Linked Data Quality Assessment and its Application to Societal Progress Measurement

Der Fakultät für Mathematik und Informatik
der Universität Leipzig
eingereichte

DISSERTATION

zur Erlangung des akademischen Grades

DOKTOR-INGENIEUR
(Dr. Ing.)

im Fachgebiet Informatik

vorgelegt

von **M.Sc. Amrapali Zaveri**

geboren am 09. November 1984 in Mumbai, Indien

Leipzig, den 17.04.2015

Bibliographic Data

Title: Linked Data Quality Assessment and its Application to Societal Progress Measurement
Author: Amrapali Zaveri
Institution: Universität Leipzig, Fakultät für Mathematik und Informatik
Statistical Information: 149 pages, 22 figures, 23 tables, 149 literature references

Abstract

In recent years, the Linked Data (LD) paradigm has emerged as a simple mechanism for employing the Web as a medium for data and knowledge integration where both documents and data are linked. Moreover, the semantics and structure of the underlying data are kept intact, making this the Semantic Web. LD essentially entails a set of best practices for publishing and connecting structure data on the Web, which allows publishing and exchanging information in an interoperable and reusable fashion. Many different communities on the Internet such as geographic, media, life sciences and government have already adopted these LD principles. This is confirmed by the dramatically growing Linked Data Web, where currently more than 50 billion facts are represented.

With the emergence of Web of Linked Data, there are several use cases, which are possible due to the rich and disparate data integrated into one global information space. Linked Data, in these cases, not only assists in building mashups by interlinking heterogeneous and dispersed data from multiple sources but also empowers the uncovering of meaningful and impactful relationships. These discoveries have paved the way for scientists to explore the existing data and uncover meaningful outcomes that they might not have been aware of previously.

In all these use cases utilizing LD, one crippling problem is the underlying *data quality*. Incomplete, inconsistent or inaccurate data affects the end results gravely, thus making them unreliable. Data quality is commonly conceived as *fitness for use*, be it for a certain application or use case. There are cases when datasets that contain quality problems, are useful for certain applications, thus depending on the use case at hand. Thus, LD consumption has to deal with the problem of getting the data into a state in which it can be exploited for real use cases. The insufficient data quality can be caused either by the LD publication process or is intrinsic to the data source itself.

A key challenge is to assess the quality of datasets published on the Web and make this quality information explicit. Assessing data quality is particularly a challenge in LD as the underlying data stems from a set of multiple, autonomous and evolving data sources. Moreover, the dynamic nature of LD makes assessing the quality crucial to

measure the accuracy of representing the real-world data. On the document Web, data quality can only be indirectly or vaguely defined, but there is a requirement for more concrete and measurable data quality metrics for LD. Such data quality metrics include correctness of facts wrt. the real-world, adequacy of semantic representation, quality of interlinks, interoperability, timeliness or consistency with regard to implicit information. Even though data quality is an important concept in LD, there are few methodologies proposed to assess the quality of these datasets.

Thus, in this thesis, we first unify 18 data quality dimensions and provide a total of 69 metrics for assessment of LD. The first methodology includes the employment of LD experts for the assessment. This assessment is performed with the help of the TripleCheckMate tool, which was developed specifically to assist LD experts for assessing the quality of a dataset, in this case DBpedia. The second methodology is a *semi-automatic* process, in which the first phase involves the detection of common quality problems by the automatic creation of an extended schema for DBpedia. The second phase involves the manual verification of the generated schema axioms. Thereafter, we employ the wisdom of the crowds i.e. workers for online crowdsourcing platforms such as Amazon Mechanical Turk (MTurk) to assess the quality of DBpedia. We then compare the two approaches (previous assessment by LD experts and assessment by MTurk workers in this study) in order to measure the feasibility of each type of the user-driven data quality assessment methodology.

Additionally, we evaluate another semi-automated methodology for LD quality assessment, which also involves human judgement. In this semi-automated methodology, selected metrics are formally defined and implemented as part of a tool, namely R2RLint. The user is not only provided the results of the assessment but also specific entities that cause the errors, which help users understand the quality issues and thus can fix them. Finally, we take into account a domain-specific use case that consumes LD and leverages on data quality. In particular, we identify four LD sources, assess their quality using the R2RLint tool and then utilize them in building the Health Economic Research (HER) Observatory. We show the advantages of this semi-automated assessment over the other types of quality assessment methodologies discussed earlier. The Observatory aims at evaluating the impact of research development on the economic and healthcare performance of each country per year. We illustrate the usefulness of LD in this use case and the importance of quality assessment for any data analysis.

Publications

This thesis is based on the following conference and journal publications, in which I have either been an author or a contributor. *At the respective chapter and section*, I have included the references to the appropriate publications. The full list of publications can be found in Appendix A.

Conference publications, peer-reviewed

- *Using Linked Data to evaluate the impact of Research and Development in Europe: a Structural Equation Model*, In Proceedings of 12th International Semantic Web Conference, 21-25 October 2013, Sydney, Australia (ISWC 2013) [Zaveri et al., 2013d]
- *User-driven Quality Evaluation of DBpedia*, In Proceedings of 9th International Conference on Semantic Systems, I-SEMANTICS '13, Graz, Austria, September 4-6, (I-Semantics 2013) [Zaveri et al., 2013a]
- *Crowdsourcing Linked Data quality assessment*, In Proceedings of 12th International Semantic Web Conference, 21-25 October 2013, Sydney, Australia (ISWC 2013) [Acosta et al., 2013]

Journal publications, peer-reviewed

- *Quality assessment methodologies for Linked Data: A Survey*, Semantic Web Journal (2015) [Zaveri et al., 2015]
- *Publishing and Interlinking the Global Health Observatory Dataset*, Semantic Web Journal (2013) [Zaveri et al., 2013b]

Journal publications, under review

- *Using Linked Data to build an Observatory of Societal Progress Indicators*, Journal of Web Semantics (2014) [Zaveri et al., 2014b]
- *Publishing and Interlinking the USPTO Patent Data*, Semantic Web Journal (2014) [Zaveri et al., 2014a]

Acknowledgments

"Der Weg ist das Ziel" - Confucius. This proverb translates to "The journey is the destination" and very aptly fits the journey of a PhD student. Thus, I would like to thank all the people who were part of my journey. First of all, I would like to thank my supervisors: Dr. Klaus-Peter Fähnrich and Dr. Sören Auer for giving me the opportunity to pursue PhD. I would like to specially thank Sören for his excellent guidance, stimulating discussions and keen observations, which helped me improve over time. Moreover, he taught me how to balance work and family efficiently. I am also indebted to my supervisor Dr. Jens Lehmann for his constant support, motivation and confidence in me that enabled me to progress in my work with each passing day. Most importantly, I would like to thank the German Academic Exchange Service (DAAD) for granting me the scholarship, without which pursuing a PhD would be more difficult.

I would also like to thank all my colleagues at AKSW for their help and engaging discussions. A big thank you to all my co-authors for believing in my ideas, seeing them through and for their work so that this thesis could be completed. I am grateful to my colleagues for their companionship that helped me pull through the several years away from home. In particular, I would like to thank Saeedeh Shekarpour, for always being there when I needed someone to talk to; Timofey Ermilov, for his jolly nature and everything-is-easy attitude; Nadine Jänicke for helping me fill out numerous German forms and translate contracts; Thomas Riechert for helping me always and making me feel as a part of his family; Konrad Höffner for being a good friend and for helping me for every little thing I asked for.

A special thanks also goes to Dr. Ricardo Pietrobon, who inspired me to pursue research and remotely guided me throughout the journey. Moreover, thanks to Anisa Rula, Maribel Acosta, Joao Ricardo Nickenig Vissoci and Cinzia Daraio, with whom working remotely, despite the time or location differences, was not only engaging but also enjoyable. I would like to thank my German language teacher Frau Zajonz for teaching me the German language well so that I could go through my daily activities effortlessly. I would also like to thank my gym instructor, Laura for keeping me fit throughout these years.

Most importantly, I would like to thank my parents, Jayshree and Jyotindra Zaveri, for their unconditional love. Also, I am grateful for the support from my family, Preksha, Soham, Aanya and Tanishi Chakravarti. A big thank you to my friends for their support, whose friendship I truly cherish, Siddha Joshi, Sunita Patil, Elena Ermilov and Annett Riechert.

Contents

1. Introduction

1.1. Linked Data and Data Quality on the Web

The World Wide Web (WWW), since its inception, has drastically altered the way we share knowledge by publishing documents as part of a global information space. This Web of Documents contains hypertext links that enables users to traverse this information using Web browsers. Despite the inarguable benefits that the Web provides, until recently the same principles that enabled the Web of Documents to expand have not been applied to data. Traditionally, data published on the Web is available in a variety of formats such as CSV or XML, or marked up as HTML tables, neglecting much of its structure and semantics. These diverse formats are not expressive enough to enable the linking of individual facts/entities in a particular document to be connected to related facts/entities in another document.

In recent years, the Linked Data (LD) paradigm [Berners-Lee, 2006] has emerged as a simple mechanism for employing the Web as a medium for data and knowledge integration where both documents and data are linked. Moreover, the semantics and structure of the underlying data are kept intact, making this the Semantic Web. LD essentially entails a set of best practices for publishing and connecting structure data on the Web, which allows publishing and exchanging information in an interoperable and reusable fashion. Many different communities on the Internet such as geographic, media, life sciences and government have already adopted these LD principles. This is confirmed by the dramatically growing Linked Data Web, where currently more than 50 billion facts are represented[1]. In particular, the amount of information in the life science domain, specifically on diseases and healthcare research, being published as Linked Data is constantly increasing. Figure 1.1 shows the part of the cloud of the Linked Data Web[2] covering the life science domain.

Earlier, in the Web of Documents, the distributed document collections and taxonomic indexing schemes hindered the ability of researchers to identify important connections that could yield new scientific insights [Boyce et al., 2014]. Now, with the Web of Linked Data, there are several use cases which are possible due to the rich and disparate data integrated into one global information space. Successful use cases of Linked Data have been in healthcare research area [Zaveri et al., 2011, Zaveri et al., 2013c], biomedical domain, e.g. for drug discovery [Williams et al., 2012, Jentzsch et al., 2009] or for detecting patterns in particular types of diseases [Zaveri et al., 2013b]. Linked Data, in these cases, not only assists in building mashups by interlinking hetero-

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

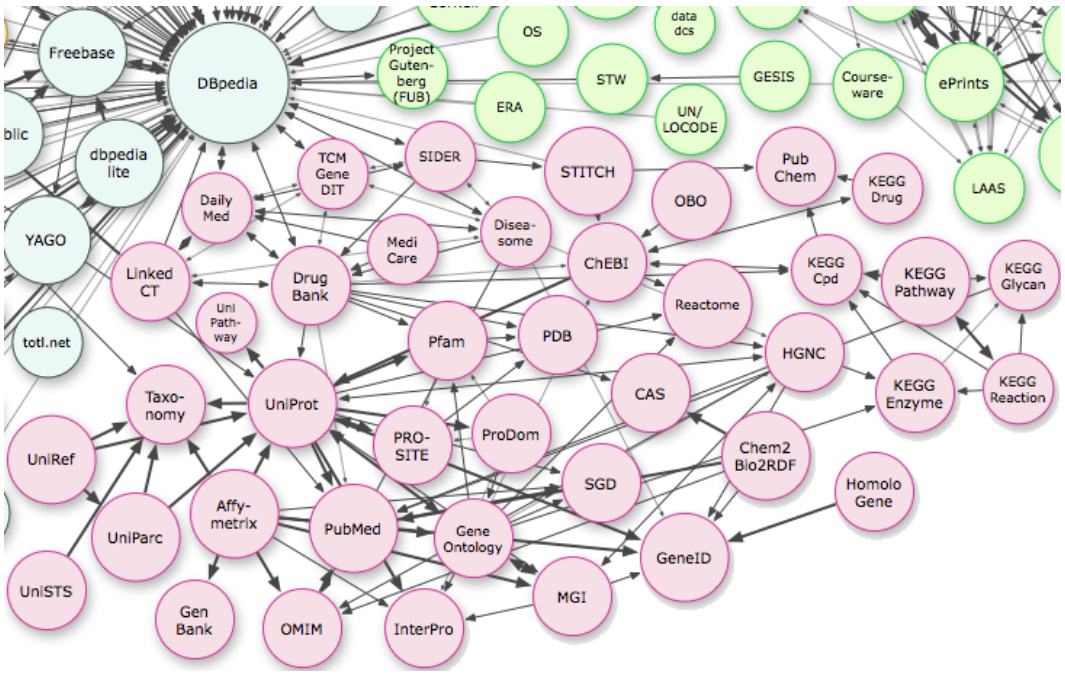


Figure 1.1.: The life science Linked Data Web.

geneous and dispersed data from multiple sources but also empowers the uncovering of meaningful and impactful relationships. These discoveries have paved the way for scientists to explore the existing data and uncover meaningful outcomes that they might not have been aware of previously.

In all these use cases utilizing LD, one crippling problem is the underlying *data quality*. Incomplete, inconsistent or inaccurate data affects the end results gravely, thus making them unreliable. In [Orr, 1998], data quality is "the measure of the agreement between the data views presented by an information system and that same data in the real world", however, it is commonly conceived as *fitness for use* [Juran, 1974, Wang and Strong, 1996] for a certain application or use case. There are cases when datasets, that contain quality problems, are useful for particular applications, thus depending on the use case at hand. Linked Data on the Web is either created from structured data sources (such as relational databases), from semi-structured sources (such as Wikipedia), or from unstructured sources (such as text). Thus, in the case of DBpedia [Lehmann et al., 2014, Morsey et al., 2012] (the LD version of Wikipedia), the quality is sufficient for providing facts about general information. However, when using this information to making important decisions, such as in case of a medical application, the quality is insufficient [Zaveri et al., 2013a]. The insufficient data quality can be caused either by the LD publication process or is intrinsic to the data source itself. Thus, LD consumption has to deal with the problem of getting the data into a state in which it can be exploited for real use cases.

A key challenge is to assess the quality of datasets published on the Web and make this quality information explicit. Assessing data quality is particularly a challenge in

LD as the underlying data stems from a set of multiple, autonomous and evolving data sources. Moreover, the dynamic nature of LD makes assessing the quality crucial to measure the accuracy of representing the real-world data. On the document Web, data quality can only be indirectly or vaguely defined, but there is a requirement for more concrete and measurable data quality metrics for LD. Such data quality metrics include correctness of facts wrt. the real-world, adequacy of semantic representation, quality of interlinks, interoperability, timeliness or consistency with regard to implicit information. Even though data quality is an important concept in LD, there are few methodologies proposed to assess the quality of these datasets. Thus, in this thesis, we investigate the following research areas:

- *different user-driven data quality assessment methodologies particularly for LD*
- *consumption of LD for a particular use case leveraging on data quality*

1.2. User Scenario

Ms. Sharma, a healthcare policy maker is interested in knowing which diseases represent the largest threat to the citizens of India and for which of these are the affordable and effective treatment options currently available. She is looking to improve health outcomes and lower the cost of the delivery of healthcare services for individuals with specific needs. Obtaining this information will help her in allocating funds appropriately to develop corresponding treatment options and to conduct clinical trials in India for that disease. Since this information (combining the threat information and the treatment effectiveness of diseases) is not explicitly present in any one source, Ms. Sharma needs to gather information from different sources.

First, she looks up the World Health Organization (WHO) website to ascertain which diseases (and their regional variations) are currently the most prevalent in developing regions in India. After manually searching through the many reports, she discovers that a particular variant of tuberculosis is becoming more and more prevalent, that is the Multi-drug-resistant tuberculosis (MDR-TB). Next, she looks up the `http://clinicaltrials.gov` website and uses the keyword "Tuberculosis" in order to find the countries where the most number of clinical trials for tuberculosis are being conducted. Her query results in a list of the clinical trials for tuberculosis containing all the information that is reported for each clinical trial such as verification date, sponsor, secondary outcomes etc., which is not relevant for his analysis. Therefore, an additional burden for her is to extract relevant information from each of the trials and store it in a separate file.

After retrieving the results from his query, she finds that some of the trials do not contain complete and accurate information, which may hamper the results. In other trails, she notices that the data is not updated and thus her analysis misses out on valuable information. Also, after extracting relevant information from the datasets, she needs to perform statistical tests on the data. But, since the datasets are not in a single format and not aligned with each other, she has to manually record the values in another file in a

format, which will help in the statistical calculation. This task of manually gathering statistical values is a very cumbersome and time-consuming process and may also lead to errors since it involves a lot of manual work. Additionally, analyzing the results will also pose problems in case any of the presumptions (such as the region, type of disease or time frame of the analysis) need to be changed. Also, she will not be able to view the disparity over time as this information is also dispersed and difficult to calculate manually. Due to these obstacles, the analysis performed consequently leads to an inappropriate allocation of funds neglecting the most threatening diseases, despite the significant amount of work Ms. Sharma spend in this case.

1.3. Challenges

We identified the following challenges in the area of data quality, specifically for LD, and of the role of data quality for optimal utilization of LD: (1) Lack of unified descriptions for data quality dimensions and metrics for Linked Data, (2) Lack of user-driven data quality assessment methodologies for Linked Data and (3) Lack of quality assessment of datasets before utilization in particular use cases, which we outline in this section.

1.3.1. Lack of unified descriptions for data quality dimensions and metrics for Linked Data

There have been several different definitions as well as classifications of data quality dimensions and metrics proposed in the literature [Wang and Strong, 1996, Wand and Wang, 1996, Redman, 1997, Naumann, 2002, Batini and Scannapieco, 2006, Jarke et al., 2010]. These concepts focus on non or semi-structured data sources. Bizer [Bizer, 2007] adapted these concepts and proposed several data quality dimensions into a classification scheme specifically for LD. Recently, however, there have been different notions of data quality in terms of the dimensions as well as the metrics that should be considered while assessing the quality of LD [Fürber and Hepp, 2011, Mendes et al., 2012b, Hogan et al., 2012, Gamble and Goble, 2011]. But, there is no consensus on the definitions of these data quality concepts or the categorization. Moreover, means of measuring these dimensions i.e. the metrics are not clearly assigned to each dimension. Thus, one is left with a myriad of data quality problems but without a guide to understand, appropriately choose and measure them.

1.3.2. Lack of user-driven data quality assessment methodologies for Linked Data

Data quality assessment is a well-known issue for data in any format, right from unstructured content to relational databases. In the case of Linked Data specifically, there have been several data quality assessment methodologies that have been proposed [Flemming, 2011, Guéret et al., 2012b, Mendes et al., 2012b]. However, these methodologies

are either very specific to a domain, thus being inapplicable to all use cases and unable to provide meaningful results for the task at hand.

There are several data quality metrics belonging to certain dimensions, which cannot be measured quantitatively, but require human judgement. Trustworthiness, relevancy, understandability are few examples, which require the user to subjectively measure the dimension. However, the existing methodologies do not involve users, be it publishers or consumers, in the assessment process. These methodologies are either fully automated, inhibiting the users from choosing the dataset of interest or semi-automated, thus demanding considerable amount of user expertise. Thus, the users are unable to choose the quality requirements of interest and are provided with results, which are hard to interpret thus leaving the user without clear insights as to how to improve the quality of the data used.

1.3.3. Lack of quality assessment of datasets before utilization in particular use cases

With a huge amount of data recently being published on the Web as LD, several different use cases are being made possible in different domains. However, one of the main obstacles for the reliability of the results of these use cases is the *data quality*. With data being either incomplete or inconsistent or in some cases, untrustworthy, these use cases are unreliable. Recent studies have shown that majority of these datasets suffer from data quality problems [Hogan et al., 2012]. However, there are very few studies that undertake quality assessment measures before utilizing the data in particular use cases. These datasets are used directly and the poor data quality significantly affects the results.

