keys. Interestingly, the optional fields *size*, *language* and *characterset* for describing a resource are hardly ever used or completed (i.e., usage or completeness less than 0.3).

Further, we observed a set of 40 extra keys on `data.gv.at` not occurring in the optional or mandatory OGD metadata fields with a high average completeness of about 0.7 and an average usage of 0.5.

CHAPTER 7

Summary & Conclusion

"The world is full of obvious things which nobody by any chance ever observes."
— Sir Arthur Conan Doyle, *The Hound of the Baskervilles*

The Open Data movement enjoys great popularity and enthusiasm mostly driven by public administration but also increasingly gaining attention the private sector. However, there are various issues that could disrupt the success of the Open Data project. Firstly, there is an issue of poor metadata quality in data portals: inadequate descriptions or classifications of datasets directly affect the usability and searchability of resources. Secondly, there is an issue of high heterogeneity of formats, licenses and taxonomies in Open Data portals. Moreover, there is an issue of heterogeneous metadata schemas, which affects the comparability, discoverability and interoperability of datasets across portals.

While first projects emerge to quantify and qualify the quality of Open Data, there exists no comprehensive quality assessment and evolution tracking targeting the mentioned issues. We have contributed to this efforts by (i) selecting and formally defining a set of objective quality metrics suitable to assess the quality of Open Data portals in an automated way, (ii) proposing a mapping of metadata keys found on different publishing software frameworks partially based on existing work [ATS15] and (iii) developing an Open Data portal monitoring and quality assessment framework, the "*Open Data Portal Watch*" platform.[1]

In order to assess the quality of Open Data portals in an objective and automated way, we have formally defined a set of quality metrics, namely *retrievability*, *usage*, *completeness*, *accuracy*, *openness* and *contactability*. We implemented the assessment of these metrics in our monitoring framework which consists of three main components:

[1] `http://data.wu.ac.at/portalwatch/`

A **fetching component** periodically retrieves dataset information of listed portals and stores the respective metadata in a document store.

The stored information is analysed by a **quality assessment component** which computes our defined quality metrics for the various datasets and portals.

A publicly available **dashboard component** displays up-to-date quality metrics for each portal using various views and charts and allows to browse in the history of the portals based on weekly snapshots.

Currently, our assessment framework monitors 126 CKAN, 102 Socrata and 11 OpenDataSoft portals. More in-depth for CKAN, all CKAN portals together consist of a total of 745k datasets and 3.12M resources. Our core findings over this set of CKAN portals can be summarized as follows:

- We found 1.91M unique and valid resource URLs. The most common file format is currently `PDF` (25.6% of the resources), followed by `HTML`, `other`[2] and `CSV` (5.8%). Looking into detail, we observed that about 80% of the PDF files can be found on in the largest 10% of all portals while 90% of the CSV files are distributed over the smallest 30% of the portals. This indicates that there are large portals in our system holding the majority of the non-machine-readable PDF files.

- We observed a strong heterogeneity across portals with respect to format descriptions, extra meta data keys and tags which cause serious challenges for a global integration of the portals. We found 218 different unique licence IDs and 1516 file format descriptions over the set of CKAN portals, indicating a very high heterogeneity caused by non-standardized metadata specifications. For instance, by looking into the 500 most common format descriptions we were able to match the term "csv" 21 times (e.g., "csv", "CSV" and ".csv").

- ∼40% of the portals provide confirmed open licenses, but there is also a group of around 27% of the portals for which we could only confirm an average license openness per dataset of less than 0.1. Similarly, ∼30% of the portals contain mainly resources in a confirmed open format.

  In detail, only three out of the 10 most-used file formats are open formats and only one format (CSV) out of these 10 can be considered as machine-readable. We conclude that there is a gap between the common definition of Open Data (i.e., the Open Definition[3]) and these observed results.

- The majority of the datasets do not provide contact information in form of email addresses or URLs. This missing provenance information involves the risk of intransparency and impaired usability of datasets.

[2]Please note, that "other" is used as the format description.
[3]`http://opendefinition.org/`, last accessed 2015-09-28

7.1 Further Work

On our agenda in order to further improve our framework and quality metrics, we prioritize the evaluation of more portals (especially including non-CKAN portals), the scalable monitoring of the actual resource content for a better accuracy metric (e.g., expanding on change frequency aspects) and further refinement of the openness metric regarding the various licences and formats.

Quality metrics based on DCAT mapping. To achieve a higher degree of comparability across the various data publishing catalogs we aim to adapt and expand our set of quality metrics. Currently the metrics are partially tailored to the CKAN software (e.g., the openness of formats or the retrievability of resources).