1.4. Research Questions and Contributions

In this section, we outline the key research questions (RQ) that address the aforementioned challenges along with our contributions towards each of them, which are:

- *RQ1: What are the existing approaches to assess the quality of Linked Data employing a conceptual framework integrating prior approaches?*
  - *RQ1.1: What are the data quality problems that each approach assesses?*
  - *RQ1.2: Which are the data quality dimensions and metrics supported by the proposed approaches?*
- *RQ2: How can we assess the quality of Linked Data using a user-driven methodology?*
  - *RQ2.1 How feasible is it to employ Linked Data experts to assess the quality issues of LD?*
  - *RQ2.2 How feasible is it to use a combination of user-driven and semi-automated methodology to assess the quality of LD?*

  - *RQ2.3 Is it possible to detect quality issues in LD data sets via crowdsourcing mechanisms?*
  - *RQ2.4 What type of crowd is most suitable for each type of quality issues?*
  - *RQ2.5 Which types of assessment errors are made by lay users and experts?*
  - *RQ2.6 How can we semi-automatically assess the quality of datasets and provide meaningful results to the user?*

- *RQ3: How can we exploit Linked Data for building the HER Observatory and ensure good data quality?*

1.4.1. Descriptions of data quality dimensions and metrics

The research question we aim to answer is:

- *RQ1: What are the existing approaches to assess the quality of Linked Data employing a conceptual framework integrating prior approaches?*

To address this question, we conducted a literature review following the systematic review procedures described in [Kitchenham, 2004, Moher et al., 2009]. As a result of the survey, we identified 30 different approaches that propose a data quality assessment methodology, specifically for LD. Further, we divide this general research question into the following sub-questions:

- *RQ1.1: What are the data quality problems that each approach assesses?*
- *RQ1.2: Which are the data quality dimensions and metrics supported by the proposed approaches?*

We first identified the problems that each of the 30 approaches addressed (RQ1.1) and then mapped these problems to a particular data quality dimension. We then unified the definitions that each approach provides and formalized them (RQ1.2) in Chapter 3 for each of the 18 identified dimensions. Additionally, we provided a total of 69 metrics for these dimensions (RQ1.2). Furthermore, we classified each metric into being qualitatively or quantitatively assessed.

1.4.2. User-driven data quality assessment methodologies

The research question we aim to answer is:

- *RQ2: How can we assess the quality of Linked Data using a user-driven methodology?*

In order to address this research question, we present three different data quality assessment methodologies, which are user-driven and/or sensitive to a use case. Firstly, we present a user-driven methodology for assessing the quality of LD resources comprising of a manual and a semi-automatic process. The research question we aim to answer is:

- *RQ2.1 How feasible is it to employ Linked Data experts to assess the quality issues of LD?*
- *RQ2.2 How feasible is it to use a combination of user-driven and semi-automated methodology to assess the quality of LD?*

In the *manual* process, the first phase includes the detection of common quality problems and their representation in a quality problem taxonomy. The second phase comprises of the evaluation of a large number of individual resources, according to the quality problem taxonomy, performed by users. This process is accompanied by a tool, namely *TripleCheckMate*, wherein a user assesses an individual resource and evaluates each fact for correctness. In this case, the user is a LD expert who is conversant with RDF. We then analyze the results to assess the feasibility of this approach (RQ2.1). In case of the *semi-automatic* process, the first phase involves the detection of common quality problems by the automatic creation of an extended schema for DBpedia. The second phase involves the generation and manual verification of schema axioms. We report results of applying this methodology to DBpedia and thus assess the feasibility of this approach (RQ2.2) in Chapter 4 .

Another means we employ for assessing the quality of LD is via *crowdsourcing*. We further break down our research question into the following:

- *RQ2.3 Is it possible to detect quality issues in LD data sets via crowdsourcing mechanisms?*
- *RQ2.4 What type of crowd is most suitable for each type of quality issues?*
- *RQ2.5 Which types of errors are made by lay users and experts?*

We utilize the wisdom of the crowd, i.e. workers from online crowdsourcing platforms such as MTurk, to assess the quality of DBpedia. We analyze the results to assess the feasibility of this approach (RQ2.3). Then, we use the results from the previous user-driven assessment (performed by LD experts) and feed them to MTurk. We then compare the two methodologies in order to determine the type of crowd as well as cost and time feasibility of the approaches (RQ2.4). We analyze the types of errors made by users and experts by comparing the results from both the assessments (RQ2.5). We report the results obtained by applying both these methodologies to DBpedia in Chapter 5.

The third assessment methodology we propose is that which implements the data quality metrics identified in our survey to provide a tool, namely *R2RLint*, to assess the quality of LD. The research question we aim to answer here is:

- *RQ2.6 How can we semi-automatically assess the quality of datasets and provide meaningful results to the user?*

This tool takes as input an RDF dump or SPARQL Protocol and RDF Query Language (SPARQL) endpoint and the various quality metrics to assess the quality of any particular dataset. The user can choose which metrics are required based on the use case. Moreover,

the user is not only provided the results of the assessment but also specific entities that cause the errors, which help users understand the quality issues and thus can fix them. We provide the specific dimensions along with detailed explanations of the implementation of the metrics in Chapter 6. The results of the quality assessment of the four datasets that are part of our use case are reported in Chapter 7. We discuss the advantages of this semi-automated quality assessment over the other types of quality assessment methodologies discussed earlier.

1.4.3. Consumption of Linked Data leveraging on data quality

The research question we aim to answer is:

- *RQ3: How can we exploit Linked Data for a particular use case and ensure good data quality?*

In response to this question, we design a use case employing Linked Data to build the HER Observatory of societal progress indicators. We choose four linked datasets and integrate them to build the HER Observatory, which determines the impact of research and technology on health and economic performance of countries per year. In order to ensure *good* data quality of the datasets, we perform semi-automated quality assessment on all the four datasets involved in the use case. We employ the R2RLint tool to perform this assessment, wherein the metrics are chosen based on the use case, thus being use case specific. Also, the user is provided with the underlying triples causing the quality problems, thus being able to improve the quality. We show the importance of the role of data quality assessment and improvement in such a use case. We provide details of the use case, results of the data quality assessment and results of the use case in Chapter 7.

1.5. Thesis Overview

As depicted in Figure 1.2, this thesis is divided into seven chapters, which are described in this section.

- **Chapter 2** introduces the concepts of the Semantic Web and its associated technologies, which constitutes the basic scientific background required for the reader to understand the thesis. The chapter introduces the reader to the fundamentals of the Semantic Web followed by discussing the RDF language and its components . Thereafter, the various RDF serialization formats (e.g. N-Triples) and the differences among them are explained. Then, the crucial topic of Semantic Web, the ontology and the various languages that can be uses to develop the ontologies are discussed. At the end, the SPARQL query language, triple stores and how they support the SPARQL language are described.
- In **Chapter 3**, 18 data quality dimensions are introduced and defined with the help of examples. These 18 dimensions have been identified as a result of the systematic

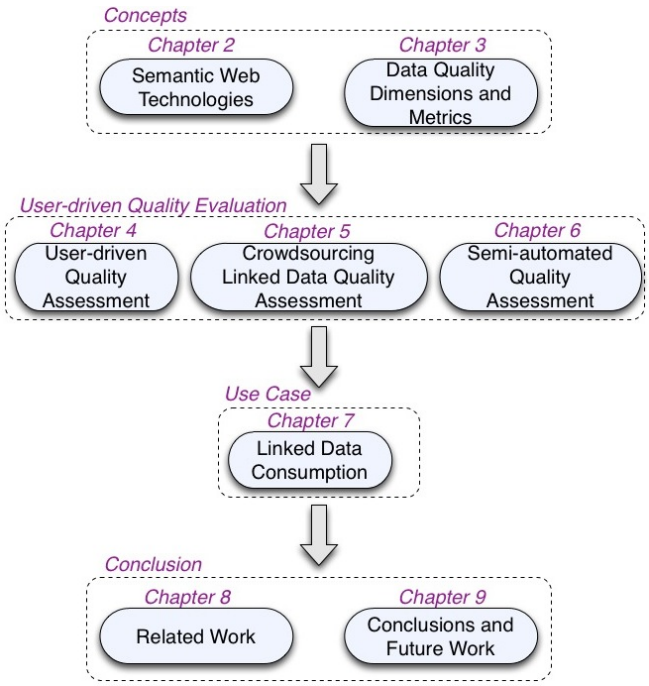


Figure 1.2.: Overview of the thesis structure.

literature review performed by us. We unify and formalize the definitions for each of them by combining those reported in previous literature. Additionally, metrics corresponding to each dimension are provided and are accompanied with a description on how they can be measured. Moreover, the metrics are classified into being either qualitative (QL) or quantitatively (QN) assessed.

- **Chapter 4** contains details of two user-driven methodologies for assessing the quality of DBpedia. The first methodology includes the employment of LD experts for the assessment. This assessment is performed with the help of the TripleCheckMate tool, which was developed specifically to assist LD experts for assessing the quality of any dataset, in this case DBpedia. The second methodology is a *semi-automatic* process, in which the first phase involves the detection of common quality problems by the automatic creation of an extended schema for DBpedia. The second phase involves the generation and manual verification of schema axioms. Results of applying both these methodologies to assess the quality of DBpedia are reported.

- In **Chapter 5**, the crowdsourcing data quality assessment methodology is outlined with the particulars of the tasks and results of employing workers from MTurk to perform the quality assessment of DBpedia. This chapter also includes a comparison between the two approaches (previous assessment by LD experts and assessment by MTurk workers in this study) in order to measure the feasibility of each type of the user-driven data quality assessment methodology.

- The **Chapter 6** proposes another semi-automated methodology for LD quality assessment, which also involves human judgement. In this semi-automated methodology, selected metrics are formally defined and implemented as part of a tool, namely R2RLint. This work was developed by a colleague from the AKSW group but built upon the metrics identified as part of the survey performed by the author of this thesis (as described in Chapter 3). The user is not only provided the results of the assessment but also specific entities that cause the errors, which help users understand the quality issues and thus can fix them. Details of these metrics along with formulae of the assessments are provided in this chapter.

- In **Chapter 7**, the details of a use case of utilizing LD for building the HER Observatory for several societal progress indicators, is presented. The use case involves the assessment of the impact of research and technology on healthcare and economic performance of each country per year. This chapter also includes the specifics of performing data quality assessment on the four datasets involved in this use case by using the semi-automated methodology (described in Chapter 6) backed by user-involvement. Thus, this chapter brings together both the challenges that this thesis addresses, that is, utilization of LD for a specific use case enhanced with the assessment of the quality of the datasets involved. We illustrate the importance of assessment of data quality in this use case and how the quality affects the end results. Additionally, we describe the advantages of this semi-automated quality assessment over the other types of quality assessment methodologies discussed earlier.

- **Chapter 8** provides an overview of the state-of-the-art in four areas that are part of this thesis. First, a discussion on the various data quality dimensions already available in the literature is presenting portraying that there is no harmony in the dimensions, their definitions and the classification that currently exists. Then, the various data quality assessment efforts undertaken are examined. These efforts are either done on the entire Web of Data, LD, a representative part of it or on particular dataset (e.g. DBpedia). We also discuss the existing efforts on crowdsourcing quality assessment undertaken. Thereafter, a qualitative comparison of 12 tools based on eight different attributes is presented. These 12 tools are identified from the 30 articles that are part of our survey. The eight attributes are (i) accessibility, (ii) licensing, (iii) automation, (iv) collaboration, (v) customizability, (vi) scalability, (vii) usability and (viii) maintenance. Finally, a discussion on the current efforts to calculate societal progress indicators is provided with the different methodologies and organisations that are involved in this process.

- Finally, **Chapter 9** summarizes the main contributions of this thesis and outlines directions for future research.

2. Semantic Web Technologies

This chapter gives a general overview of the Semantic Web. It describes the basic concepts, different RDF serialization formats as well as the ontology and its languages in detail. This chapter is mainly based on [Yu, 2007][1].

The rest of the chapter is organized as follows: In Section 2.1, we define Semantic Web as a whole. In Section 2.2, we describe RDF and its advantages. In Section 2.2.1, Section 2.2.2 and Section 2.2.3 we describe the basic elements of RDF in more detail. In Section 2.2.4, we introduce the RDF serialization formats. In Section 2.3, we define the term Ontology. In Section 2.3.1, we describe the ontology languages. In Section 2.4, we describe the SPARQL query language. Finally, in Section 2.5, we explain triplestores.

2.1. The Semantic Web Vision

With the widespread adoption of the World Wide Web, it has become a common place to share information around the world. This current Web infrastructure supports a distributed network of web pages that can refer to one another with global links called Uniform Resource Locators (URLs). However, the main idea of the Semantic Web is to support a distributed Web not at the level of the data rather at the level of the representation. The idea is that instead of having one web page indicate another, one data item can indicate another using global references called Uniform Resource Identifiers (URIs). The data model used by the Semantic Web infrastructure to represent this distributed web of data is called the Resource Description Framework (described in Section 2.2).

There are many different definitions of the Semantic Web. Tim Berners-Lee, the inventor of the World Wide Web, defined it as "not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation." [Berners-Lee et al., 2001] In other words, Semantic Web allows the machines not only to present data but also to process it.

There is a dedicated team of people at the World Wide Web consortium working towards improving, extending and standardizing the Semantic Web, and many languages, publications, tools have already been developed (e.g. [Tramp et al., 2010]). W3C has defined Semantic Web as "the idea of having data on the Web defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration, and reuse of data across various applications." [World Wide Web Consortium,

[1] Standard components of the Semantic Web and definitions for them are taken from this book as they follow the defined standards and are widely used. Examples for each component are provided by the author.

2009] In other words, Semantic Web is the machine-readable Web and can be thought of as an efficient way of representing the data on the World Wide Web or as a globally linked database.

2.2. Resource Description Framework (RDF)

The basic representation languages of the Semantic Web are RDF, RDFS, and Web Ontology Language (OWL), with RDF serving as the foundation. RDF is an XML-based language for describing information contained in a Web resource. This Web resource can be anything, for example a Web page or a Web site. RDF is the basic building block for supporting the Semantic Web, and is same as HTML is for the conventional Web. RDF relies heavily on the infrastructure of the Web, using many of its familiar and proven features, while extending them to provide a foundation for a distributed network of data.

The properties of RDF are:

- RDF is a language recommended by W3C [World Wide Web Consortium, 2004], which serves in managing the distributed data.
- RDF is capable of describing any fact (resource) independent of any domain.
- RDF provides a basis for *coding*, *exchanging*, and *reusing* structured (meta)data.
- RDF is structured; i.e. it is machine-readable. Machines can do useful operations with the knowledge expressed in RDF.
- RDF allows *interoperability* among applications by exchanging *machine understandable* information on the Web.

RDF has several basic elements, namely *Resource*, *Property* and *Statement*, which are discussed in the following subsections.

2.2.1. Resource

In the Semantic Web, we refer to the things in the world that are described by an RDF expression as resources (or entities or things). The resource can be a Web site, a person or anything else that one wants to talk about. Resource is identified by a **U**niform **R**esource **I**dentifier (**URI**). The rationale of using URIs is that the name of a resource must be globally unique.

In fact, the URLs, commonly used for accessing Web sites, are simply a subset of URIs. URIs take the same format as URLs, for example, `http://aksw.org/AmrapaliZaveri` and in fact the URL is just a special case of the URI. The main reason behind this is that the domain name used in the URL is guaranteed to be unique, therefore the uniqueness of the resource is ensured. Any two Web applications in the world can refer to the same thing by referencing the same URI. Unlike URLs, URIs may or may not refer to an actual Web site or a Web page.

2.2.2. Property

Property is a resource that has a name and can also be used to describe some specific characteristic, attribute, aspect or relation of the given resource. For instance, http://xmlns.com/foaf/0.1/name, denotes the name of some thing. In other words, this property relates a resource representing a thing to its name as shown in Figure 2.1.

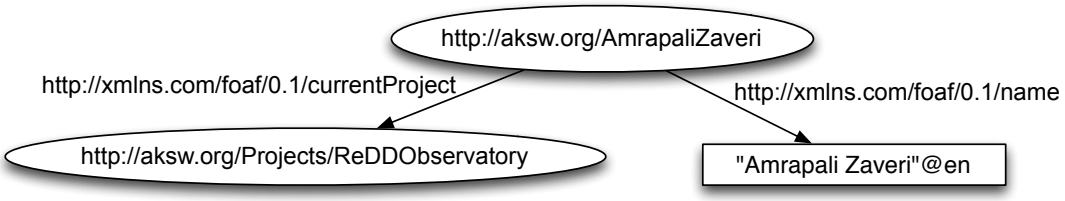


Figure 2.1.: RDF statement represented as a directed graph.

2.2.3. Statement

An RDF Statement is used to describe properties of resources. It is also called a triple and has the following format:

<resource (subject)> <property (predicate)> <property value (object)>.

The property value (object) can be a string, literal or another resource referenced by the URI. For example:

<http://aksw.org/AmrapaliZaveri>
<http://xmlns.com/foaf/0.1/currentProject>
<http://aksw.org/Projects/ReDDObservatory>.

This RDF statement simply states "The subject identified by http://aksw.org/AmrapaliZaveri has a property identified by http://xmlns.com/foaf/0.1/currentProject, whose value is equal to http://aksw.org/Projects/ReDDObservatory". This means that the person "Amrapali Zaveri" has a "currentProject" which is "ReDD-Observatory".

Another example:

<http://aksw.org/AmrapaliZaveri>
<http://xmlns.com/foaf/0.1/name>
"Amrapali Zaveri"@en.

This RDF statement states "The subject identified by http://aksw.org/AmrapaliZaveri has the property identified by http://xmlns.com/foaf/0.1/name, whose value is equal to "Amrapali Zaveri". This means that the person "Amrapali Zaveri" has a "name" whose value is "Amrapali Zaveri" and the trailing "@en" is the English language tag. In fact, RDF statements can also be expressed as directed graphs, as shown in Figure 2.1.

Subject	Predicate	Object
aksw:AmrapaliZaveri	rdf:type	foaf:Person
aksw:AmrapaliZaveri	foaf:age	"29"^^xsd:int
aksw:AmrapaliZaveri	foaf:skypeID	"amrapaliz"
aksw:AmrapaliZaveri	foaf:birthday	"1984-01-01"^^xsd:date
aksw:AmrapaliZaveri	foaf:name	"Amrapali Zaveri"@en
aksw:AmrapaliZaveri	foaf:currentProject	akswProject:ReDDObservatory
akswProject:ReDDObservatory	foaf:homepage	<http://redd.aksw.org>

Table 2.1.: Sample RDF statements.

It is to be noted here that the subject or the object or both can be an anonymous resource, called a "blank node". Blank nodes are used basically when the key purpose of a specific resource is to provide a context for some other properties to appear. In order to distinguish a blank node from the others, the RDF parser generates an *internal* unique identifier for each blank node. In other words, this identifier given to the blank node helps in identifying the node in a certain RDF document and the URI given to a resource is assured to be globally unique.

Since a URIs can be large, there is a short format for writing them i.e. by using a prefix. For instance, if we use `http://aksw.org/` as a prefix and give it a label e.g. `aksw`, then resource `http://aksw.org/AmrapaliZaveri` can be written as `aksw:AmrapaliZaveri`. Similarly, if `http://xmlns.com/foaf/0.1/` is used as a prefix with label `foaf`, then the properties `http://xmlns.com/foaf/0.1/name` and `http://xmlns.com/foaf/0.1/currentProject`, can be written as `foaf:name` and `foaf:currentProject` in short form. This format is very useful in writing human-readable RDF statements.

Whenever more triples describing a specific resource are added, the machine gets more knowledge about that resource. Table 2.1 shows more RDF statements about Amrapali Zaveri. This means that the resource of Amrapali Zaveri is the subject of other statements, which give more details about that resource. It should be noted that the object of a particular statement can be in turn the subject of other statement(s), e.g. Amrapali Zaveri has a current project identified by URI `akswProject:ReDDObservatory` and the knowledge base contains more information about that project as well. Also, it should be noted that the object of the second and fifth statement (a number and a date) has a trailing datatype. This small knowledge base can also be viewed as a directed graph as shown in Figure 2.2.

Using these simple RDF statements one can pose complex queries to the machine, e.g. "What is the homepage of Amrapali Zaveri's current project?".

2.2.4. RDF Serialization Formats

Serializing RDF data is a very crucial issue since different platforms and environments work better with different data formats. The issue of representing RDF in text not only arise in books and documents about RDF; it also arises when we want to publish data in RDF on the Web. In response to this need, there are several formats for serializing

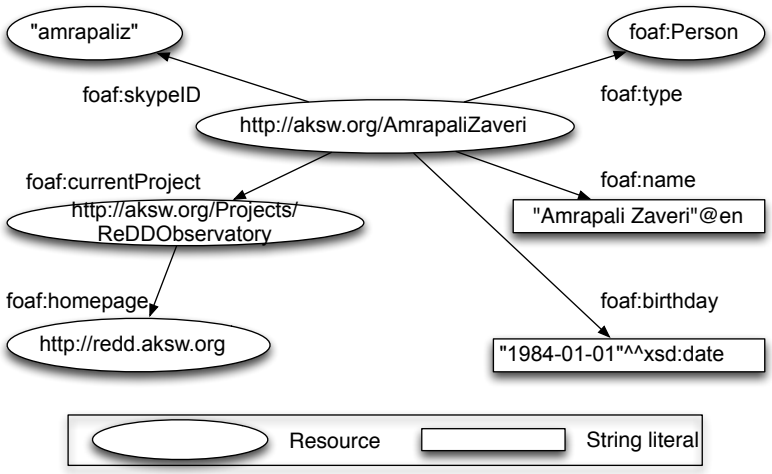


Figure 2.2.: Small knowledge base about Amrapali Zaveri represented as a graph.

```
1 <http://aksw.org/AmrapaliZaveri> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <
  http://xmlns.com/foaf/0.1/Person> .
2 <http://aksw.org/AmrapaliZaveri> <http://xmlns.com/foaf/0.1/age> "29"^^<http://www.w3.
  org/2001/XMLSchema#int> .
3 <http://aksw.org/AmrapaliZaveri> <http://xmlns.com/foaf/0.1/skypeID> "amrapaliz" .
4 <http://aksw.org/AmrapaliZaveri> <http://xmlns.com/foaf/0.1/birthday> "1984-01-01"^^<
  http://www.w3.org/2001/XMLSchema#date> .
5 <http://aksw.org/AmrapaliZaveri> <http://xmlns.com/foaf/0.1/currentProject> <http://
  aksw.org/Projects/ReDDObservatory> .
6 <http://aksw.org/AmrapaliZaveri> <http://xmlns.com/foaf/0.1/name> "Amrapali Zaveri"@en
   .
7 <http://aksw.org/Projects/ReDDObservatory> <http://xmlns.com/foaf/0.1/homepage> <http
  ://redd.aksw.org> .
```

Figure 2.3.: Sample N-Triples format.

RDF data such as N-Triples, RDF/XML, N3 and Turtle. Each of these is discussed along with an example in the following sections.

2.2.4.1. N-Triples

N-Triples is a simple line-based RDF serialization format and corresponds most directly to the raw RDF triples. It refers to resources using their fully unabbreviated URIs. Each RDF triple is written as a separate line, each URI between angle brackets (<and>) and terminated by a period (.). Typically files with N-Triples have the .nt extension [Grant and Beckett, 2004]. Figure 2.3 indicates our sample triples encoded in N-Triples format.

2.2.4.2. RDF/XML

RDF/XML represents RDF triples in XML format [Beckett, 2004]. The RDF/XML format is more convenient for machines than N-Triples since the traditional XML

```
<rdf:RDF xmlns:log="http://www.w3.org/2000/10/swap/log#" xmlns:rdf="http://www.w3.org
/1999/02/22-rdf-syntax-ns#">
  <rdf:Description rdf:about="http://aksw.org/Projects/ReDDObservatory">
    <homepage xmlns="http://xmlns.com/foaf/0.1/" rdf:resource="http://redd.aksw.org"/>
  </rdf:Description>

  <Person xmlns="http://xmlns.com/foaf/0.1/" rdf:about="http://aksw.org/AmrapaliZaveri
">
    <currentProject rdf:resource="http://aksw.org/Projects/ReDDObservatory"/>
    <birthday rdf:datatype="http://www.w3.org/2001/XMLSchema#date">1984-01-01</
birthDate>
    <age rdf:datatype="http://www.w3.org/2001/XMLSchema#int">29</age>
    <skypeID xmlns="http://dbpedia.org/property/" xml:lang="en">amrapaliz</skypeID>
    <name xmlns="http://dbpedia.org/property/" xml:lang="en">Amrapali Zaveri</name>
  </Person>
</rdf:RDF>
```

Figure 2.4.: Sample RDF/XML format.

```
@prefix aksw: <http://aksw.org/> .
@prefix akswProject: <http://aksw.org/Projects/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

aksw:AmrapaliZaveri a foaf:Person;
    foaf:age "29"^^xsd:int;
    foaf:currentProject akswProject:ReDDObservatory;
    foaf:birthday "1984-01-01"^^xsd:date;
    foaf:skypeID "amrapaliz";
    foaf:name "Amrapali Zaveri"@en .

akswProject:OntoWiki foaf:homepage <http://redd.aksw.org> .
```

Figure 2.5.: Sample N3 format.

format is commonly adopted and there are a variety of libraries available that simplify interaction with this format. Figure 2.4 shows our RDF example in RDF/XML format. Files containing RDF/XML data have .rdf as the file extension.

2.2.4.3. N3

N3 stands for Notation3 and is a shorthand notation for representing RDF graphs. N3 was designed to be easily read by humans and it is not an XML-compliant language [Berners-Lee and Connolly, 2011]. Figure 2.5 shows our RDF example in N3 format. Files containing RDF data in N3 format normally have a .n3 extension.

2.2.4.4. Turtle

The Turtle serialisation format is a subset of N3. Turtle stands for Terse RDF Triple Language. Turtle files have a .ttl extension [Dave and Berners-Lee, 2011]. This particular serialization is popular among developers of the Semantic Web.

2.3. Ontology

W3C defines an ontology as "the terms used to describe and represent an area of knowledge." [Heflin, 2004].

This definition has several aspects that should be discussed. First, the definition states that an ontology is used to describe and represent an area of knowledge. In other words, an ontology is domain specific; it does not represents all knowledge areas, but one specific area of knowledge. A domain is simply a specific subject area or sphere of knowledge, such as literature, medicine, education, etc.

Second, the ontology contains terms and relationships among those terms. Terms are also called classes or concepts; these words are interchangeable. The relationships between these classes can be expressed by using a hierarchy, i.e. superclasses represent higher-level concepts and subclasses represent finer concepts. The finer concepts inherit all the features and attributes that the higher concepts have.

Third, in addition to the aforementioned relationships among classes, there is another level of relationship expressed by using a special group of terms called properties. These property terms describe various features and attributes of the concepts and they can also be used to associate different classes together. Thus, the relationships among classes are not only superclass or subclass relationships, but relationships expressed in terms of properties as well.

In other words, an ontology defines a set of classes (e.g. "Person", "Book", "Writer"), and their hierarchy, i.e. which class is a subclass of another one (e.g. "Writer" is a subclass of "Person"). The ontology also defines how these classes interact with each other, i.e. how different classes are connected to each other via properties (e.g. a "Book" has an author of type "Writer").

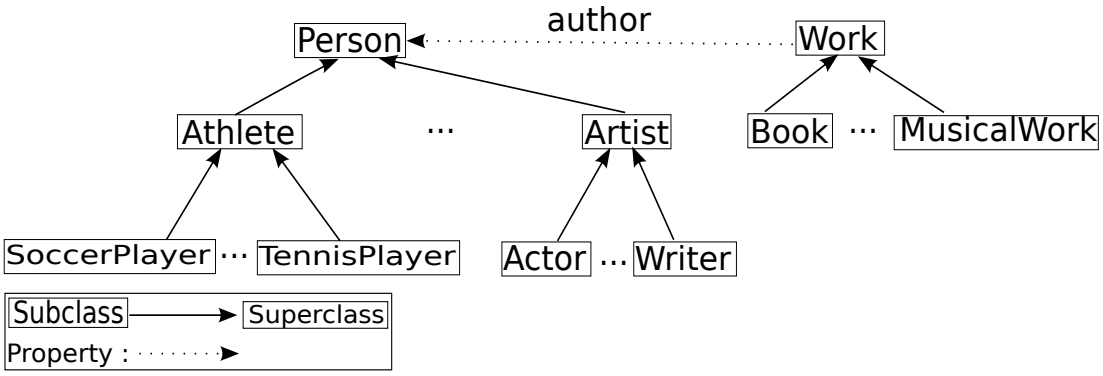


Figure 2.6.: Excerpt of the DBpedia ontology.

Figure 2.6 shows an excerpt of the ontology representing DBpedia[2]. This ontology shows that there is a class called "Writer" which is a subclass of the class "Artist", which in turn a subclass of "Person". *William Shakespeare, Johann Wolfgang von Goethe*, and *Dan Brown* are candidate instances of the class "Writer". The same applies to the

[2]http://dbpedia.org/

class “Work” and its subclasses. Note that there is a property called “author” relating an instance of class “Work” to an instance of the class “Person” i.e. it relates a work to its author. For instance, the book titled “First Folio” is an instance of classes “Work” and “Book”, and related via property “author” to its author “William Shakespeare”, which is an instance of the classes “Person”, “Artist” and “Writer”.

The main benefits of using an ontology are that it:

- enables a shared and common understanding about certain key concepts in a domain,
- facilitates a way for reuse of domain knowledge,
- makes the domain assumptions explicit and
- provides a way to combine knowledge and semantics in such a way that machines can understand it.

2.3.1. Ontology Languages

The question now is “What are the languages used to create ontologies?”. There are several languages, which can be used to encode ontologies such as Resource Description Framework Schema (RDFS) and OWL.

2.3.1.1. RDFS

RDFS is an ontology language, which can be used to create a vocabulary for describing classes, subclasses and properties of RDF resources and it is a W3C recommendation [Brickley and Guha, 2004]. The RDFS language also associates the properties with the classes it defines. RDFS can add semantics to RDF predicates and resources, i.e. it defines the meaning of a given term by specifying its properties and what kinds of objects these properties can have. It is worth noting here that RDFS is written in RDF, so any RDFS document is a legal RDF document.

2.3.1.2. OWL

The Web Ontology Language (OWL), built on RDFS, is used to create ontologies and is also a W3C recommendation [Bechhofer et al., 2004]. We can say that **OWL = RDFS + new constructs for expressiveness**. All classes and properties provided by RDFS can be used in OWL ontologies. OWL and RDFS have the same purpose, which is defining classes, properties and relations among these classes. OWL has an advantage over RDFS, which is its capability to express more complex relationships.

Due to its expressiveness power, most ontology developers use OWL to develop their ontologies. For example, an ontology developer can create a new class as the union or intersection of two or more classes using the expressive power of OWL. With OWL one can also declare that two classes are representing the same thing. For instance, consider the case that there are two separate ontologies created by different developers.

```
<http://dbpedia.org/ontology/Person> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
 <http://www.w3.org/2002/07/owl#Class> .
<http://dbpedia.org/ontology/Artist> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
 <http://www.w3.org/2002/07/owl#Class> .
<http://dbpedia.org/ontology/Artist> <http://www.w3.org/2000/01/rdf-schema#subClassOf>
 <http://dbpedia.org/ontology/Person> .
<http://dbpedia.org/ontology/Writer> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
 <http://www.w3.org/2002/07/owl#Class> .
<http://dbpedia.org/ontology/Writer> <http://www.w3.org/2000/01/rdf-schema#subClassOf>
 <http://dbpedia.org/ontology/Artist> .
<http://dbpedia.org/ontology/Work> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <
http://www.w3.org/2002/07/owl#Class> .
<http://dbpedia.org/ontology/Book> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <
http://www.w3.org/2002/07/owl#Class> .
<http://dbpedia.org/ontology/Book> <http://www.w3.org/2000/01/rdf-schema#subClassOf> <
http://dbpedia.org/ontology/Work> .
<http://dbpedia.org/ontology/author> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
 <http://www.w3.org/2002/07/owl#ObjectProperty> .
<http://dbpedia.org/ontology/author> <http://www.w3.org/2000/01/rdf-schema#domain> <
http://dbpedia.org/ontology/Work> .
<http://dbpedia.org/ontology/author> <http://www.w3.org/2000/01/rdf-schema#range> <
http://dbpedia.org/ontology/Person> .
```

Figure 2.7.: OWL representation of a part of an ontology in N-Triples format.

In the first ontology there is a class called “Poet” and in the other ontology there is a class called “PoetryWriter”. In fact, these classes are equivalent to each other and in RDFS one cannot declare that these classes are equivalent, but with OWL one can.

OWL provides some powerful features for properties as well. For example, in OWL one can declare that two properties are the inverse of each other, (e.g. author, and isAuthorOf). Figure 2.7 indicates a part of our ontology expressed in OWL.

Note that for the property author we have defined two properties domain and range. The domain property defines the class of instances, which can be the subject of that property (author property), while the range property defines the class of instances, which can be the object of that property.

OWL has many powerful features, interested readers can find more about these feature in [Bechhofer et al., 2004].

2.4. SPARQL Query Language

“The **S**PARQL **P**rotocol **A**nd **R**DF **Q**uery **L**anguage (SPARQL) is the W3C standard query language and protocol for RDF.” [Clark et al., 2008]. SPARQL allows the user to write queries that consist of triple patterns, conjunctions (logical “and”), disjunctions (logical “or”) and/or a set of optional patterns [Wikipedia, 2013]. Examples of these optional patterns are: `FILTER`, `REGEX` and `LANG`.

The SPARQL query specifies the pattern(s) that the resulting data should satisfy. The results of SPARQL queries can be result sets or RDF graphs. SPARQL has four query forms, specifically `SELECT`, `CONSTRUCT`, `ASK` and `DESCRIBE` [Prud’hommeaux and Seaborne, 2008].

```
1  PREFIX aksw: <http://aksw.org/>
2  PREFIX foaf: <http://xmlns.com/foaf/0.1/>
3  SELECT ?homepage
4    WHERE {aksw:AmrapaliZaveri foaf:currentProject ?project.
5    ?project foaf:homepage ?homepage. }
```

Figure 2.8.: SPARQL query to get the homepage of Amrapali Zaveri's current project.

Let us take an example to clarify the usage of SPARQL. Assume that we want to ask the query "What is the homepage of Amrapali Zaveri's current project?" to our small knowledge base. Figure 2.8 shows a SPARQL query to get information about the homepage of Amrapali Zaveri's current project.

In Figure 2.8, lines 1 and 2 define prefixes in order to write URIs in their short forms. Line 3 declares the variables that should be rendered to the output of that query, which is only one variable `?homepage`. Note that SPARQL variables start either with a question mark "?", or with a dollar sign "$". Line 4 states that for the statement with subject `aksw:AmrapaliZaveri` and property `foaf:currentProject`, we want the value of its object to be assigned to a variable called `?project`. Upon execution, this variable will take the value of `akswProject:ReDDObservatory`. In line 5, we want the variable `?project` which now has the value `akswProject:ReDDObservatory`, to be the subject of the next statement. In other words, the statement will be `akswProject:ReDDObservatory foaf:homepage ?homepage`. Now, variable `?homepage` is the only unknown variable of the statement, and it will take the value `http://redd.aksw.org`. Eventually, this value will be rendered to the output.

2.5. Triplestore

The crucial question here is "How do we store RDF data for efficient and quick access?". Basically, RDF data is stored in triplestores. A triplestore is a software program capable of storing and indexing RDF data efficiently, in order to enable querying this data easily and effectively. A triplestore for RDF data is like Relational Database Management System (DBMS) for relational databases.

Most triplestores support SPARQL query language for querying RDF data. As there are several DBMSs in the wild, such as Oracle[3], MySQL[4] and SQL Server[5], similarly there are several triplestores. Virtuoso [Erling and Mikhailov, 2009], Sesame [Broekstra et al., 2002] and BigOWLIM [Bishop et al., 2011] are typical examples of triplestores for desktop and server computers. DBpedia, for example, uses Virtuoso as the underlying triplestore.

[3] http://www.oracle.com/us/products/database/overview/index.html
[4] http://www.mysql.com
[5] http://www.microsoft.com/en-us/sqlserver/default.aspx

3. Linked Data Quality Dimension and Metrics

In this chapter, we first describe the basic concepts of data quality and then present a list of 18 quality dimensions and 69 metrics that can be applied for quality assessment of LD. These dimensions and metrics have been identified as a result of a literature review conducted in order to identify the approaches for assessing the quality of LD. As a result of the systematic literature review, as described in [Zaveri et al., 2015], a total of 30 articles (Table 3.1) were identified that proposed methodologies, dimensions and metrics for quality assessment of LD. We unify and define each dimension and provide different means to measure them (metrics) along with an example for each. The occurrences of each dimension in the 30 core articles are illustrated in Table 3.8. These dimensions and metrics form the core of this thesis as they are used in formulating the quality problem taxonomy (Chapter 4), which in turn is used to select the types of quality issues that are presented to the MTurk workers (Chapter 5). Also, specific metrics identified as a result of this survey are implemented as part of a tool (Chapter 6) and used to assess the quality of four datasets that are part of our use case (Chapter 7). This chapter is based on [Zaveri et al., 2015].

3.1. Conceptualization

3.1.1. Data Quality

Data quality is commonly conceived as a multi-dimensional construct with a popular definition "'fitness for use' [Juran, 1974]". Data quality may depend on various factors (dimensions or characteristics) such as accuracy, timeliness, completeness, relevancy, objectivity, believability, understandability, consistency, conciseness, availability and verifiability [Wang and Strong, 1996].

In terms of the Semantic Web, there exist different means of assessing data quality. The process of measuring data quality is supported by quality related metadata as well as data itself. On the one hand, provenance (as a particular case of metadata) information, for example, is an important concept to be considered when assessing the trustworthiness of datasets [Lei et al., 2007a]. On the other hand, the notion of link quality is another important aspect that is introduced in LD, where it is automatically detected whether a link is useful or not [Guéret et al., 2012a]. It is to be noted that *data* and *information* are interchangeably used in the literature.

Table 3.1.: List of the selected papers.

Citation	Title
Gil et al., 2002 [Gil and Ratnakar, 2002]	Trusting Information Sources One Citizen at a Time
Golbeck et al., 2003 [Golbeck et al., 2003]	Trust Networks on the Semantic Web
Mostafavi et al., 2004 [Mostafavi et al., 2004]	An ontology-based method for quality assessment of spatial data bases
Golbeck, 2006 [Golbeck, 2006]	Using Trust and Provenance for Content Filtering on the Semantic Web
Gil et al., 2007 [Gil and Artz, 2007]	Towards content trust of Web resources
Lei et al., 2007 [Lei et al., 2007b]	A framework for evaluating semantic metadata
Hartig, 2008 [Hartig, 2008]	Trustworthiness of Data on the Web
Bizer et al., 2009 [Bizer and Cyganiak, 2009]	Quality-driven information filtering using the WIQA policy framework
Böhm et al., 2010 [Böhm et al., 2010]	Profiling linked open data with ProLOD
Chen et al., 2010 [Chen and Garcia, 2010]	Hypothesis generation and data quality assessment through association mining
Flemming, 2010 [Flemming, 2011]	Assessing the quality of a Linked Data source
Hogan et al.,2010 [Hogan et al., 2010]	Weaving the Pedantic Web
Shekarpour et al., 2010 [Shekarpour and Katebi, 2010]	Modeling and evaluation of trust with an extension in semantic web
Fürber et al.,2011 [Fürber and Hepp, 2011]	SWIQA – a semantic web information quality assessment framework
Gamble et al., 2011 [Gamble and Goble, 2011]	Quality, Trust and Utility of Scientific Data on the Web: Towards a Joint Model
Jacobi et al., 2011 [Jacobi et al., 2011]	Rule-Based Trust Assessment on the Semantic Web
Bonatti et al., 2011 [Bonatti et al., 2011]	Robust and scalable linked data reasoning incorporating provenance and trust annotations
Ciancaglini et al., 2012 [Dezani-Ciancaglini et al., 2012]	Tracing where and who provenance in Linked Data: a calculus
Guéret et al., 2012 [Guéret et al., 2012a]	Assessing Linked Data Mappings Using Network Measures
Hogan et al., 2012 [Hogan et al., 2012]	An empirical survey of Linked Data conformance
Mendes et al., 2012 [Mendes et al., 2012b]	Sieve: Linked Data Quality Assessment and Fusion
Rula et al., 2012 [Rula et al., 2012]	Capturing the Age of Linked Open Data: Towards a Dataset-independent Framework
Acosta et al., 2013 [Acosta et al., 2013]	Crowdsourcing Linked Data Quality Assessment
Zaveri et al., 2013 [Zaveri et al., 2013a]	User-driven Quality evaluation of DBpedia
Albertoni et al., 2013 [Albertoni and Perez, 2013]	Assessing Linkset Quality for Complementing Third-Party Datasets
Feeney et al., 2014 [Feeney et al., 2014]	Improving curated web-data quality with structured harvesting and assessment
Kontokostas et al., 2014 [Kontokostas et al., 2014]	Test-driven Evaluation of Linked Data Quality
Paulheim et al., 2014 [Paulheim and Bizer, 2014]	Improving the Quality of Linked Data Using Statistical Distributions
Ruckhaus et al., 2014 [Ruckhaus et al., 2014]	Analyzing Linked Data Quality with LiQuate
Wienand et al., 2014 [Wienand and Paulheim, 2014]	Detecting Incorrect Numerical Data in DBpedia

3.1.2. Data Quality Problems

Bizer et al. [Bizer and Cyganiak, 2009] relate data quality problems to those arising in web-based information systems, which integrate information from different providers. For Mendes et al. [Mendes et al., 2012b], the problem of data quality is related to values being in conflict between different data sources as a consequence of the diversity of the data. Flemming [Flemming, 2011], on the other hand, implicitly explains the data quality problems in terms of *data diversity*. Hogan et al. [Hogan et al., 2010, Hogan et al., 2012] discuss about *errors*, *noise*, *difficulties* or *modelling issues*, which are prone to the non-exploitations of the data from the applications. Thus, the term *data quality problem* refers to a set of issues that can affect the potentiality of the applications that use the data.