Improving and extending DCAT mapping. We will further research solutions to deal with the high heterogeneity of metadata keys and values in and across portals. A first approach would the extension of the static mapping by the most commonly used metadata keys. For instance, by mapping a set of conceptually similar CKAN *extra* keys to a corresponding DCAT property.

RDFizing & Interlinking data sources. By mapping various metadata schemas to DCAT (and therefore enabling RDF query functionality across datasets) we are able to interlink data from different portals.

Similarly, one can think of interlinking the published resources itself across different data portals. A large amount of open data sources are information published from relational databases or MS Excel sheets and are commonly represented in so called character-separate-value formats (e.g., CSV, TSV, etc.). In order to enable query functionality over the resources of the portals one has to transform this CSV files to RDF and interlink it with existing knowledge graphs (e.g., LODPilot Core, WikiData or DBPedia).

(Semi-)Automatic Quality Improvement. An interesting application of the acquired information in our system would be the automated addition and correction of respectively missing and wrong metadata. An approach would be to integrate an automated quality improvement in an existing data publishing framework, e.g., in form of a CKAN plugin. A respective plugin for CKAN could provide the functionality to automatically suggest values for certain missing metadata fields, which could be computed by respective heuristics.

Another feature of such a plugin could be an automated consistency checking of existing fields in comparison to actual values computed by such heuristics. Such a tool could detect mistakes like mismatches between file types specified in metadata and actual resources, resource size, encodings etc.

A complementary approach towards quality improvement on open data portals would be the integration of community-driven methods, i.e., using crowd-sourcing to involve data producers and consumers in the improvement process.

Bibliography

[ATS15] Ahmad Assaf, Raphaël Troncy, and Aline Senart. HDL - Towards a harmonized dataset model for open data portals. In *PROFILES 2015, 2nd International Workshop on Dataset Profiling & Federated Search for Linked Data, ESWC15*, Portoroz, Slovenia, May 2015.

[Aus14] Cooperation Open Government Data Austria. OGD Metadaten 2.3. `http://reference.e-government.gv.at/fileadmin/_migrated/content_uploads/OGD-Metadaten_2_3_2015_02_19_EN.pdf`, 2014.

[BBL08] David Beckett and Tim Berners-Lee. Turtle - Terse RDF Triple Language. `http://www.w3.org/TeamSubmission/turtle/`, January 2008.

[BCFM09] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. Methodologies for Data Quality Assessment and Improvement. *ACM Comput. Surv.*, 41(3):16:1–16:52, July 2009.

[BETL12] Katrin Braunschweig, Julian Eberius, Maik Thiele, and Wolfgang Lehner. The State of Open Data - Limits of Current Open Data Platforms. In *Proceedings of the International World Wide Web Conference, WWW 2012, Lyon, France*, 2012.

[BL06] Tim Berners-Lee. Linked data, 2006. `http://www.w3.org/DesignIssues/LinkedData.html`, 2006.

[Bla14] Mattias Blaim. Quality and Compatibility of License Information in Open Data portals, 2014. Bachelor Thesis at Vienna University of Economics and Business.

[BMS11] John Carlo Bertot, Patrice McDermott, and Ted Smith. Measurement of open government: Metrics and process. In *Proceedings of the Annual Hawaii International Conference on System Sciences*, pages 2491–2499, 2011.

[Bra14] T. Bray. The JavaScript Object Notation (JSON) Data Interchange Format. Internet Engineering Task Force (IETF) RFC 7159, March 2014.

[Com09] Creative Commons. CC0 1.0 Universal, 2009. https://creativecommons.org/publicdomain/zero/1.0/legalcode.

[DFC+01] Erik Duval, Eddy Forte, Kris Cardinaels, Bart Verhoeven, Rafael Van Durm, Koen Hendrikx, Maria Wentland Forte, Norbert Ebel, Maciej Macowicz, Ken Warkentyne, et al. The Ariadne knowledge pool system. *Communications of the ACM*, 44(5):72–78, 2001.

[DHSW02] Erik Duval, Wayne Hodgins, Stuart Sutton, and Stuart L Weibel. Metadata principles and practicalities. *D-lib Magazine*, 8(4):16, 2002.

[GMe13] Paul Groth and Luc Moreau (eds.). PROV-Overview. An Overview of the PROV Family of Documents. http://www.w3.org/TR/prov-overview/, April 2013.

[HK08] Mark Henley and Richard Kemp. Open Source Software: An introduction. *Computer Law & Security Review*, 24(1):77 – 85, 2008.