3.1.3. Data Quality Dimensions and Metrics

Data quality assessment involves the measurement of quality *dimensions* or *criteria* that are relevant to the consumer. The dimensions can be considered as the characteristics of a dataset. A data quality assessment *metric*, *measure* or *indicator* is a procedure for measuring a data quality dimension [Bizer and Cyganiak, 2009]. These metrics are heuristics that are designed to fit a specific assessment situation [Leo Pipino and Rybold, 2005]. Since dimensions are rather abstract concepts, the assessment metrics rely on quality *indicators* that allow the assessment of the quality of a data source w.r.t the criteria [Flemming, 2011]. An assessment score is computed from these indicators using a scoring function.

There are a number of studies, which have identified, defined and grouped data quality dimensions into different classifications [Wang and Strong, 1996, Wand and Wang, 1996, Redman, 1997, Naumann, 2002, Batini and Scannapieco, 2006, Jarke et al., 2010]. For example, Bizer et al. [Bizer and Cyganiak, 2009], classified the data quality dimensions into three categories according to the type of information that is used as a quality dimension: (i) Content Based – information content itself; (ii) Context Based – information about the context in which information was claimed; (iii) Rating Based – based on the ratings about the data itself or the information provider. However, we identify further dimensions and classify the dimensions into the (i) Accessibility (ii) Intrinsic (iii) Contextual and (iv) Representational groups.

3.1.4. Data Quality Assessment Methodology

A data quality assessment methodology is defined as the process of evaluating if a piece of data meets the information consumers need in a specific use case. The process involves measuring the quality dimensions that are relevant to the user and comparing the assessment results with the user's quality requirements [Bizer and Cyganiak, 2009].

3.2. Linked Data Quality dimensions

After analyzing the 30 selected approaches in detail, we identified a core set of 18 different data quality dimensions that can be applied to assess the quality of LD. We grouped the identified dimensions according to the classification introduced in [Wang and Strong, 1996]:

- Accessibility dimensions
- Intrinsic dimensions
- Contextual dimensions
- Representational dimensions

We further re-examine the dimensions belonging to each group and change their membership according to the LD context. In this section, we unify, formalize and adapt the definition for each dimension according to LD. For each dimension, we identify metrics and report them too. In total, 69 metrics are provided for all the 18 dimensions. Furthermore, we classify each metric as being quantitatively or qualitatively assessed. Quantitatively (QN) measured metrics are those that are quantified or for which a concrete value (score) can be calculated. Qualitatively (QL) measured metrics are those which cannot be quantified and depend on the users perception of the respective metric.

In general, a group captures the same essence for the underlying dimensions that belong to that group. However, these groups are not strictly disjoint but can partially overlap since there exist trade-offs between the dimensions of each group as described in Section 3.2.5. Additionally, we provide a general use case scenario and specific examples for each of the dimensions. In certain cases, the examples point towards the quality of the information systems such as search engines (e.g. performance) and in other cases, about the data itself.

Use case scenario. Since data quality is conceived as "fitness for use", we introduce a specific use case that will allow us to illustrate the importance of each dimension with the help of an example. The use case is about an intelligent flight search engine, which relies on aggregating data from several datasets. The search engine obtains information about airports and airlines from an airline dataset (e.g. *OurAirports*[1], *OpenFlights*[2]). Information about the location of countries, cities and particular addresses is obtained from a spatial dataset (e.g. *LinkedGeoData*[3]). Additionally, aggregators pull all the information related to flights from different booking services (e.g. *Expedia*[4]) and represent this information as RDF. This allows a user to query the integrated dataset for a flight between any start and end destination for any time period. We will use this scenario throughout as an example to explain each quality dimension through a quality issue.

3.2.1. Accessibility dimensions

The dimensions belonging to this category involve aspects related to the access, authenticity and retrieval of data to obtain either the entire or some portion of the data (or from another linked dataset) for a particular use case. There are five dimensions that are part of this group, which are *availability, licensing, interlinking, security* and *performance*. Table 3.2 displays metrics for these dimensions and provides references to the original literature.

3.2.1.1. Availability

Flemming [Flemming, 2011] referred to availability as the proper functioning of all access methods. The other articles [Hogan et al., 2010, Hogan et al., 2012] provide

[1] http://thedatahub.org/dataset/ourairports
[2] http://thedatahub.org/dataset/open-flights
[3] http://linkedgeodata.org
[4] http://www.expedia.com/

Table 3.2.: Data quality metrics related to accessibility dimensions (type QN refers to a quantitative metric, QL to a qualitative one).

Dimension	Abr	Metric	Description	Type
Availability	A1	accessibility of the SPARQL endpoint and the server	checking whether the server responds to a SPARQL query [Flemming, 2011]	QN
	A2	accessibility of the RDF dumps	checking whether an RDF dump is provided and can be downloaded [Flemming, 2011]	QN
	A3	dereferenceability of the URI	checking (i) for dead or broken links i.e. when an HTTP-GET request is sent, the status code `404 Not Found` is not be returned (ii) that useful data (particularly RDF) is returned upon lookup of a URI, (iii) for changes in the URI i.e the compliance with the recommended way of implementing redirections using the status code `303 See Other` [Flemming, 2011, Hogan et al., 2010]	QN
	A4	no misreported content types	detect whether the HTTP response contains the header field stating the appropriate content type of the returned file e.g. `application/rdf+xml` [Hogan et al., 2010]	QN
	A5	dereferenced forward-links	dereferenceability of all forward links: all available triples where the local URI is mentioned in the subject (i.e. the description of the resource) [Hogan et al., 2012]	QN
Licensing	L1	machine-readable indication of a license	detection of the indication of a license in the VoID description or in the dataset itself [Flemming, 2011, Hogan et al., 2012]	QN
	L2	human-readable indication of a license	detection of a license in the documentation of the dataset [Flemming, 2011, Hogan et al., 2012]	QN
	L3	specifying the correct license	detection of whether the dataset is attributed under the same license as the original [Flemming, 2011]	QN
Interlinking	I1	detection of good quality interlinks	(i) detection of (a) interlinking degree, (b) clustering coefficient, (c) centrality, (d) open sameAs chains and (e) description richness through sameAs by using network measures [Guéret et al., 2012a], (ii) via crowdsourcing [Acosta et al., 2013, Zaveri et al., 2013a]	QN
	I2	existence of links to external data providers	detection of the existence and usage of external URIs (e.g. using `owl:sameAs` links) [Hogan et al., 2012]	QN
	I3	dereferenced back-links	detection of all local in-links or back-links: all triples from a dataset that have the resource's URI as the object [Hogan et al., 2012]	QN
Security	S1	usage of digital signatures	by signing a document containing an RDF serialization, a SPARQL result set or signing an RDF graph [Carroll, 2003, Flemming, 2011]	QN
	S2	authenticity of the dataset	verifying authenticity of the dataset based on a provenance vocabulary such as the author and his contributors, the publisher of the data and its sources, if present in the dataset [Flemming, 2011]	QL
Performance	P1	usage of slash-URIs	checking for usage of slash-URIs where large amounts of data is provided [Flemming, 2011]	QN
	P2	low latency	(minimum) delay between submission of a request by the user and reception of the response from the system [Flemming, 2011]	QN
	P3	high throughput	(maximum) no. of answered HTTP-requests per second [Flemming, 2011]	QN
	P4	scalability of a data source	detection of whether the time to answer an amount of ten requests divided by ten is not longer than the time it takes to answer one request [Flemming, 2011]	QN

metrics for this dimension.

Definition 1 (Availability)**.** *Availability of a dataset is the extent to which data (or some portion of it) is present, obtainable and ready for use.*

Metrics. The metrics identified for availability are:

- A1: checking whether the server responds to a SPARQL query [Flemming, 2011]
- A2: checking whether an RDF dump is provided and can be downloaded [Flemming, 2011]
- A3: detection of dereferenceability of URIs by checking:
  - for dead or broken links [Hogan et al., 2010], i.e. that when an HTTP-GET request is sent, the status code `404 Not Found` is not returned [Flemming, 2011]
  - that useful data (particularly RDF) is returned upon lookup of a URI [Hogan et al., 2010]
  - for changes in the URI, i.e. compliance with the recommended way of implementing redirections using the status code `303 See Other` [Flemming, 2011]
- A4: detect whether the HTTP response contains the header field stating the appropriate content type of the returned file, e.g. `application/rdf+xml` [Hogan et al., 2010]
- A5: dereferenceability of all forward links: all available triples where the local URI is mentioned in the subject (i.e. the description of the resource) [Hogan et al., 2012]

Example. Let us consider the case in which a user looks up a flight in our flight search engine. She requires additional information such as car rental and hotel booking at the destination, which is present in another dataset and interlinked with the flight dataset. However, instead of retrieving the results, she receives an error response code `404 Not Found`. This is an indication that the requested resource cannot be dereferenced and is therefore unavailable. Thus, with this error code, she may assume that either there is no information present at that specified URI or the information is unavailable.

3.2.1.2. Licensing

Licensing is a new quality dimensions not considered for relational databases but mandatory in the LD world. Flemming [Flemming, 2011] and Hogan et al. [Hogan et al., 2012] both stated that each RDF document should contain a license under which the content can be (re)used, in order to enable information consumers to use the data under clear legal terms. Additionally, the existence of a machine-readable indication

(by including the specifications in a VoID[5] description) as well as a human-readable indication of a license are important not only for the permissions a license grants but as an indication of which requirements the consumer has to meet [Flemming, 2011]. Although both these studies do not provide a formal definition, they agree on the use and importance of licensing in terms of data quality.

Definition 2 (Licensing)**.** *Licensing is defined as the granting of permission for a consumer to re-use a dataset under defined conditions.*

Metrics. The metrics identified for licensing are:

- L1: machine-readable indication of a license in the VoID description or in the dataset itself [Flemming, 2011, Hogan et al., 2012]
- L2: human-readable indication of a license in the documentation of the dataset [Flemming, 2011, Hogan et al., 2012]
- L3: detection of whether the dataset is attributed under the same license as the original [Flemming, 2011]

Example. Since our flight search engine aggregates data from several existing data sources, a clear indication of the license allows the search engine to re-use the data from the airlines websites. For example, the LinkedGeoData dataset is licensed under the Open Database License[6], which allows others to copy, distribute and use the data and produce work from the data allowing modifications and transformations. Due to the presence of this specific license, the flight search engine is able to re-use this dataset to pull geo-spatial information and feed it to the search engine.

3.2.1.3. Interlinking

Interlinking is a relevant dimension in LD since it supports data integration. Interlinking is provided by RDF triples that establish a link between the entity identified by the subject with the entity identified by the object. Through the typed RDF links, data items are effectively interlinked. Even though the core articles in this survey do not contain a formal definition for interlinking, they provide metrics for this dimension [Guéret et al., 2012a, Hogan et al., 2010, Hogan et al., 2012].

Definition 3 (Interlinking)**.** *Interlinking refers to the degree to which entities that represent the same concept are linked to each other, be it within or between two or more data sources.*

Metrics. The metrics identified for interlinking are:

- I1: detection of:

[5] http://vocab.deri.ie/void
[6] http://opendatacommons.org/licenses/odbl/

  - interlinking degree: how many hubs there are in a network[7] [Guéret et al., 2012a]
  - clustering coefficient: how dense is the network [Guéret et al., 2012a]
  - centrality: indicates the likelihood of a node being on the shortest path between two other nodes [Guéret et al., 2012a]
  - whether there are open sameAs chains in the network [Guéret et al., 2012a]
  - how much value is added to the description of a resource through the use of sameAs edges [Guéret et al., 2012a]

- I2: detection of the existence and usage of external URIs (e.g. using `owl:sameAs` links) [Hogan et al., 2010, Hogan et al., 2012]

- I3: detection of all local in-links or back-links: all triples from a dataset that have the resource's URI as the object [Hogan et al., 2012]

Example. In our flight search engine, the instance of the country `"United States"` in the airline dataset should be interlinked with the instance `"America"` in the spatial dataset. This interlinking can help when a user queries for a flight, as the search engine can display the correct route from the start destination to the end destination by correctly combining information for the same country from both datasets. Since names of various entities can have different URIs in different datasets, their interlinking can help in disambiguation.

3.2.1.4. Security

Flemming [Flemming, 2011] referred to security as "the possibility to restrict access to the data and to guarantee the confidentiality of the communication between a source and its consumers". Additionally, Flemming referred to the verifiability dimension as the means a consumer is provided with to examine the data for correctness. Thus, security and verifiability point towards the same quality dimension i.e. to avoid alterations of the dataset and verify its correctness.

Definition 4 (Security)**.** *Security is the extent to which data is protected against alteration and misuse.*

Metrics. The metrics identified for security are:

- S1: using digital signatures to sign documents containing an RDF serialization, a SPARQL result set or signing an RDF graph [Flemming, 2011]

- S2: verifying authenticity of the dataset based on provenance information such as the author and his contributors, the publisher of the data and its sources (if present in the dataset) [Flemming, 2011]

[7] In [Guéret et al., 2012a], a network is described as a set of facts provided by the graph of the Web of Data, excluding blank nodes.

Example: In our use case, if we assume that the flight search engine obtains flight information from arbitrary airline websites, there is a risk for receiving incorrect information from malicious websites. For instance, an airline or sales agency website can pose as its competitor and display incorrect flight fares. Thus, by this spoofing attack, this airline can prevent users to book with the competitor airline. In this case, the use of standard security techniques such as digital signatures allows verifying the identity of the publisher.

3.2.1.5. Performance

Performance is a dimension that has an influence on the quality of the information system or search engine, not on the dataset itself. Flemming [Flemming, 2011] states "the performance criterion comprises aspects of enhancing the performance of a source as well as measuring of the actual values". Also, response-time and performance point towards the same quality dimension.

Definition 5 (Performance)**.** *Performance refers to the efficiency of a system that binds to a large dataset, that is, the more performant a data source is the more efficiently a system can process data.*

Metrics. The metrics identified for performance are:

- P1: checking for usage of slash-URIs where large amounts of data is provided[8] [Flemming, 2011]
- P2: low latency[9]: (minimum) delay between submission of a request by the user and reception of the response from the system [Flemming, 2011]
- P3: high throughput: (maximum) number of answered HTTP-requests per second [Flemming, 2011]
- P4: scalability: detection of whether the time to answer an amount of ten requests divided by ten is not longer than the time it takes to answer one request [Flemming, 2011]

Example. In our use case, the performance may depend on the type and complexity of the query by a large number of users. Our flight search engine can perform well by considering response-time when deciding which sources to use to answer a query.

3.2.1.6. Intra-relations

The dimensions in this group are related with each other as follows: performance (response-time) of a system is related to the availability dimension. A dataset can

[8] http://www.w3.org/wiki/HashVsSlash

[9] Latency is the amount of time from issuing the query until the first information reaches the user [Naumann, 2002].

perform well only if it is available and has low response time. Also, interlinking is related to availability because only if a dataset is available, it can be interlinked and these interlinks can be traversed. Additionally, the dimensions security and licensing are related since providing a license and specifying conditions for re-use helps secure the dataset against alterations and misuse.

3.2.2. Intrinsic dimensions

Intrinsic dimensions are those that are independent of the user's context. There are five dimensions that are part of this group, which are *syntactic validity, semantic accuracy, consistency, conciseness* and *completeness*. These dimensions focus on whether information correctly (syntactically and semantically), compactly and completely represents the real world and whether information is logically consistent in itself. Table 3.3 provides metrics for these dimensions along with references to the original literature.

3.2.2.1. Syntactic validity

Fürber et al. [Fürber and Hepp, 2011] classified accuracy into syntactic and semantic accuracy. He explained that a "value is syntactically accurate, when it is part of a legal value set for the represented domain or it does not violate syntactical rules defined for the domain". Flemming [Flemming, 2011] defined the term validity of documents as "the valid usage of the underlying vocabularies and the valid syntax of the documents". We thus associate the validity of documents defined by Flemming to syntactic validity. We similarly distinguish between the two types of accuracy defined by Fürber et al. and form two dimensions: *Syntactic validity* (syntactic accuracy) and *Semantic accuracy*. Additionally, Hogan et al. [Hogan et al., 2010] identify syntax errors such as RDF/XML syntax errors, malformed datatype literals and literals incompatible with datatype range, which we associate with syntactic validity. The other articles [Acosta et al., 2013, Feeney et al., 2014, Kontokostas et al., 2014, Wienand and Paulheim, 2014, Zaveri et al., 2013a] provide metrics for this dimension.

Definition 6 (Syntactic validity)**.** *Syntactic validity is defined as the degree to which an RDF document conforms to the specification of the serialization format.*

Metrics. The metrics identified for syntactic validity are:

- SV1: detecting syntax errors using (i) validators [Flemming, 2011, Hogan et al., 2010], (ii) via crowdsourcing [Acosta et al., 2013, Zaveri et al., 2013a]
- SV2: detecting use of:
  - explicit definition of the allowed values for a certain datatype, (ii) syntactic rules [Fürber and Hepp, 2011], (iii) detecting whether the data conforms to the specific RDF pattern and that the "types" are defined for specific resources [Kontokostas et al., 2014], (iv) use of different outlier techniques and clustering for detecting wrong values [Wienand and Paulheim, 2014]

Table 3.3.: Data quality metrics related to intrinsic dimensions (type QN refers to a quantitative metric, QL to a qualitative one).

Dimension	Abr	Metric	Description	Type
Syntactic validity	SV1	no syntax errors of the documents	detecting syntax errors using (i) validators [Flemming, 2011, Hogan et al., 2010], (ii) via crowdsourcing [Acosta et al., 2013, Zaveri et al., 2013a]	QN
	SV2	syntactically accurate values	by (i) use of explicit definition of the allowed values for a datatype, (ii) syntactic rules [Fürber and Hepp, 2011], (iii) detecting whether the data conforms to the specific RDF pattern and that the "types" are defined for specific resources [Kontokostas et al., 2014], (iv) use of different outlier techniques and clustering for detecting wrong values [Wienand and Paulheim, 2014]	QN
	SV3	no malformed datatype literals	detection of ill-typed literals, which do not abide by the lexical syntax for their respective datatype that can occur if a value is (i) malformed, (ii) is a member of an incompatible datatype [Feeney et al., 2014, Hogan et al., 2010]	QN
Semantic accuracy	SA1	no outliers	by (i) using distance-based, deviation-based and distribution-based methods [Bizer and Cyganiak, 2009, Feeney et al., 2014], (ii) using the statistical distributions of a certain type to assess the statement's correctness [Paulheim and Bizer, 2014]	QN
	SA2	no inaccurate values	by (i) using functional dependencies between the values of two or more different properties [Fürber and Hepp, 2011], (ii) comparison between two literal values of a resource [Kontokostas et al., 2014], (iii) via crowdsourcing [Acosta et al., 2013, Zaveri et al., 2013a]	QN
	SA3	no inaccurate annotations, labellings or classifications	$1 - \frac{\text{inaccurate instances}}{\text{total no. of instances}} * \frac{\text{balanced distance metric}}{\text{total no. of instances}}$ [Lei et al., 2007b]	QN
	SA4	no misuse of properties	by using profiling statistics, which support the detection of discordant values or misused properties and facilitate to find valid formats for specific properties [Böhm et al., 2010]	QN
	SA5	detection of valid rules	ratio of the number of semantically valid rules to the number of nontrivial rules [Chen and Garcia, 2010]	QN
Consistency	CS1	no use of entities as members of disjoint classes	$\frac{\text{no. of entities described as members of disjoint classes}}{\text{total no. of entities described in the dataset}}$ [Flemming, 2011, Hogan et al., 2010, Kontokostas et al., 2014]	QN
	CS2	no misplaced classes or properties	using entailment rules that indicate the position of a term in a triple [Feeney et al., 2014, Hogan et al., 2010]	QN
	CS3	no misuse of `owl:DatatypeProperty` or `owl:ObjectProperty`	detection of misuse of `owl:DatatypeProperty` or `owl:ObjectProperty` through the ontology maintainer [Hogan et al., 2010]	QN
	CS4	members of `owl:DeprecatedClass` or `owl:DeprecatedProperty` not used	detection of use of members of `owl:DeprecatedClass` or `owl:DeprecatedProperty` through the ontology maintainer or by specifying manual mappings from deprecated terms to compatible terms [Feeney et al., 2014, Hogan et al., 2010]	QN
	CS5	valid usage of inverse-functional properties	(i) by checking the uniqueness and validity of the inverse-functional values [Hogan et al., 2010], (ii) by defining a SPARQL query as a constraint [Kontokostas et al., 2014]	QN
	CS6	absence of ontology hijacking	detection of the re-definition by third parties of external classes/properties such that reasoning over data using those external terms is affected [Hogan et al., 2010]	QN
	CS7	no negative dependencies/correlation among properties	using association rules [Böhm et al., 2010]	QN
	CS8	no inconsistencies in spatial data	through semantic and geometric constraints [Mostafavi et al., 2004]	QN
	CS9	correct domain and range definition	the attribution of a resource's property (with a certain value) is only valid if the resource (domain), value (range) or literal value (rdfs ranged) is of a certain type - detected by use of SPARQL queries as a constraint [Kontokostas et al., 2014]	QN
	CS10	no inconsistent values	detection by the generation of a particular set of schema axioms for all properties in a dataset and the manual verification of these axioms [Zaveri et al., 2013a]	QN

Table 3.4.: Data quality metrics related to intrinsic dimensions (type QN refers to a quantitative metric, QL to a qualitative one).

Dimension	Abr	Metric	Description	Type
Conciseness	CN1	high intensional conciseness	$\frac{\text{no. of unique properties/classes of a dataset}}{\text{total no. of properties/classes in a target schema}}$ [Mendes et al., 2012b]	QN
	CN2	high extensional conciseness	(i) $\frac{\text{no. of unique objects of a dataset}}{\text{total number of objects representations in the dataset}}$ [Mendes et al., 2012b], (ii) $1 - \frac{\text{total no. of instances that violate the uniqueness rule}}{\text{total no. of relevant instances}}$ [Fürber and Hepp, 2011, Kontokostas et al., 2014, Lei et al., 2007b]	QN
	CN3	usage of unambiguous annotations/labels	$1 - \frac{\text{no. of ambiguous instances}}{\text{no. of instances contained in the semantic metadata set}}$ [Lei et al., 2007b, Ruckhaus et al., 2014]	QN
Completeness	CM1	schema completeness	$\frac{\text{no. of classes and properties represented}}{\text{total no. of classes and properties}}$ [Fürber and Hepp, 2011, Mendes et al., 2012b]	QN
	CM2	property completeness	(i) $\frac{\text{no. of values represented for a specific property}}{\text{total no. of values for a specific property}}$ [Feeney et al., 2014, Fürber and Hepp, 2011], (ii) exploiting statistical distributions of properties and types to characterize the property and then detect completeness [Paulheim and Bizer, 2014]	QN
	CM3	population completeness	$\frac{\text{no. of real-world objects are represented}}{\text{total no. of real-world objects}}$ [Feeney et al., 2014, Fürber and Hepp, 2011, Mendes et al., 2012b]	QN
	CM4	interlinking completeness	(i) $\frac{\text{no. of instances in the dataset that are interlinked}}{\text{total no. of instances in a dataset}}$ [Guéret et al., 2012a, Ruckhaus et al., 2014], (ii) calculating percentage of mappable types in a datasets that have not yet been considered in the linksets when assuming an alignment among types [Albertoni and Perez, 2013]	QN

  - syntactic rules (type of characters allowed and/or the pattern of literal values) [Fürber and Hepp, 2011]

- SV3: detection of ill-typed literals, which do not abide by the lexical syntax for their respective datatype that can occur if a value is (i) malformed, (ii) is a member of an incompatible datatype [Feeney et al., 2014, Hogan et al., 2010]

Example. In our use case, let us assume that the ID of the flight between Paris and New York is A123 while in our search engine the same flight instance is represented as A231. Since this ID is included in one of the datasets, it is considered to be syntactically accurate since it is a valid ID (even though it is incorrect).

3.2.2.2. Semantic accuracy

Fürber et al. [Fürber and Hepp, 2011] classified accuracy into syntactic and semantic accuracy. He explained that values are semantically accurate when they represent the correct state of an object. Based on this definition, we also considered the problems of *spurious annotation* and *inaccurate annotation* (inaccurate labeling and inaccurate classification) identified in Lei et al. [Lei et al., 2007b] related to the semantic accuracy dimension. The other articles [Acosta et al., 2013, Bizer and Cyganiak, 2009, Böhm et al., 2010, Chen and Garcia, 2010, Feeney et al., 2014, Kontokostas et al., 2014, Paulheim and Bizer, 2014, Zaveri et al., 2013a] provide metrics for this dimension.

Definition 7 (Semantic accuracy)**.** *Semantic accuracy is defined as the degree to which data values correctly represent the real world facts.*

Metrics. The metrics identified for semantic accuracy are:

- SA1: detection of outliers by (i) using distance-based, deviation-based and distribution-based methods [Bizer and Cyganiak, 2009, Feeney et al., 2014], (ii) using the statistical distributions of a certain type to assess the statement's correctness [Paulheim and Bizer, 2014]
- SA2: detection of inaccurate values by (i) using functional dependencies[10] [Fürber and Hepp, 2011] between the values of two or more different properties [Fürber and Hepp, 2011], (ii) comparison between two literal values of a resource [Kontokostas et al., 2014], (iii) via crowdsourcing [Acosta et al., 2013, Zaveri et al., 2013a]
- SA3: detection of inaccurate annotations[11], labellings[12] or classifications[13] using the formula:
  $1 - \frac{\text{inaccurate instances}}{\text{total no. of instances}} * \frac{\text{balanced distance metric}}{\text{total no. of instances}}$[14] [Lei et al., 2007b]
- SA4: detection of misuse of properties[15] by using profiling statistics, which support the detection of discordant values or misused properties and facilitate to find valid values for specific properties [Böhm et al., 2010]
- SA5: ratio of the number of semantically valid rules [16] to the number of nontrivial rules[17] [Chen and Garcia, 2010]

Example. Let us assume that the ID of the flight between Paris and New York is A123, while in our search engine the same flight instance is represented as A231 (possibly manually introduced by a data acquisition error). In this case, the instance is semantically inaccurate since the flight ID does not represent its real-world state i.e. A123.

3.2.2.3. Consistency

Hogan et al. [Hogan et al., 2010] defined consistency as "no contradictions in the data". Another definition was given by Mendes et al. [Mendes et al., 2012b] that "a dataset is consistent if it is free of conflicting information". The other articles [Böhm et al., 2010, Feeney et al., 2014, Flemming, 2011, Hogan et al., 2010, Kontokostas et al.,

[10] Functional dependencies are dependencies between the values of two or more different properties.
[11] Where an instance of the semantic metadata set can be mapped back to more than one real world object or in other cases, where there is no object to be mapped back to an instance.
[12] Where mapping from the instance to the object is correct but not properly labeled.
[13] In which the knowledge of the source object has been correctly identified by not accurately classified.
[14] Balanced distance metric is an algorithm that calculates the distance between the extracted (or learned) concept and the target concept [Maynard et al., 2006].
[15] Properties are often misused when no applicable property exists.
[16] Valid rules are generated from the real data and validated against a set of principles specified in the semantic network.
[17] The intuition is that the larger a dataset is, the more closely it should reflect the basic domain principles and the semantically incorrect rules will be generated.

2014, Mostafavi et al., 2004, Zaveri et al., 2013a] provide metrics for this dimension. However, it should be noted that for some languages such as OWL DL, there are clearly defined semantics, including clear definitions of what inconsistency means. In description logics, model based semantics are used: A knowledge base is a set of axioms. A model is an interpretation, which satisfies all axioms in the knowledge base. A knowledge base is consistent if and only if it has a model [Baader et al., 2003].

Definition 8 (Consistency)**.** *Consistency means that a knowledge base is free of (logical/formal) contradictions with respect to particular knowledge representation and inference mechanisms.*

Metrics. A straightforward way to check for consistency is to load the knowledge base into a reasoner and check whether it is consistent. However, for certain knowledge bases (e.g. very large or inherently inconsistent ones) this approach is not feasible. Moreover, most OWL reasoners specialize in the OWL (2) DL sublanguage as they are internally based on description logics. However, it should be noted that Linked Data does not necessarily conform to OWL DL and, therefore, those reasoners cannot directly be applied. Some of the important metrics identified in the literature are:

- CS1: detection of use of entities as members of disjoint classes using the formula: $\frac{\text{no. of entities described as members of disjoint classes}}{\text{total no. of entities described in the dataset}}$ [Flemming, 2011, Hogan et al., 2010, Kontokostas et al., 2014]
- CS2: detection of misplaced classes or properties[18] using entailment rules that indicate the position of a term in a triple [Feeney et al., 2014, Hogan et al., 2010]
- CS3: detection of misuse of `owl:DatatypeProperty` or `owl:ObjectProperty` through the ontology maintainer[19] [Hogan et al., 2010]
- CS4: detection of use of members of `owl:DeprecatedClass` or `owl:DeprecatedProperty` through the ontology maintainer or by specifying manual mappings from deprecated terms to compatible terms [Feeney et al., 2014, Hogan et al., 2010]
- CS5: detection of bogus `owl:InverseFunctionalProperty` values (i) by checking the uniqueness and validity of the inverse-functional values [Hogan et al., 2010], (ii) by defining a SPARQL query as a constraint [Kontokostas et al., 2014]
- CS6: detection of the re-definition by third parties of external classes/properties (ontology hijacking) such that reasoning over data using those external terms is not affected [Hogan et al., 2010]
- CS7: detection of negative dependencies/correlation among properties using association rules [Böhm et al., 2010]

[18] For example, a URI defined as a class is used as a property or vice-a-versa.
[19] For example, attribute properties used between two resources and relation properties used with literal values.

- CS8: detection of inconsistencies in spatial data through semantic and geometric constraints [Mostafavi et al., 2004]
- CS9: the attribution of a resource's property (with a certain value) is only valid if the resource (domain), value (range) or literal value (rdfs ranged) is of a certain type - detected by use of SPARQL queries as a constraint [Kontokostas et al., 2014]
- CS10: detection of inconsistent values by the generation of a particular set of schema axioms for all properties in a dataset and the manual verification of these axioms [Zaveri et al., 2013a]

Example. Let us assume a user looking for flights between Paris and New York on the 21st of December, 2013. Her query returns the following results:

Flight	From	To	Arrival	Departure
A123	Paris	NewYork	14:50	22:35
A123	Paris	London	14:50	22:35

The results show that the flight number A123 has two different destinations[20] at the same date and same time of arrival and departure, which is inconsistent with the ontology definition that one flight can only have one destination at a specific time and date. This contradiction arises due to inconsistency in data representation, which is detected by using inference and reasoning.

3.2.2.4. Conciseness

Mendes et al. [Mendes et al., 2012b] classified conciseness into schema and instance level conciseness. On the schema level (intensional), "a dataset is concise if it does not contain redundant attributes (two equivalent attributes with different names)". Thus, intensional conciseness measures the number of unique schema elements (i.e. properties and classes) of a dataset in relation to the overall number of schema elements in a schema. On the data (instance) level (extensional), "a dataset is concise if it does not contain redundant objects (two equivalent objects with different identifiers)". Thus, extensional conciseness measures the number of unique objects in relation to the overall number of objects in the dataset. This definition of conciseness is very similar to the definition of 'uniqueness' defined by Fürber et al. [Fürber and Hepp, 2011] as the "degree to which data is free of redundancies, in breadth, depth and scope". This comparison shows that uniqueness and conciseness point to the same dimension. Redundancy occurs when there are *equivalent* schema elements with different names/identifiers (in case of intensional conciseness) and when there are *equivalent* objects (instances) with different identifiers (in case of extensional conciseness) present in a dataset [Lei et al., 2007b]. Kontokostas et al.[Kontokostas et al., 2014] provide metrics for this dimension.

Definition 9 (Conciseness)**.** *Conciseness refers to the minimization of redundancy of entities at the schema and the data level. Conciseness is classified into (i) intensional*

[20] Under the assumption that we can infer that NewYork and London are different entities or, alternatively, make the unique name assumption.

conciseness (schema level) which refers to the case when the data does not contain redundant schema elements (properties and classes) and (ii) extensional conciseness (data level) which refers to the case when the data does not contain redundant objects (instances).

Metrics. The metrics identified for conciseness are:

- CN1: intensional conciseness measured by $\frac{\text{no. of unique properties/classes of a dataset}}{\text{total no. of properties/classes in a target schema}}$ [Mendes et al., 2012b]
- CN2: extensional conciseness measured by:
  - $\frac{\text{no. of unique instances of a dataset}}{\text{total number of instances representations in the dataset}}$ [Mendes et al., 2012b],
  - $1 - \frac{\text{total no. of instances that violate the uniqueness rule}}{\text{total no. of relevant instances}}$ [Fürber and Hepp, 2011, Kontokostas et al., 2014, Lei et al., 2007b]
- CN3: detection of unambiguous annotations using the formula: $1 - \frac{\text{no. of ambiguous instances}}{\text{no. of instances contained in the semantic metadata set}}$[21] [Lei et al., 2007b, Ruckhaus et al., 2014]

Example. In our flight search engine, an example of intensional conciseness would be a particular flight, say `A123`, being represented by two different properties in the same dataset, such as `http://flights.org/airlineID` and `http://flights.org/name`. This redundancy ('airlineID' and 'name' in this case) can ideally be solved by fusing the two properties and keeping only one unique identifier. On the other hand, an example of extensional conciseness is when both these identifiers of the same flight have the same information associated with them in both the datasets, thus duplicating the information.

3.2.2.5. Completeness

Fürber et al. [Fürber and Hepp, 2011] classified completeness into: (i) Schema completeness, which is the degree to which classes and properties are not missing in a schema; (ii) Column completeness, which is a function of the missing property values for a specific property/column; and (iii) Population completeness, which refers to the ratio between classes represented in an information system and the complete population. Mendes et al. [Mendes et al., 2012b] distinguished completeness on the schema and the data level. On the schema level, a dataset is complete if it contains all of the attributes needed for a given task. On the data (i.e. instance) level, a dataset is complete if it contains all of the necessary objects for a given task. The two types of completeness defined in Mendes et al. can be mapped to the two categories (i) Schema completeness and (iii) Population completeness provided by Fürber et al. Additionally, we introduce the category interlinking completeness, which refers to the degree to which instances in

[21] Detection of an instance mapped back to more than one real world object leading to more than one interpretation.

the dataset are interlinked [Guéret et al., 2012a]. Albertoni et al. [Albertoni and Perez, 2013] define interlinking completeness as "linkset completeness as the degree to which links in the linksets are not missing." The other articles [Feeney et al., 2014, Paulheim and Bizer, 2014, Ruckhaus et al., 2014] provide metrics for this dimension.

Definition 10 (Completeness)**.** *Completeness refers to the degree to which all required information is present in a particular dataset. In terms of LD, completeness comprises of the following aspects: (i) Schema completeness, the degree to which the classes and properties of an ontology are represented, thus can be called "ontology completeness", (ii) Property completeness, measure of the missing values for a specific property, (iii) Population completeness is the percentage of all real-world objects of a particular type that are represented in the datasets and (iv) Interlinking completeness, which has to be considered especially in LD, refers to the degree to which instances in the dataset are interlinked.*

Metrics. The metrics identified for completeness are:

- CM1: schema completeness $\frac{\text{no. of classes and properties represented}}{\text{total no. of classes and properties}}$ [Fürber and Hepp, 2011, Mendes et al., 2012b]
- CM2: property completeness (i) $\frac{\text{no. of values represented for a specific property}}{\text{total no. of values for a specific property}}$ [Feeney et al., 2014, Fürber and Hepp, 2011], (ii) exploiting statistical distributions of properties and types to characterize the property and then detect completeness [Paulheim and Bizer, 2014]
- CM3: population completeness $\frac{\text{no. of real-world objects are represented}}{\text{total no. of real-world objects}}$ [Feeney et al., 2014, Fürber and Hepp, 2011, Mendes et al., 2012b]
- CM4: interlinking completeness
  (i) $\frac{\text{no. of instances in the dataset that are interlinked}}{\text{total no. of instances in a dataset}}$ [Guéret et al., 2012a, Ruckhaus et al., 2014], (ii) calculating percentage of mappable types in a datasets that have not yet been considered in the linksets when assuming an alignment among types [Albertoni and Perez, 2013]

It should be noted that in this case, users should assume a closed-world-assumption where a gold standard dataset is available and can be used to compare against the converted dataset.

Example. In our use case, the flight search engine contains complete information to include all the airports and airport codes such that it allows a user to find an optimal route from the start to the end destination (even in cases when there is no direct flight). For example, the user wants to travel from Santa Barbara to San Francisco. Since our flight search engine contains interlinks between these close airports, the user is able to locate a direct flight easily.

3.2.2.6. Intra-relations

The dimensions in this group are related to each other as follows: Data can be semantically accurate by representing the real world state but still can be inconsistent. However, if we merge accurate datasets, we will most likely get fewer inconsistencies than merging inaccurate datasets. On the other hand, being syntactically valid does not necessarily mean that the value is semantically accurate. Moreover, if a dataset is complete, tests for syntactic validity, semantic accuracy and consistency checks need to be performed to determine if the values have been completed correctly. Additionally, the conciseness dimension is related to the completeness dimension since both point towards the dataset having all, however unique (non-redundant) information. However, if data integration leads to duplication of instances, it may lead to contradictory values thus leading to inconsistency [Bleiholder and Naumann, 2008].

3.2.3. Contextual dimensions

Contextual dimensions are those that highly depend on the context of the task at hand. There are four dimensions that are part of this group, namely *relevancy, trustworthiness, understandability* and *timeliness*. These dimensions along with their corresponding metrics and references to the original literature are presented in Table 3.5 and Table 3.6.

3.2.3.1. Relevancy

Flemming [Flemming, 2011] defined amount-of-data as the "criterion influencing the usability of a data source". Thus, since the amount-of-data dimension is similar to the relevancy dimension, we merge both dimensions. Bonatti et al. [Bonatti et al., 2011] provides a metric for this dimension. The other articles [Acosta et al., 2013, Zaveri et al., 2013a] provide metrics for this dimension.

Definition 11 (Relevancy)**.** *Relevancy refers to the provision of information which is in accordance with the task at hand and important to the users' query.*

Metrics. The metrics identified for relevancy are:

- R1: obtaining relevant data by: (i) ranking (a numerical value similar to PageRank), which determines the centrality of RDF documents and statements [Bonatti et al., 2011]), (ii) via crowdsourcing [Acosta et al., 2013, Zaveri et al., 2013a]
- R2: measuring the coverage (i.e. number of entities described in a dataset) and level of detail (i.e. number of properties) in a dataset to ensure that the data retrieved is appropriate for the task at hand [Flemming, 2011]

Example. When a user is looking for flights between any two cities, only relevant information i.e. departure and arrival airports, starting and ending time, duration and cost per person should be provided. Some datasets, in addition to relevant information, also contain much irrelevant data such as car rental, hotel booking, travel insurance etc. and

Table 3.5.: Data quality metrics related to contextual dimensions (type QN refers to a quantitative metric, QL to a qualitative one).

Dimension	Abr	Metric	Description	Type
Relevancy	R1	relevant terms within meta-information attributes	obtaining relevant data by (i) ranking (a numerical value similar to PageRank), which determines the centrality of RDF documents and statements [Bonatti et al., 2011], (ii) via crowdsourcing [Acosta et al., 2013, Zaveri et al., 2013a]	QN
	R2	coverage	measuring the coverage (i.e. number of entities described in a dataset) and level of detail (i.e. number of properties) in a dataset to ensure that the data retrieved is appropriate for the task at hand [Flemming, 2011]	QN
Trustworthiness	T1	trustworthiness of statements	computing statement trust values based on: (i) provenance information which can be either unknown or a value in the interval [−1,1] where 1: absolute belief, −1: absolute disbelief and 0: lack of belief/disbelief [Feeney et al., 2014, Hartig, 2008] (ii) opinion-based method, which use trust annotations made by several individuals [Gil and Ratnakar, 2002, Hartig, 2008] (iii) provenance information and trust annotations in Semantic Web-based social-networks [Golbeck, 2006] (iv) annotating triples with provenance data and usage of provenance history to evaluate the trustworthiness of facts [Dezani-Ciancaglini et al., 2012]	QN
	T2	trustworthiness through reasoning	using annotations for data to encode two facets of information [Bonatti et al., 2011]: (i) blacklists (indicates that the referent data is known to be harmful) (ii) authority (a boolean value which uses the Linked Data principles to conservatively determine whether or not information can be trusted)	QN
	T3	trustworthiness of statements, datasets and rules	using trust ontologies that assigns trust values that can be transferred from known to unknown data using: (i) content-based methods (from content or rules) and (ii) metadata-based methods (based on reputation assignments, user ratings, and provenance, rather than the content itself) [Jacobi et al., 2011]	QN
	T4	trustworthiness of a resource	computing trust values between two entities through a path by using: (i) a propagation algorithm based on statistical techniques (ii) in case there are several paths, trust values from all paths are aggregated based on a weighting mechanism [Shekarpour and Katebi, 2010]	QN
	T5	trustworthiness of the information provider	computing trustworthiness of the information provider by: (i) construction of decision networks informed by provenance graphs [Gamble and Goble, 2011] (ii) checking whether the provider/contributor is contained in a list of trusted providers [Bizer and Cyganiak, 2009]	QN
			(iii) indicating the level of trust for the publisher on a scale of 1−9 [Gil and Artz, 2007, Golbeck et al., 2003]	QL
	T6	trustworthiness of information provided (content trust)	checking content trust based on associations (e.g. anything having a relationship to a resource such as author of the dataset) that transfers trust from content to resources [Gil and Artz, 2007]	QL
	T7	reputation of the dataset	assignment of explicit trust ratings to the dataset by humans or analyzing external links or page ranks [Mendes et al., 2012b]	QL
Understandability	U1	human-readable labelling of classes, properties and entities as well as presence of metadata	detection of human-readable labelling of classes, properties and entities as well as indication of metadata (e.g. name, description, website) of a dataset [Feeney et al., 2014, Flemming, 2011, Hogan et al., 2012]	QN
	U2	indication of one or more exemplary URIs	detect whether the pattern of the URIs is provided [Flemming, 2011]	QN
	U3	indication of a regular expression that matches the URIs of a dataset	detect whether a regular expression that matches the URIs is present [Flemming, 2011]	QN
	U4	indication of an exemplary SPARQL query	detect whether examples of SPARQL queries are provided [Flemming, 2011]	QN
	U5	indication of the vocabularies used in the dataset	checking whether a list of vocabularies used in the dataset is provided [Flemming, 2011]	QN
	U6	provision of message boards and mailing lists	checking the effectiveness and the efficiency of the usage of the mailing list and/or the message boards [Flemming, 2011]	QL

Table 3.6.: Data quality metrics related to contextual dimensions (continued) (type QN refers to a quantitative metric, QL to a qualitative one).