[Int92] International Organizational for Standardization. ISO 9000: International standards for quality management, 1992.

[Int08] International Organizational for Standardization. ISO 32000-1:2008. Document management — Portable document format — Part 1: PDF 1.7, 2008.

[JCZ12] Marijn Janssen, Yannis Charalabidis, and Anneke Zuiderwijk. Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, 29(4):258–268, 2012.

[JV97] Matthias Jarke and Yannis Vassiliou. Data warehouse quality: A review of the DWQ project. In *Second Conference on Information Quality (IQ 1997)*, pages 299–313, 1997.

[KAU+13] Tobias Käfer, Ahmed Abdelrahman, Jürgen Umbrich, Patrick O'Byrne, and Aidan Hogan. Observing linked data dynamics. In *The Semantic Web: Semantics and Big Data: 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings*, volume 7882, page 213. Springer, 2013.

[KC] Graham Klyne and Jeremy J. Carroll. Resource Description Framework (RDF): Concepts and Abstract Syntax. Technical report.

[KCN13] Jan Kučera, Dušan Chlapek, and Martin Nečaský. Open Government Data Catalogs: Current Approaches and Quality Perspective. In *EGOVIS/EDEM*, pages 152–166. Springer-Verlag Berlin Heidelberg, 2013.

[Lau08] Philippe Laurent. The GPLv3 and Compatibility Issues. European Open source Lawyers Event 2008, September 2008.

[LN14] Juho Lindman and Linus Nyman. The Businesses of Open Data and Open Source: Some Key Similarities and Differences. *Technology Innovation Management Review*, (January):12–17, 2014.

[ME14] Fadi Maali and John Erickson. Data Catalog Vocabulary (DCAT). `http://www.w3.org/TR/vocab-dcat/`, January 2014.

[MG13] Luc Moreau and Paul Groth. Provenance: an Introduction to PROV. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 3(4):1–129, 2013.

[Mol11] Jennifer C. Molloy. The Open Knowledge Foundation: Open Data Means Better Science. *PLoS Biol*, 9(12):e1001195, 12 2011.

[MR08] Peter Murray-Rust. Open Data in Science. *Serials Review*, 34(1):52–64, 2008.

[NH02] Marc Najork and Allan Heydon. High-performance web crawling. In *Handbook of Massive Data Sets*, volume 4 of *Massive Computing*, pages 25–45. Springer US, 2002.

[Nor07] Ray P Norris. How to Make the Dream Come True: The Astronomers' Data Manifesto. *Data Science Journal*, 6:S116—-S124, 2007.

[NP12] Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

[OD09] Xavier Ochoa and Erik Duval. Automatic evaluation of metadata quality in digital repositories. *International Journal on Digital Libraries*, 10(2):67–91, 2009.

[O'H11] Kieron O'Hara. Transparent government, not transparent citizens: a report on privacy and transparency for the cabinet office. Technical report, September 2011.

[Ope12] Open Knowledge Foundation. The Open Data Handbook. `http://opendatahandbook.org/`, 2012.

[Ors09] Peter Orszag. Open Government Directive. `https://www.whitehouse.gov/open/documents/open-government-directive`, 2009. Memorandum for the Heads of Executive Departments and Agencies.

[PLW02] Leo L Pipino, Yang W Lee, and Richard Y Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002.

[RHS14] Konrad Johannes Reiche, Edzard Höfig, and Ina Schieferdecker. Assessment and Visualization of Metadata Quality for Open Government Data. In *International Conference for E-Democracy and Open Government*, 2014.

[SB+08] Alma Swan, Sheridan Brown, et al. *To share or not to share: Publication and quality assurance of research data outputs: Main report.* Research Information Network, 2008.

[Sha05] Yakov Shafranovich. Common format and mime type for comma-separated values (csv) files. Internet RFC 4180, October 2005.

[SKL14] Manu Sporny, Gregg Kellogg, and Markus Lanthaler. JSON-LD 1.0A JSON-based Serialization for Linked Data. `http://www.w3.org/TR/json-ld/`, January 2014.

[SLW97] Diane M Strong, Yang W Lee, and Richard Y Wang. Data quality in context. *Communications of the ACM*, 40(5):103–110, 1997.

[The12] The Unicode Consortium, editor. *The Unicode Standard, Version 6.1 — Core Specification.* The Unicode Consortium, 2012.

[UMP15] Jürgen Umbrich, Nina Mrzelj, and Axel Polleres. Towards capturing and preserving changes on the web of data. In *Proceedings of the First DIACHRON Workshop on Managing the Evolution and Preservation of the Data Web co-located with 12th European Semantic Web Conference (ESWC 2015), Portorož, Slovenia, May 31, 2015.*, pages 50–65, 2015.