Dimension	Abr	Metric	Description	Type
Timeliness	TI1	freshness of datasets based on currency and volatility	$\max\{0, 1 - \frac{currency}{volatility}\}$ [Hartig and Zhao, 2009], which gives a value in a continuous scale from 0 to 1, where score of 1 implies that the data is timely and 0 means it is completely outdated thus unacceptable. In the formula, volatility is the length of time the data remains valid [Fürber and Hepp, 2011] and currency is the age of the data when delivered to the user [Feeney et al., 2014, Mendes et al., 2012b, Rula et al., 2012]	QN
	TI2	freshness of datasets based on their data source	detecting freshness of datasets based on their data source by measuring the distance between last modified time of the data source and last modified time of the dataset [Fürber and Hepp, 2011, Mendes et al., 2012a]	QN

as a consequence a lot of irrelevant extra information is provided. Providing irrelevant data distracts service developers and potentially users and also wastes network resources. Instead, restricting the dataset to only flight related information simplifies application development and increases the likelihood to return only relevant results to users.

3.2.3.2. Trustworthiness

Trustworthiness is a crucial topic due to the availability and the high volume of data from varying sources on the Web of Data. Jacobi et al. [Jacobi et al., 2011], similar to Pipino et al., referred to trustworthiness as a subjective measure of a user's belief that the data is "true". Gil et al. [Gil and Artz, 2007] used reputation of an entity or a dataset either as a result from direct experience or recommendations from others to establish trust. Ciancaglini et al. [Dezani-Ciancaglini et al., 2012] state "the degree of trustworthiness of the triple will depend on the trustworthiness of the individuals involved in producing the triple and the judgement of the consumer of the triple." We consider reputation as well as objectivity as part of the trustworthiness dimension. Other articles [Bonatti et al., 2011, Dezani-Ciancaglini et al., 2012, Feeney et al., 2014, Gamble and Goble, 2011, Gil and Ratnakar, 2002, Golbeck, 2006, Golbeck et al., 2003, Hartig, 2008, Mendes et al., 2012b, Shekarpour and Katebi, 2010] provide metrics for assessing trustworthiness.

Definition 12 (Trustworthiness)**.** *Trustworthiness is defined as the degree to which the information is accepted to be correct, true, real and credible.*

Metrics. The metrics identified for trustworthiness are:

- T1: computing statement trust values based on:
  - provenance information which can be either unknown or a value in the interval [−1,1] where 1: absolute belief, −1: absolute disbelief and 0: lack of belief/disbelief [Feeney et al., 2014, Hartig, 2008]

  - opinion-based method, which use trust annotations made by several individuals [Gil and Ratnakar, 2002, Hartig, 2008]
  - provenance information and trust annotations in Semantic Web-based social-networks [Golbeck, 2006]
  - annotating triples with provenance data and usage of provenance history to evaluate the trustworthiness of facts [Dezani-Ciancaglini et al., 2012]

- T2: using annotations for data to encode two facets of information:
  - blacklists (indicates that the referent data is known to be harmful) [Bonatti et al., 2011] and
  - authority (a boolean value which uses the Linked Data principles to conservatively determine whether or not information can be trusted) [Bonatti et al., 2011]

- T3: using trust ontologies that assigns trust values that can be transferred from known to unknown data [Jacobi et al., 2011] using:
  - content-based methods (from content or rules) and
  - metadata-based methods (based on reputation assignments, user ratings, and provenance, rather than the content itself)

- T4: computing trust values between two entities through a path by using:
  - a propagation algorithm based on statistical techniques [Shekarpour and Katebi, 2010]
  - in case there are several paths, trust values from all paths are aggregated based on a weighting mechanism [Shekarpour and Katebi, 2010]

- T5: computing trustworthiness of the information provider by:
  - construction of decision networks informed by provenance graphs [Gamble and Goble, 2011]
  - checking whether the provider/contributor is contained in a list of trusted providers [Bizer and Cyganiak, 2009]
  - indicating the level of trust for the publisher on a scale of $1 - 9$ [Gil and Artz, 2007, Golbeck et al., 2003]

- T6: checking content trust[22] based on associations (e.g. anything having a relationship to a resource such as author of the dataset) that transfer trust from content to resources [Gil and Artz, 2007]

- T7: assignment of explicit trust ratings to the dataset by humans or analyzing external links or page ranks [Mendes et al., 2012b]

[22] Content trust is a trust judgement on a particular piece of information in a given context [Gil and Artz, 2007].

Example. In our flight search engine use case, if the flight information is provided by trusted and well-known airlines then a user is more likely to trust this information than when an unknown travel agency provides it. Generally information about a product or service (e.g. a flight) can be trusted when it is directly published by the producer or service provider (e.g. the airline). On the other hand, if a user retrieves information from a previously unknown source, she can decide whether to believe this information by checking whether the source is well-known or if it is contained in a list of trusted providers.

3.2.3.3. Understandability

Flemming [Flemming, 2011] related understandability to the comprehensibility of data i.e. the ease with which human consumers can understand and utilize the data. Thus, comprehensibility can be interchangeably used with understandability. Hogan et al. [Hogan et al., 2012] specified the importance of providing human-readable metadata "for allowing users to visualize, browse and understand RDF data, where providing labels and descriptions establishes a baseline". Feeney et al. [Feeney et al., 2014] provide a metric for this dimension.

Definition 13 (Understandability)**.** *Understandability refers to the ease with which data can be comprehended without ambiguity and be used by a human information consumer.*

Metrics. The metrics identified for understandability are:

- U1: detection of human-readable labelling of classes, properties and entities as well as indication of metadata (e.g. name, description, website) of a dataset [Feeney et al., 2014, Flemming, 2011, Hogan et al., 2012]
- U2: detect whether the pattern of the URIs is provided [Flemming, 2011]
- U3: detect whether a regular expression that matches the URIs is present [Flemming, 2011]
- U4: detect whether examples of SPARQL queries are provided [Flemming, 2011]
- U5: checking whether a list of vocabularies used in the dataset is provided [Flemming, 2011]
- U6: checking the effectiveness and the efficiency of the usage of the mailing list and/or the message boards [Flemming, 2011]

Example. Let us assume that a user wants to search for flights between Boston and San Francisco using our flight search engine. From the data related to Boston in the integrated dataset for the required flight, the following URIs and a label is retrieved:

- `http://rdf.freebase.com/ns/m.049jnng`
- `http://rdf.freebase.com/ns/m.043j22x`

- "Boston Logan Airport"@en

For the first two items no human-readable label is available, therefore the machine is only able to display the URI as a result of the users query. This does not represent anything meaningful to the user besides perhaps that the information is from Freebase. The third entity, however, contains a human-readable label, which the user can easily understand.

3.2.3.4. Timeliness

Gamble et al. [Gamble and Goble, 2011] defined timeliness as "a comparison of the date the annotation was updated with the consumer's requirement". The timeliness dimension is motivated by the fact that it is possible to have current data that is actually incompetent because it reflects a state of the real world that is too old for a specific usage. According to the timeliness dimension, data should ideally be recorded and reported as frequently as the source values change and thus never become outdated. Other articles [Feeney et al., 2014, Fürber and Hepp, 2011, Hartig and Zhao, 2009, Mendes et al., 2012b, Rula et al., 2012] provide metrics for assessing timeliness.

Definition 14. *Timeliness measures how up-to-date data is relative to a specific task.*

Metrics. The metrics identified for timeliness are:

- TI1: detecting freshness of datasets based on currency and volatility using the formula:
$$\max\{0, 1 - \frac{currency}{volatility}\}$$
[Hartig and Zhao, 2009], which gives a value in a continuous scale from 0 to 1, where a score of 1 implies that the data is timely and 0 means it is completely outdated and thus unacceptable. In the formula, currency is the age of the data when delivered to the user [Feeney et al., 2014, Mendes et al., 2012b, Rula et al., 2012] and volatility is the length of time the data remains valid [Fürber and Hepp, 2011]

- TI2: detecting freshness of datasets based on their data source by measuring the distance between the last modified time of the data source and last modified time of the dataset [Fürber and Hepp, 2011]

Example. Consider a user checking the flight timetable for her flight from city A to city B. Suppose that the result is a list of triples comprising of the description of the resource A such as the connecting airports, the time of departure and arrival, the terminal, the gate, etc. This flight timetable is updated every 10 minutes (volatility). Assume there is a change of the flight departure time, specifically a delay of one hour. However, this information is communicated to the control room with a slight delay. They update this information in the system after 30 minutes. Thus, the timeliness constraint of updating the timetable within 10 minutes is not satisfied which renders the information out-of-date.

Table 3.7.: Data quality metrics related to representational dimensions (type QN refers to a quantitative metric, QL to a qualitative one).

Dimension	Abr	Metric	Description	Type
Representational-conciseness	RC1	keeping URIs short	detection of long URIs or those that contain query parameters [Feeney et al., 2014, Hogan et al., 2012]	QN
	RC2	no use of prolix RDF features	detection of RDF primitives i.e. RDF reification, RDF containers and RDF collections [Feeney et al., 2014, Hogan et al., 2012]	QN
Interoperability	IO1	re-use of existing terms	detection of whether existing terms from all relevant vocabularies for that particular domain have been reused [Hogan et al., 2012]	QL
	IO2	re-use of existing vocabularies	usage of relevant vocabularies for that particular domain [Flemming, 2011]	QL
Interpretability	IN1	use of self-descriptive formats	identifying objects and terms used to define these objects with globally unique identifiers [Feeney et al., 2014]	QN
	IN2	detecting the interpretability of data	detecting the use of appropriate language, symbols, units, datatypes and clear definitions [Flemming, 2011, Pipino et al., 2002]	QL
	IN3	invalid usage of undefined classes and properties	detection of invalid usage of undefined classes and properties (i.e. those without any formal definition) [Hogan et al., 2010]	QN
	IN4	no misinterpretation of missing values	detecting the use of blank nodes [Hogan et al., 2012]	QN
Versatility	V1	provision of the data in different serialization formats	checking whether data is available in different serialization formats [Flemming, 2011]	QN
	V2	provision of the data in various languages	checking whether data is available in different languages [Auer et al., 2010, Flemming, 2011, Labra Gayo et al., 2012]	QN

3.2.3.5. Intra-relations

The dimensions in this group are related to each other as follows: Data is of high relevance if data is current for the user needs. The timeliness of information thus influences its relevancy. On the other hand, if a dataset has current information, it is considered to be trustworthy. Moreover, to allow users to properly understand information in a dataset, a system should be able to provide sufficient relevant information.

3.2.4. Representational dimensions

Representational dimensions capture aspects related to the design of the data such as the *representational-conciseness, interoperability, interpretability* as well as *versatility*. Table 3.7 displays metrics for these four dimensions along with references to the original literature.

3.2.4.1. Representational-conciseness

Hogan et al. [Hogan et al., 2010, Hogan et al., 2012] provide benefits of using shorter URI strings for large-scale and/or frequent processing of RDF data thus encouraging the use of concise representation of the data. Moreover, they emphasized that the use of RDF reification should be avoided "as the semantics of reification are unclear and as reified statements are rather cumbersome to query with the SPARQL query language".

Definition 15 (Representational-conciseness)**.** *Representational-conciseness refers to*

the representation of the data, which is compact and well formatted on the one hand and clear and complete on the other hand.

Metrics. The metrics identified for representational-conciseness are:

- RC1: detection of long URIs or those that contain query parameters [Feeney et al., 2014, Hogan et al., 2012]
- RC2: detection of RDF primitives i.e. RDF reification, RDF containers and RDF collections [Feeney et al., 2014, Hogan et al., 2012]

Example. Our flight search engine represents the URIs for the destination compactly with the use of the airport codes. For example, LEJ is the airport code for Leipzig, therefore the URI is `http://airlines.org/LEJ`. Such short representation of the URIs helps users share and memorize them easily.

3.2.4.2. Interoperability

Hogan et al. [Hogan et al., 2012] state that the re-use of well-known terms to describe resources in a uniform manner increases the interoperability of data published in this manner and contributes towards the interoperability of the entire dataset. The definition of "uniformity", which refers to the re-use of established formats to represent data as described by Flemming [Flemming, 2011], is also associated to the interoperability of the dataset.

Definition 16 (Interoperability)**.** *Interoperability is the degree to which the format and structure of the information conforms to previously returned information as well as data from other sources.*

Metrics. The metrics identified for interoperability are:

- IO1: detection of whether existing terms from all relevant vocabularies for that particular domain have been reused [Hogan et al., 2012]
- IO2: usage of relevant vocabularies for that particular domain [Flemming, 2011]

Example. Let us consider different airline datasets using different notations for representing the geo-cordinates of a particular flight location. While one dataset uses the WGS 84 geodetic system, another one uses the GeoRSS points system to specify the location. This makes querying the integrated dataset difficult, as it requires users and the machines to understand the heterogeneous schema. Additionally, with the difference in the vocabularies used to represent the same concept (in this case the co-ordinates), consumers are faced with the problem of how the data can be interpreted and displayed.

3.2.4.3. Interpretability

Hogan et al. [Hogan et al., 2010, Hogan et al., 2012] specify that the ad-hoc definition of classes and properties as well use of blank nodes makes the automatic integration of data less effective and forgoes the possibility of making inferences through reasoning. Thus, these features should be avoided in order to make the data much more interpretable. The other articles [Feeney et al., 2014, Flemming, 2011] provide metrics for this dimension.

Definition 17 (Interpretability)**.** *Interpretability refers to technical aspects of the data, that is, whether information is represented using an appropriate notation and whether the machine is able to process the data.*

Metrics. The metrics identified for interpretability are:

- IN1: identifying objects and terms used to define these objects with globally unique identifiers [Feeney et al., 2014]
- IN2: detecting the use of appropriate language, symbols, units, datatypes and clear definitions [Flemming, 2011]
- IN3: detection of invalid usage of undefined classes and properties (i.e. those without any formal definition) [Hogan et al., 2010]
- IN4: detecting the use of blank nodes[23] [Hogan et al., 2012]

Example. Consider our flight search engine and a user that is looking for a flight from `Mumbai` to `Boston` with a two day stop-over in `Berlin`. The user specifies the dates correctly. However, since the flights are operated by different airlines, thus different datasets, they have a different way of representing the date. In the first leg of the trip, the date is represented in the format `dd/mm/yyyy` whereas in the other case, the date is represented as `mm/dd/yy`. Thus, the machine is unable to correctly interpret the data and cannot provide an optimal result for this query. This lack of consensus in the format of the date hinders the ability of the machine to interpret the data and thus provide the appropriate flights.

3.2.4.4. Versatility

Flemming [Flemming, 2011] defined versatility as the "alternative representations of the data and its handling."

Definition 18 (Versatility)**.** *Versatility refers to the availability of the data in different representations and in an internationalized way.*

Metrics. The metrics identified for versatility are:

[23] Blank nodes are not recommended since they cannot be externally referenced.

- V1: checking whether data is available in different serialization formats [Flemming, 2011]
- V2: checking whether data is available in different languages [Flemming, 2011]

Example. Consider a user who does not understand English but only Chinese and wants to use our flight search engine. In order to cater to the needs of such a user, the dataset should provide labels and other language-dependent information in Chinese so that any user has the capability to understand it.

3.2.4.5. Intra-relations

The dimensions in this group are related as follows: Interpretability is related to the interoperability of data since the consistent representation (e.g. re-use of established vocabularies) ensures that a system will be able to interpret the data correctly [Ding and Finin, 2006]. Versatility is also related to the interpretability of a dataset as the more different forms a dataset is represented in (e.g. in different languages), the more interpretable a dataset is. Additionally, concise representation of the data allows the data to be interpreted correctly.

3.2.5. Inter-relationships between dimensions

The 18 data quality dimensions explained in the previous sections are not independent from each other but correlations exist among them. In this section, we describe the inter-relations between the 18 dimensions, as shown in Figure 3.1. If some dimensions are considered more important than others for a specific application (or use case), then favoring the more important ones will result in downplaying the influence of others. The inter-relationships help to identify which dimensions should possibly be considered together in a certain quality assessment application. Hence, investigating the relationships among dimensions is an interesting problem, as shown by the following examples.

First, relationships exist between the dimensions trustworthiness, semantic accuracy and timeliness. When assessing the trustworthiness of a LD dataset, the semantic accuracy and the timeliness of the dataset should be assessed. Frequently the assumption is made that a publisher with a high reputation will produce data that is also semantically accurate and current, when in reality this may not be so.

Second, relationships occur between timeliness and the semantic accuracy, completeness and consistency dimensions. On the one hand, having semantically accurate, complete or consistent data may require time and thus timeliness can be negatively affected. Conversely, having timely data may cause low accuracy, incompleteness and/or inconsistency. Based on quality preferences given by an application, a possible order of quality can be as follows: timely, consistent, accurate and then complete data. For instance, a list of courses published on a university website might be first of all timely, secondly consistent and accurate, and finally complete. Conversely, when considering an e-banking application, first of all it is preferred that data is accurate, consistent and

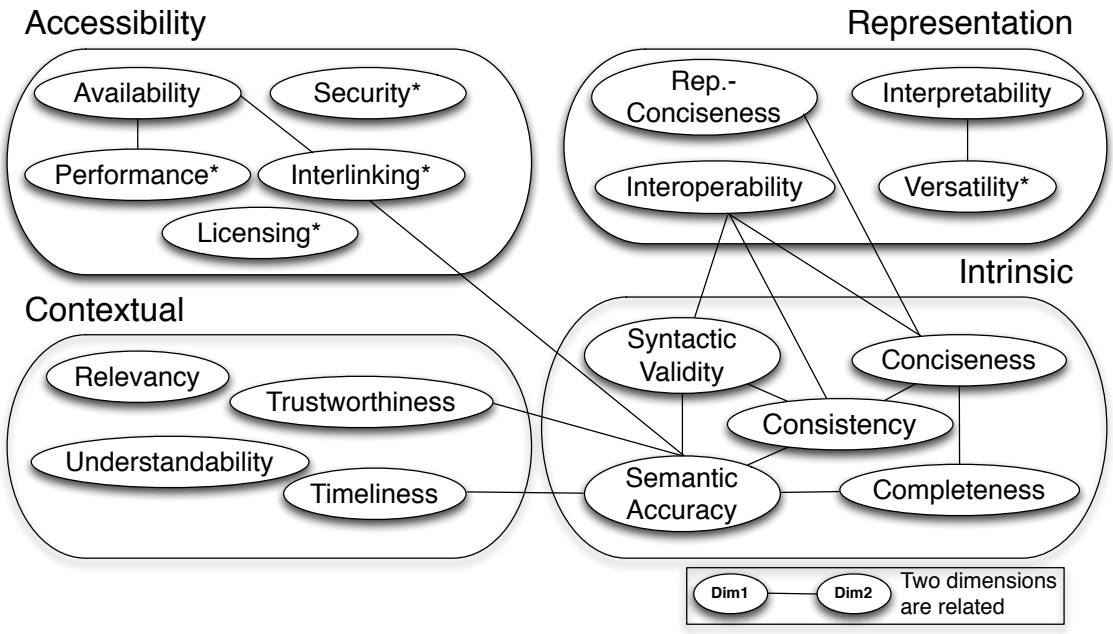


Figure 3.1.: Linked Data quality dimensions and the relations between them. The dimensions marked with '*' are specific for Linked Data.

complete as stringent requirements and only afterwards timely since delays are allowed in favour of correctness of data provided.

The representational-conciseness dimension (belonging to the representational group) and the conciseness dimension (belonging to the intrinsic group) are also closely related with each other. On the one hand, representational-conciseness refers to the conciseness of *representing* the data (e.g. short URIs) while conciseness refers to the compactness of the *data itself* (no redundant attributes and objects). Both dimensions thus point towards the compactness of the data. Moreover, representational-conciseness not only allows users to understand the data better but also provides efficient processing of frequently used RDF data (thus affecting performance). On the other hand, Hogan et al. [Hogan et al., 2012] associated performance to the issue of "using prolix RDF features" such as (i) reification, (ii) containers and (iii) collections. These features should be avoided as they are cumbersome to represent in triples and can prove to be expensive to support in data intensive environments.

Additionally, the interoperability dimension (belonging to the representational group) is inter-related with the consistency dimension (belonging to the intrinsic group), because the invalid re-usage of vocabularies (mandated by the interoperability dimension) may lead to inconsistency in the data. The versatility dimension, also part of the representational group, is related to the accessibility dimension since provision of data via different means (e.g. SPARQL endpoint, RDF dump) inadvertently points towards the different ways in which data can be accessed. Additionally, versatility (e.g. providing data in different languages) allows a user to understand the information better, thus also relates

to the understandability dimension. Furthermore, there exists an inter-relation between the conciseness and the relevancy dimensions. Conciseness frequently positively affects relevancy since removing redundancies increases the proportion of relevant data that can be retrieved.

The interlinking dimension is associated with the semantic accuracy dimension. It is important to choose the correct similarity relationship such as *same, matches, similar* or *related* between two entities to capture the most appropriate relationship [Halpin et al., 2010] thus contributing towards the semantic accuracy of the data. Additionally, interlinking is directly related to the interlinking completeness dimension. However, the interlinking dimension focuses on the quality of the interlinks whereas the interlinking completeness focus on the presence of *all* relevant interlinks in a dataset.

These sets of non-exhaustive examples of inter-relations between the dimensions belonging to different groups indicates the interplay between them and show that these dimensions are to be considered differently in different data quality assessment scenarios.

3.3. Summary

In this section, we provided a total of 18 quality dimensions and 69 metrics that can be applied for quality assessment of LD identified from the 30 core articles of our survey. In particular, we provided a definition and an example for each of the 18 dimensions. Additionally, different metrics were identified for each dimension and were furthermore classified into being qualitatively or quantitatively assessed. The 18 dimensions were classified into four groups and the intra as well as the inter relations between the dimensions were discussed.

Table 3.8.: Occurrences of the 18 data quality dimensions in each of the included approaches.

Article / Dimensions	Availability	Licensing	Interlinking	Security	Performance	Syntactic validity	Semantic accuracy	Consistency	Conciseness	Completeness	Relevancy	Trustworthiness	Understandability	Timeliness	Rep.-conciseness	Interoperability	Interpretability	Versatility
Bizer et al.,2009																		
Flemming,2010	✔	✔		✔	✔	✔							✔			✔		✔
Böhm et al.,2010							✔	✔										
Chen et al.,2010							✔											
Ciancaglini et al.,2012												✔						
Guéret et al,2012			✔							✔								
Hogan et al.,2010	✔					✔		✔										
Hogan et al.,2012	✔	✔	✔										✔		✔	✔	✔	
Lei et al.,2007							✔		✔									
Mendes et al., 2012								✔	✔	✔		✔		✔				
Mostafavi et al., 2004								✔										
Fürber et al.,2011						✔	✔		✔	✔				✔				
Rula et al., 2012														✔				
Hartig, 2008												✔						
Gamble et al., 2011												✔						
Shekarpour et al., 2010												✔						
Golbeck, 2006												✔						
Gil et al., 2002												✔						
Golbeck et al., 2003												✔						
Gil et al., 2007												✔						
Jacobi et al., 2011												✔						
Bonatti et al., 2011											✔	✔						
Acosta et al.,2013			✔			✔	✔											
Zaveri et al.,2013			✔			✔	✔				✔							
Albertoni et al.,2013										✔								
Feeney et al.,2014						✔	✔			✔		✔	✔	✔	✔		✔	
Kontokostas et al.,2014						✔	✔	✔	✔									
Paulheim et al.,2014										✔								
Ruckhaus et al.,2014									✔	✔								
Wienand et al.,2014							✔											

4. User-Driven Linked Data Quality Evaluation

On the Data Web, we have varying quality of information covering various domains. There are a large number of high quality datasets (in particular in the life-sciences domain), which are carefully curated over decades and recently published on the Web. There are, however, also many datasets, which were extracted from unstructured and semi-structured information or are the result of some crowdsourcing process, where large numbers of users contribute small parts. DBpedia [Auer and Lehmann, 2007, Lehmann et al., 2009] is actually an example for both - a dataset extracted from the result of a crowdsourcing process. Hence, we assess and present results of assessing the quality of DBpedia in this chapter, which is based on [Zaveri et al., 2013a].

Quality usually means fitness for a certain use case [Juran, 1974]. Hence, even datasets with quality problems might be useful for certain applications, as long as the quality is in the required range. In the case of DBpedia, for example, the data quality is perfectly sufficient for enriching Web search with facts or suggestions about common sense information, such as entertainment topics. In such a scenario, where the DBpedia background knowledge can be, for example, used to show the movies Angelina Jolie was starring in and actors she played with it is rather neglectable if, in relatively few cases, a movie or an actor is missing. For developing a medical application, on the other hand, the quality of DBpedia is probably completely insufficient. Please note, that also on the traditional document-oriented Web we have varying quality of the information and still the Web is perceived to be extremely useful by most people. Consequently, a key challenge is to determine the quality of datasets published on the Web and make this quality information explicit. Other than on the document Web where information quality can be only indirectly defined, we can have much more concrete and measurable data quality indicators for structured information, such as correctness of facts, adequacy of semantic representation or degree of coverage.

In this chapter, we devise a data quality assessment methodology, which comprises of a manual and a semi-automatic process. We empirically assess, based on this methodology, the data quality of one of the major knowledge hubs on the Data Web – DBpedia. In the *manual* process, the first phase includes the detection of common quality problems and their representation in a quality problem taxonomy. The second phase comprises of the evaluation of a large number of individual resources, according to the quality problem taxonomy, performed by users. Here we would like to clarify the use of crowdsource used in this study. Crowdsourcing involves the creating if Human Intelligent Tasks (HIT), submitting them to a crowdsourcing platform (e.g. Amazon Mechanical

Turk[1]) and providing a (financial) reward for each HIT [Howe, 2006]. However, we use the broader-sense of the word as a large-scale problem-solving approach by which a problem is divided into several smaller tasks (assessing the quality of each triple, in this case) that can be independently solved by a large group of people. Each represented fact is evaluated for correctness by each user and, if found problematic, annotated with one of 17 pre-defined quality criteria. This process is accompanied by a tool, namely *TripleCheckMate*, wherein a user assesses an individual resource and evaluates each fact for correctness. In this case, the user is a LD expert who is conversant with RDF. In case of the *semi-automatic* process, the first phase involves the detection of common quality problems by the automatic creation of an extended schema for DBpedia. The second phase involves the generation and manual verification of schema axioms.

The semi-automatic process involves the generation and verification of schema axioms, which yielded a total of $222,982$ triples that have a high probability to be incorrect. We find that while a substantial number of problems exist, the overall quality is with a less than 11.93% error rate relatively high. With this study we not only aim to assess the quality of DBpedia but also to adopt a methodology to improve the quality in future versions by regularly providing feedback to the DBpedia maintainers to fix these problems.

Our main contributions are:

- a crowdsourcing based methodology for data quality assessment (Section 4.1),
- a comprehensive quality issue taxonomy comprising common knowledge extraction problems (Section 4.2),
- a crowdsourcing based data quality assessment tool (Section 4.3),
- an empirical data quality analysis of the DBpedia dataset performed using crowdsourcing (Section 5.3) and
- a semi-automated evaluation of data quality problems in DBpedia (Section 5.3).

We conclude with an outlook on future work in Section 7.5.

4.1. Assessment Methodology

In this section, we describe a generalized methodology for the assessment and subsequent data quality improvement of resources belonging to a dataset. The assessment methodology we propose is depicted in Figure 4.1. This methodology consists of the following four steps: 1. Resource selection, 2. Evaluation mode selection, 3. Resource evaluation and 4. Data quality improvement. In the following, we describe these steps in more detail.

[1] http://mturk.com

Step I: Resource selection. In this first step, the resources belonging to a particular dataset are selected. This selection can be performed in three different ways:

- *Per Class:* select resources belonging to a particular class
- *Completely random:* a random resource from the dataset
- *Manual:* a resource selected manually from the dataset

Choosing resources per class (e.g. animal, sport, place etc.) gives the user the flexibility to choose resources belonging to only those classes she is familiar with. However, when choosing resources from a class, the selection should be made in proportion to the number of instances of that class. Random selection, on the other hand, ensures an unbiased and uniform coverage of the underlying dataset. In the manual selection option, the user is free to select resources with problems that she has perhaps previously identified.

Step II: Evaluation mode selection. The assignment of the resources to a person or machine, selected in Step I, can be accomplished in the following three ways:

- *Manual:* the selected resources are assigned to a person (or group of individuals) who will then proceed to manually evaluate the resources individually.
- *Semi-automatic:* selected resources are assigned to a semi-automatic tool, which performs data quality assessment employing some form of user feedback.
- *Automatic:* the selected resources are given as input to an automatic tool, which performs the quality assessment without any user involvement.

For the semi-automatic evaluation, machine learning can be applied as shown in [Bühmann and Lehmann, 2012] and provided by the *DL-Learner framework* [Lehmann, 2009, Lehmann and Hitzler, 2010], where the workflow can be as follows: (i) based on the instance data, generate OWL axioms which can also be seen as restrictions[2], e.g. learn characteristics (irreflexivity, (inverse) functionality, asymmetry) of properties as well as definitions and disjointness of classes in the knowledge base; (ii) ask queries via SPARQL or a reasoner for violations of theses restrictions, e.g. in case of an irreflexive property, triples where subject and object are the same would indeed violate the characteristic of the irreflexivity. In the automatic case, a possible approach is to check for inconsistencies and other modelling problems as, e.g., described in [Lehmann and Bühmann, 2010].

[2] A local Unique Name Assumption is used therefore, i.e. every named individual is assumed to be different from every other, unless stated explicitly otherwise

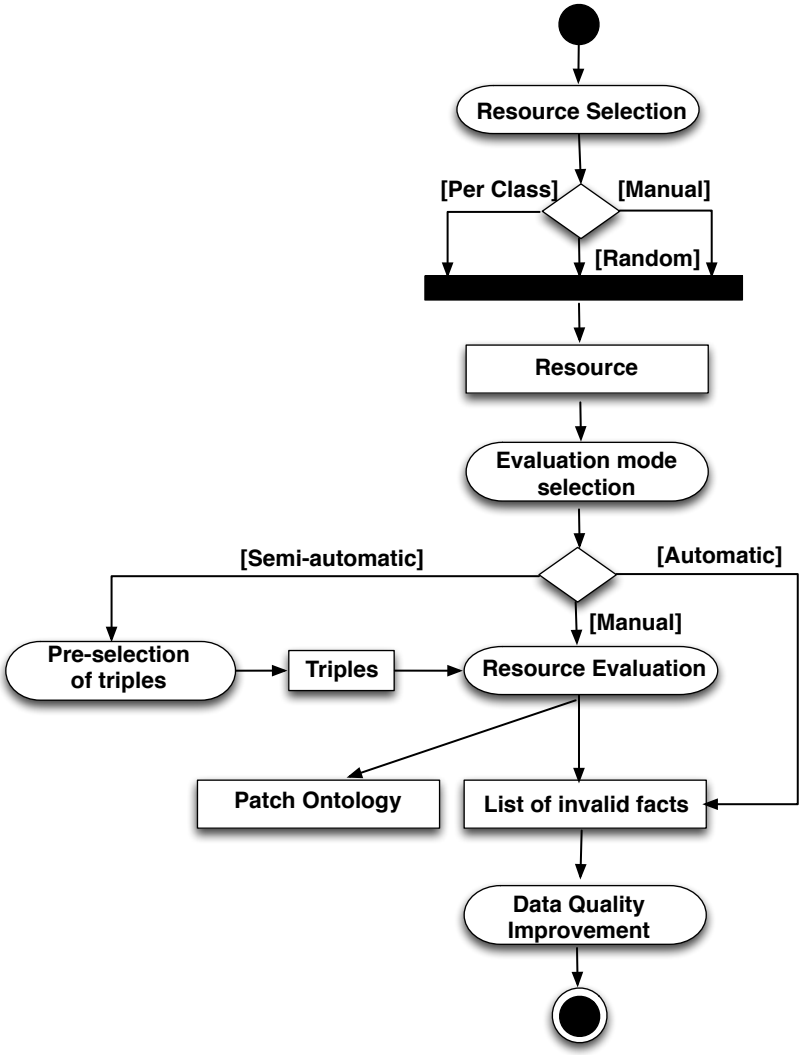


Figure 4.1.: Workflow of the data quality assessment methodology.

Step III: Resource evaluation. In case of manual assignment of resources, the person (or group of individuals) evaluates each resource individually to detect the potential data quality problems. In order to support this step, a quality assessment tool can be used which allows a user to evaluate each individual triple belonging to a particular resource. If, in case of Step II, the selected resources are assigned to a semi-automatic tool, the tool points to triples likely to be wrong. For example, domain or range problems are identified by the tool and then assigned to a person to verify the correctness of the results.

Step IV: Data quality improvement. After the evaluation of resources and identification of potential quality problems, the next step is to improve the data quality. There are at least two ways to perform an improvement:

- Direct: editing the triple, identified to contain the problem, with the correct value

- Indirect: using the Patch Request Ontology[3] [Knuth et al., 2012], which allows gathering user feedbacks about erroneous triples.

4.2. Quality Problem Taxonomy

A systematic review done in [Zaveri et al., 2015] identified a number of different data quality dimensions (criteria) applicable to LD. After carrying out an initial data quality assessment on DBpedia (as part of the first phase of the manual assessment methodology cf. subsubsection 4.4.1.1), the problems identified were mapped to this list of identified dimensions. In particular, *Accuracy, Relevancy, Representational-consistency and Interlinking* were identified to be problems affecting a large number of DBpedia resources. Additionally, these dimensions were further divided into categories and sub-categories.

Table 4.1 gives an overview of these data quality dimensions along with their categories and sub-categories. We indicate whether the problems are automatically detectable (column D) and fixable (column F). The ones marked with a ✔in column D refer to those categories that can be automatically identified such as invalid datatypes (`"1981-01-01T00:00:00+02:00"^^xsd:gYear`), irrelevant properties (`dbpprop:imageCaption`) or dead links. The column F refers to those categories that can be automatically amended, like fixing an invalid datatype (`"1981"^^xsd:gYear`) or removing triples with irrelevant properties and dead links. If the problem is fixable, we determined whether the problem can be fixed by amending the (i) extraction framework (E), (ii) the mappings wiki (M) or (iii) Wikipedia itself (W).

Moreover, the table specifies whether the problems are specific to DBpedia (marked with a ✔) or could potentially occur in any RDF dataset. For example, the sub-category *Special template not properly recognized* is a problem that occurs only in DBpedia due to the presence of specific keywords in Wikipedia articles that do not cite any references or resources (e.g. {{Unreferenced stub—auto=yes}}). On the other hand, the problems that are not DBpedia specific can occur in any other datasets. In the following, we provide the quality problem taxonomy and discuss each of the dimensions along with its categories and sub-categories in detail by providing examples.

4.2.1. Accuracy

Accuracy is defined as *the extent to which data is correct, that is, the degree to which it correctly represents the real world facts and is also free of error* [Zaveri et al., 2015]. We further classify this dimension into the categories (i) object incorrectly extracted, (ii) datatype problems and (iii) implicit relationship between attributes.

Object incorrectly extracted. This category refers to those problems, which arise when the object value of a triple is flawed. This may occur when the value is either (i)

[3] http://141.89.225.43/patchr/ontologies/patchr.ttl#

incorrectly extracted, (ii) incompletely extracted or (iii) the special template in Wikipedia is not recognized:

- *Object value is incorrectly extracted*, e.g.:
  `dbpedia:Oregon_Route_238 dbpprop:map "238.0"^^http://dbpedia.org/datatype/second.`
  This resource about state highway Oregon Route 238 has the incorrect property 'map' with value 238. In Wikipedia the attribute 'map' refers to the image name as the value: map=Oregon Route 238.svg. The DBpedia property only extracted the value 238 from his attribute value and gave it the datatype 'second' assuming it is a time value, which is incorrect.

- *Object value is incompletely extracted*, e.g.:
  `dbpedia:Dave_Dobbyn dbpprop:dateOfBirth "3" ^^xsd:integer.` In this example, only the day of birth of a person is extracted and mapped to the 'dateofBirth' property when it should have been the entire date i.e. day, month and year. Thus, the object value is not completely extracted.

- *Special template not properly recognized*, e.g.:
  `dbpedia:328_Gudrun dbpprop:auto "yes"@en.`
  Certain article classifications in Wikipedia (such as "This article does not cite any references or sources.") are performed via special templates (e.g. {{Unreferenced stub—auto=yes}}). Such templates should be listed on a black-list and omitted by the DBpedia extraction in order to prevent non-meaningful triples.

Datatype problems. This category refers to those triples which are extracted with an incorrect datatype for a typed literal.

- *Datatype incorrectly extracted*, e.g.:
  `dbpedia:Stephen_Fry dbpedia-owl:activeYearsStartYear "1981-01-01T00:00:00+02:00"^^xsd:gYear.` In this case, the DBpedia ontology datatype property `activeYearsStartYear` has `xsd:gYear` as range. Although the datatype declaration is correct, it is formatted as `xsd:dateTime`. The expected value is `"1981"^^xsd:gYear`.

Implicit relationship between attributes. This category of problems may arise due to (i) representation of one fact in several attributes, (ii) several facts encoded in one attribute or (iii) an attribute value computed from another attribute value in Wikipedia.

- *One fact is encoded in several attributes*, e.g.:
  `dbpedia:Barlinek dbpprop:postalCodeType "Postal code"@en.` In this example, the value of the postal code of the town of Barlinek is encoded in two attributes 'postal code type = Postal code' and 'postalcode = 74-320'. DBpedia extracts both these attributes separately instead of combining them together to produce one triple, such as:
  `dbpedia:Barlinek dbpprop:postalCode "74-320"@en.`

- *Several facts are encoded in one attribute*, e.g.:
  `dbpedia:Picathartes dbpedia-owl:synonym Galgulus "Wagler, 1827 (non Brisson, 1760:preoccupied)"@en.`
  In this example, even though the triple is not incorrect, it contains two pieces of information. Only the first word is the synonym, the rest of the value is a reference to that synonym. In Wikipedia, this fact is represented as ““synonyms = “Galgulus” ⟨*small*⟩ Wagler, 1827 (“non” [[Mathurin Jacques Brisson—Brisson]], 1760: [[Coracias—preoccupied]])/⟨/*small*⟩””. The DBpedia framework should ideally recognize this and separate these facts into several triples.

- *Attribute value computed from another attribute value*, e.g.:
  `dbpedia:Barlinek dbpprop:populationDensityKm "auto"@en.` In Wikipedia, this attribute is represented as “population density km2 = auto”. The word “auto” is an indication in Wikipedia that the value associated to that attribute should be computed “automatically”. In this case, the population density is computed automatically by dividing the population by area.

4.2.2. Relevancy

Relevancy refers to *the provision of information which is in accordance with the task at hand and important to the users' query* [Zaveri et al., 2015]. The only category *Irrelevant information extracted* of this dimension can be further sub-divided into the following sub-categories: (i) extraction of attributes containing layout information, (ii) image related information, (iii) redundant attribute values and (iv) other irrelevant information.

- *Extraction of attributes containing layout information*, e.g.:
  `dbpedia:Lærdalsøyri dbpprop:pushpinLabelPosition "bottom"@en.` Information related to layout of a page in Wikipedia, such as the position of the label on a pushpin map relative to the pushpin coordinate marker, in this example specified as ”bottom”, is irrelevant when extracted in DBpedia.