[US 15] US Office of Management and Budget. Overview of the Federal Performance Framework. Annual Performance Plans, and Annual Program Performance Reports. `http://www.whitehouse.gov/sites/default/files/omb/assets/a11_current_year/s200.pdf`, 2015.

[vLG10] Jörn von Lucke and Christian Geiger. Open Government Data – Frei verfügbare Daten des öffentlichen Sektors. *Gutachten, Deutsche Telekom Institute for Connected Cities, Zeppelin University, Friedrichshafen*, 2010.

[Win13] Joss Winn. Open data and the Academy: An Evaluation of CKAN for Research Data Management. In *In: IASSIST 2013*, Cologne, 2013.

[WKLW98] Stuart Weibel, John Kunze, Carl Lagoze, and Misha Wolf. Dublin core metadata for resource discovery. Technical report, 1998.

[Wor07] World Wide Web Consortium. Definition of Open Standards. `http://www.w3.org/2005/09/dd-osd.html`, September 2007.

[Wor15] World Wide Web Foundation. Open Data Barometer, January 2015.

[ZJC+12] Anneke Zuiderwijk, Marijn Janssen, Sunil Choenni, Ronald Meijer, and Roexsana Sheikh Alibaks. Socio-technical Impediments of Open Data. *Electronic Journal of e-Government*, 10(2):156–172, 2012.

[ZMLW14] Hongwei Zhu, Stuart E Madnick, Yang W Lee, and Richard Y Wang. Data and Information Quality Research: Its Evolution and Future. In *Computing Handbook, 3rd ed. (2)*, pages 16: 1–20. 2014.

[ZRM+14] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality Assessment Methodologies for Linked Open Data (Under Review). *Semantic Web Journal*, 2014. This article is still under review.

Glossary

Accuracy The extent to which certain resource metadata accurately describe the resource. 44

API In the context of data portals, an application programming interface (API) is usually a set of HTTP requests allowing to programatically access, edit, add and delete resources on the portal. The (possible) response messages are in semi-structured formats, e.g. JSON or XML. 14

attribution *Attribution* licenses state that re-users must give attribution to the source of the content. 22

CKAN is an open-source data portal platform developed by the Open Knowledge Foundation. The software is increasingly popular among cities, governments and private data provider worldwide. 15

Completeness The extent to which the used metadata keys contain information. 43

Contactability The extent to which the data publisher provide a contact information. 48

Creative Commons Creative Commons is an US nonprofit-organisation developing copyright-licenses known as Creative Commons licenses. Theses licenses are free of charge to the public. The organisation aims to provide an easy-to-use way of making work available for others to share and to build upon legally. 22, 89

DCAT DCAT is a W3C metadata recommendation for publishing data on the Web defined in RDF. 33, 89

Machine Readable Formats A machine-readable format is a format that can be processed in an automated and structured way. 16

Open Data Catalog A open data catalog is a single point of access to distribute data. 14

Open Data Portal A open data portal is a catalog which allows user to automatically retrieve and publish open data. 14

Open Format An open format is a format with a freely available specification and without restrictions upon its use. 17

OpenDataSoft The France-based company OpenDataSoft distributes a commercial data portal. Its few instances can be found mainly in France.. 16

Openness The extent to which the license and available formats are suitable to classify a dataset as open. 47

public domain A dataset is said to be *in the public domain* if it is not owned by a particular company or person and is free to use for anyone. 22, 23

Retrievability The extent to which dataset information in portal and the listed resources can be retrieved by a human or software agent. 42

share-alike *Share-alike* is the requirement to publish any derived content under the same license. 22

Socrata Socrata is a company offering a range of database and discovery services for governmental data. In the context of this thesis *Socrata* referrers to the data catalog product. 15

Usage The extent to which metadata keys are used to describe a dataset. 43

Acronyms

CC Creative Commons. 22

CKAN Comprehensive Knowledge Archive Network. 15

CSV comma-separated values. 18

DCAT Data Catalog Vocabulary. 33

HTTP Hypertext Transfer Protocol. 42

JSON JavaScript Object Notation. 19

LD Linked Data. 13

LOD Linked Open Data. 13

LOGD Linked Open Government Data. 13

OGD Open Government Data. 11

PDF Portable Document Format. 18

RDF Resource Description Framework. 13, 19

URI Uniform Resource Identifier. 13, 19

URL Uniform Resource Locator. 19