- *Image related information*, e.g.:
  `dbpedia:Three-banded_Plover dbpprop:imageCaption "At Masai Mara National Reserve, Kenya"@en.` Extraction of an image caption or name of the image is irrelevant in DBpedia as the image is not displayed for any DBpedia resource.

- *Redundant attributes value*, e.g.:
  The resource `dbpedia:Niedersimmental_District` contains the redundant properties `dbpedia-owl:thumbnail, foaf:depiction, dbpprop:imageMap` with the same value ”Karte Bezirk Niedersimmental 2007.png” as the object.

Dimension	Category	Sub-category	D	F	DBpedia specific
Accuracy	Triple incorrectly extracted	Object value is incompletely extracted	–	E	–
		Object value is incompletely extracted	–	E	–
		Special template not properly recognized	✔	E	✔
	Datatype problems	Datatype incorrectly extracted	✔	E	–
	Implicit relationship between attributes	One fact encoded in several attributes	–	M	✔
		Several facts encoded in one attribute	–	E	–
		Attribute value computed from another attribute value	–	E + M	✔
Relevancy	Irrelevant information extracted	Extraction of attributes containing layout information	✔	E	✔
		Redundant attribute values	✔	–	–
		Image related information	✔	E	✔
		Other irrelevant information	✔	E	–
Represensational-Consistency	Representation of number values	Inconsistency in representation of number values	✔	W	–
Interlinking	External links	External websites	✔	W	–
	Interlinks with other datasets	Links to Wikimedia	✔	E	–
		Links to Freebase	✔	E	–
		Links to Geospecies	✔	E	–
		Links generated via Flickr wrapper	✔	E	–

Table 4.1.: Data quality dimensions, categories and sub-categories identified in the DBpedia resources. Detectable (column D) means problem detection can be automized. Fixable (column F) means the issue is solvable by amending either the extraction framework (E), the mappings wiki (M) or Wikipedia (W). The last column marks the dataset specific subcategories.

- *Other irrelevant information*, e.g.: `dbpedia:IBM_Personal_Computer dbpedia:Template:Infobox_information_appliance "type"@en.` Information regarding a templates infobox information, in this case, with an object value as “type” is completely irrelevant.

4.2.3. Representational-consistency

Representational-consistency is defined as *the degree to which the format and structure of information conforms to previously returned information and other datasets.* [Zaveri et al., 2015] and has the following category:

- *Representation of number values*, e.g.: `dbpedia:Drei_Flüsse_Stadion dbpprop:seatingCapacity "20"^^xsd:integer.` In Wikipedia, the seating capacity for this stadium has the value "20.000", but in DBpedia the value displayed is only 20. This is because the value is inconsistently represented with a dot after the first two decimal places instead of a comma.

4.2.4. Interlinking

Interlinking is defined as *the degree to which entities that represent the same concept are linked to each other* [Zaveri et al., 2015]. This type of problem is recorded when links to external websites or external data sources are either incorrect, do not show any information or are expired. We further classify this dimension into the following categories:

- *External websites:* Wikipedia usually contains links to external web pages such as, for example, the home page of a company or a music band. It may happen that these links are either incorrect, do not work or are unavailable.
- *Interlinks with other datasets:* LD mandates interlinks between datasets. These links can either be incorrectly mapped or may not contain useful information. These problems are recorded in the following sub-categories: 1. Links to Wikimedia, 2. Links to Freebase, 3. Links to Geospecies, 4. Links generated via Flickr wrapper.

4.3. A Crowdsourcing Quality Assessment Tool

In order to assist several users in assessing the quality of a resource, we developed the *TripleCheckMate* tool[4] aligned with the methodology described in Section 4.1, in particular with Steps 1–3. To use the tool, the user is required to authenticate herself, which not only prevents spam but also helps in keeping track of her evaluations. After authenticating herself, she proceeds with the selection of a resource (Step 1). She is provided with three options: (i)*per class*, (ii)*completely random* and (iii)*manual* (as described in Step I of the assessment methodology).

After selecting a resource, the user is presented with a table showing each triple belonging to that resource on a single row. Step 2 involves the user evaluating each triple and checking whether it contains a data quality problem. The link to the original

[4]available at http://github.com/AKSW/TripleCheckMate

Wikipedia page for the chosen resource is provided on top of the page, which facilitates the user to check against the original values. If the triple contains a problem, she checks the box "is wrong". Moreover, she is provided with a taxonomy of pre-defined data quality problems where she assigns each incorrect triple to a problem. If the detected problem does not match any of the existing types, she has the option to provide a new type and extend the taxonomy. After evaluating one resource, the user saves the evaluation and proceeds to choosing another random resource and follows the same procedure.

Another important feature of the tool is to allow measuring of inter-rater agreements. That is, when a user selects a random method (*Any* or *Class*) to choose a resource, there is a 50% probability that she is presented with a resource that was already evaluated by another user. This probability as well as the number of evaluations per resource is configurable. Allowing many users evaluating a single resource not only helps to determine whether incorrect triples are recognized correctly but also to determine incorrect evaluations (e.g. incorrect classification of problem type or marking correct triples as incorrect), especially when crowdsourcing the quality assessment of resources. One important feature of the tool is that although it was built for DBpedia, it is parametrizable to accept any endpoint and, with very few adjustments in the database back-end (i.e. ontology classes and problem types) one could use it for any LD dataset (open or closed).

4.4. Evaluation of DBpedia Data Quality

4.4.1. Evaluation Methodology

4.4.1.1. Manual Methodology

We performed the assessment of the quality of DBpedia in two phases: *Phase I: Problem detection and creation of taxonomy* and *Phase II: Evaluation via crowdsourcing*.

Phase I: Creation of quality problem taxonomy. In the first phase, two researchers independently assessed the quality of 20 DBpedia resources each. During this phase an initial list of data quality problems, which occurred in each resource, was identified. These identified problems were mapped to the different quality dimensions from [Zaveri et al., 2015]. After analyzing the root cause of these problems, a refinement of the quality dimensions was done to obtain a finer classification of the dimensions. This classification of the dimensions into sub-categories resulted in a total of 17 types of data quality problems (cf. Table 4.1) as described in Section 4.2.

Phase II: Crowdsourcing quality assessment. In the second phase, we crowdsourced the quality evaluation wherein we invited researchers who are familiar with RDF to use the TripleCheckMate tool (described in Section 4.3). First, each user after authenticating oneself, chooses a resource by one of three options mentioned in Section 4.1. Thereafter, the extracted facts about that resource are shown to the user. The user then looks at each individual fact and records whether it contains a data quality problem and maps it to the

type of quality problem.

4.4.1.2. Semi-automatic Methodology

We applied the semi-automatic method (cf. Section 4.1), which consists of two steps: (i) the generation of a particular set of schema axioms for all properties in DBpedia and (ii) the manual verification of the axioms.

Step I: Automatic creation of an extended schema. In this step, the enrichment functionality of DL-Learner [Bühmann and Lehmann, 2012] for SPARQL endpoints was applied. Thereby for all properties in DBpedia, axioms expressing the (inverse) functional, irreflexive and asymmetric characteristic were generated, with a minimum confidence value of 0.95. For example, for the property `dbpedia-owl:firstWin`, which is a relation between Formula One racers and grand prix, axioms for all four mentioned types were generated: Each Formula One racer has only one first win in his career (functional), each grand prix can only be won by one Formula One racer (inverse functional). It is not possible to use the property `dbpedia-owl:firstWin` in both directions (asymmetric), and the property is also irreflexive.

Step II: Manual evaluation of the generated axioms. In the second step, we used at most 100 random axioms per axiom type and manually verified whether this axiom is appropriate. To focus on possible data quality problems, we restricted the evaluation data to axioms where at least one violation can be found in the knowledge base. Furthermore, we tried to facilitate the evaluation by taking also the target context into account, i.e. if it exists we consider the definition, domain and range as well as one random sample for a violation. When evaluating the inverse functionality for the property `dbpedia-owl:firstWin`, we can therefore make use of the following additional information:

```
Domain: dbpedia-owl:FormulaOneRacer Range: dbpedia-owl:GrandPrix
Sample Violation:
 dbpedia:Fernando_Alonso dbpedia-owl:firstWin dbpedia:2003_Hungarian_Grand_Prix.
 dbpedia:WikiProject_Formula_One dbpedia-owl:firstWin dbpedia:2003
     _Hungarian_Grand_Prix.
```

4.4.2. Evaluation Results

Manual Methodology. An overview of the evaluation results is shown in Table 4.2[5]. Overall, only 16.5% of all resources were not affected by any problems. On average, there were 5.69 problems per resource and 2.24 problems excluding errors in the *dbprop namespace*[6] [Lehmann et al., 2009]. While the vast majority of resources have problems, it should also be remarked that each resource has 47.19 triples on average, which is higher than in most other Linked Open Data (LOD) datasets. The tool was configured to allow two evaluations per resource and this resulted to a total of 268 inter-evaluations.

[5] Also available at: http://aksw.org/Projects/DBpediaDQ
[6] http://dbpedia.org/property/

Total no. of users	58
Total no. of distinct resources evaluated	521
Total no. of resources evaluated	792
Total no. of distinct resources without problems	86
Total no. of distinct resources with problems	435
Total no. of distinct incorrect triples	2928
Total no. of distinct incorrect triples in the *dbprop* namespace	1745
Total no. of inter-evaluations	268
No. of resources with evaluators having different opinions	89
Resource-based inter-rater agreement (Cohen's Kappa)	0.34
Triple-based inter-rater agreement (Cohen's Kappa)	0.38
No. of triples evaluated for correctness	700
No. of triples evaluated to be correct	567
No. of triples evaluated incorrectly	133
% of triples correctly evaluated	81
Average no. of problems per resource	5.69
Average no. of problems per resource in the *dbprop* namespace	3.45
Average no. of triples per resource	47.19
% of triples affected	11.93
% of triples affected in the *dbprop* namespace	7.11

Table 4.2.: Overview of the manual quality evaluation.

We computed the inter-rater agreement for those resources, which were evaluated by two persons by adjusting the observed agreement with agreement by chance as done in Cohen's kappa[7]. The inter-rater agreement results – 0.34 for resource agreement and 0.38 for triple agreement – indicate that the same resource should be evaluated more than twice in future evaluations. To assess the accuracy of the crowdsourcing evaluation, we took a random sample of 700 assessed triples (out of the total 2928) and evaluated them for correctness based on the formula in [Krejcie and Morgan, 1970] intended to be a representative of all the assessed triples. Additionally, we assumed a margin of 3.5% of error, which is a bound that we can place on the difference between the estimated correctness of the triples and the true value, and a 95% confidence level, which is the measure of how confident we are in that margin of error[8]. From these 700 triples, 133 were evaluated incorrectly resulting in about 81% of triples correctly evaluated.

Table 4.3 shows the total number of problems, the distinct resources and the percentage of affected triples for each problem type. Overall, the most prevalent problems, such as broken external links are outside the control of the DBpedia extraction framework. After that, several extraction and mapping problems that occur frequently mainly affecting accuracy, can be improved by manually adding mappings or possibly by improving the extraction framework.

[7] http://en.wikipedia.org/wiki/Cohen%27s_kappa
[8] http://research-advisors.com/tools/SampleSize.htm

Criteria	IT	DR	AT %
Accuracy			
Object incorrectly extracted	32	14	2.69
Object value is incorrectly extracted	259	121	23.22
Object value is incompletely extracted	229	109	20.92
Special template not recognized	14	12	2.30
Datatype problems	7	6	1.15
Datatype incorrectly extracted	356	131	25.14
Implicit relationship between attributes	8	4	0.77
One fact is encoded in several attributes	670	134	25.72
Several facts encoded in one attribute	87	54	10.36
Value computed from another value	14	14	2.69
Accuracy unassigned	31	11	2.11
Relevancy			
Irrelevant information extracted	204	29	5.57
Extraction of layout information	165	97	18.62
Redundant attributes value	198	64	12.28
Image related information	121	60	11.52
Other irrelevant information	110	44	8.45
Relevancy unassigned	1	1	0.19
Representational-consistency			
Representation of number values	29	8	1.54
Representational-consistency unassigned	5	2	0.38
Interlinking			
External websites (URLs)	222	100	19.19
Interlinks with other datasets (URIs)	2	2	0.38
Links to Wikimedia	138	71	13.63
Links to Freebase	99	99	19.00
Links to Geospecies	0	0	0.00
Links generated via Flickr wrapper	135	135	25.91
Interlinking unassigned	3	3	0.58

Table 4.3.: Detected number of problem for each of the defined quality problems. IT = Incorrect triples, DR = Distinct resources, AT = Affected triples.

When looking at the detectable and fixable problems from Table 4.1, in light of their prevalence, we expect that approximately one third of the problems can be automatically detected and two thirds are fixable by improving the DBpedia extraction framework. In particular, implicitly related attributes can be properly extracted with a new extractor, which can be configured using the DBpedia Mappings Wiki. As a result, we expect that the improvement potential is that the problem rate in DBpedia can be reduced from 11.93% to 5.81% (calculated by subtracting 7.11% from 11.93% reported in Table 4.2). After revising the DBpedia extraction framework, we will perform subsequent quality

assessments using the same methodology in order to realize and demonstrate these improvements.

Semi-automatic Methodology. The evaluation results in Table 4.4 show that for the irreflexive case all 24 properties that would lead to at least one violation should indeed be declared as irreflexive. Applying the irreflexive characteristic would therefore help to find overall 236 critical triples, for e.g. `dbpedia:2012_Coppa_ Italia_Final dbpedia-owl:followingEvent dbpedia:2012_Coppa_Italia_Final`, which is not meaningful as no event is the following event of itself. For asymmetry, we got 81 approved properties, for example, containing `dbpedia-owl:starring` with domain `Work` and range `Actor`. Compared with this, there are also some properties where asymmetry is not always appropriate, e.g. `dbpedia-owl:influenced`.

Characteristic	#Properties		Correct	#Violations			
	Total	Violated		Min.	Max.	Avg.	Total
Irreflexivity	142	24	24	1	133	9.8	236
Asymmetry	500	144	81	1	628	16.7	1358
Functionality	739	671	76	1	91581	2624.7	199,480
Inverse Functionality	52	49	13	8	18,236	1685.2	21,908

Table 4.4.: Results of the semi-automatic evaluation. The table shows the total number of properties that have been suggested to have the given characteristic by Step I of the semi-automatic methodology, the number of properties that would lead to at least one violation when applying the characteristic, the number of properties where the characteristic is meaningful (manually evaluated) and some metrics for the number of violations.

Functionality, i.e. having at most one value of a property, can be applied to 76 properties. During the evaluation, we observed invalid facts such as, for example, two different values `2600.0` and `1630.0` for the density of the moon `Himalia`. We spotted overall 199,480 errors of this type in the knowledge base. As the result of the inverse functionality evaluation, we obtained 13 properties where the object in the triple should only be related to one unique subject, e.g. there should only be one Formula One racer which won a particular grand prix, which is implicit when using the property `dbpedia-owl:lastWin`.

4.5. Summary

To the best of our knowledge, this study is the first comprehensive empirical quality analysis for more than 500 resources of a large LD dataset extracted from crowdsourced content. We found that a substantial number of problems exist and the overall quality,

with a 11.93% error rate, is moderate. Moreover, the semi-automatic analysis revealed more than $200,000$ violations of property characteristics. The detailed analysis of data quality problems occurring in DBpedia allows us to devise and implement corresponding mitigation strategies. Many of the problems found can be firstly automatically detected and secondly avoided by (i) improving existing extractors, (ii) developing new ones (e.g. for implicitly related attributes) or (iii) improving and extending mappings and extraction hints on the DBpedia Mappings Wiki. With this study, we not only aim to assess the quality of this sample of DBpedia resources but also adopt an agile methodology to improve the quality in future versions by regularly providing feedback to the DBpedia maintainers to fix these problems. We plan to improve the DBpedia extraction framework along these detected problems and periodically revisit the quality analysis (in regular intervals) in order to demonstrate possible improvements.

In addition to the quality analysis of DBpedia, we devised a generic methodology for LD quality analysis, derived a comprehensive taxonomy of extraction quality problems and developed a crowdsourcing tool, which can assist in the evaluation. All these contributions can be reused for analyzing any other extracted dataset (even by domain experts). In sum, we showed that the employment of Linked Data experts to assess the quality issues of LD is feasible to a certain extent as it can be time-consuming and costly. We illustrated that a combination of user-driven and semi-automated methodology to perform LD quality assessment is a feasible means of performing quality assessment.

5. Crowdsourcing Linked Data Quality Assessment

Many would agree that LD is one of the most important technological developments in data management of the last decade. However, one of the less positive aspects of this great success story is related to the varying quality of LD sources, which often poses serious problems to developers aiming to seamlessly consume and integrate LD in their applications. This state of the affairs is the result of a combination of data- and process-related factors. Keeping aside the factual flaws of the original sources, the array of data sources that may be subject to RDFication is highly heterogeneous in terms of format, organization and vocabulary. As a direct consequence, some kinds of data tend to be more challenging to translate into RDF than others, leading to errors in the LD provisioning process. Some of the quality issues hence produced (e.g. missing values) can be easily repaired automatically, but others require manual intervention. In this chapter, we look into the use of crowdsourcing as a data curation strategy that is cost-efficient and accurate in terms of the level of granularity of the errors to be spotted. This chapter is based on [Acosta et al., 2013]. This is a joint work, where I have contributed as a co-author in the following tasks:

- creation of a taxonomy of data quality problems occurring in LD
- creation of the gold standard dataset
- comparison and evaluation of results from the crowdsourcing tasks with the gold standard dataset
- calculation and comparison of results from the crowdsourcing tasks with the baseline results

We analyzed the most common quality problems encountered in LD sources and classified them according to the extent to which they are likely to be amenable to a specific form of crowdsourcing. Based on this analysis, we implemented a quality assessment methodology for LD that leverages the wisdom of the crowds in the following ways: (i) we first launched a **contest** targeting an expert crowd of LD researchers and enthusiasts in order to *find* and classify erroneous RDF triples; and then (ii) published the outcome of this contest as **paid microtasks** on Amazon Mechanical Turk (MTurk)[1] in order to *verify* the issues spotted by the experts [Bernstein et al., 2010].

[1] https://www.mturk.com/

These two crowdsourcing approaches have advantages and disadvantages. Each approach makes specific assumptions about the audiences they address (the 'crowd') and the skills of the potential contributors. A contest reaches out to the crowd to solve a given problem and rewards the best ideas; it exploits competition and intellectual challenge as main drivers for participation. The idea, originating from open innovation, has been employed in many domains, from creative industries to sciences, for tasks of varying complexity (from designing logos to building sophisticated algorithms). We applied this contest-based model to mobilize an expert crowd consisting of researchers and LD enthusiasts to discover and classify quality issues in DBpedia. The participant who covered the highest number of DBpedia resources won a prize.

Microtask crowdsourcing traditionally covers a different set of scenarios. Tasks primarily rely on basic human abilities, including visual and audio cognition, as well as natural language understanding and communication (sometimes in different languages) and less on acquired skills (such as subject-matter knowledge). As such, a great share of the tasks addressed via microtask platforms like MTurk could be referred to as 'routine' tasks – recognizing objects in images, transcripting audio and video material and text editing. To be more efficient than traditional outsourcing (or even in-house resources), the tasks need to be highly parallelized. This means that the actual work is executed by a high number of contributors in a decentralized fashion; this not only leads to significant improvements in terms of time of delivery, but also offers a means to cross-check the accuracy of the answers (as each task is typically assigned to more than one person) and reward the workers according to their performance and productivity. We applied microtask crowdsourcing as a fast and cost-efficient way to examine the errors spotted by the expert crowd who participated in the contest. More concretely, we looked into three types of quality problems the experts found in DBpedia: (i) object values incorrectly or incompletely extracted; (ii) data type incorrectly extracted; and (iii) incorrect links between DBpedia entities and related sources on the Web. The underlying data was translated into HITs, the unit of work in MTurk, which were handled by workers on the MTurk platform.

We empirically evaluated how this methodology – based on a mixed crowdsourcing approach – could efficiently spot quality issues in DBpedia. The results show that the two styles of crowdsourcing are complementary and that crowdsourcing-enabled quality assessment is a promising and affordable yet a limited way to enhance the quality of LD sets, which, in the long run, may address many of the problems that fundamentally constrain the usability of the Web of Data in real-world applications.

5.1. Linked Data Quality Issues

The Web of Data spans a network of data sources of varying quality. There are a large number of high-quality data sets, for instance, in the life-science domain, which are the result of decades of thorough curation and have been recently made available as

LD[2]. Other data sets, however, have been (semi-)automatically translated to RDF from their primary sources, or via crowdsourcing in a decentralized process involving a large number of contributors. Probably the best example of a data set produced in this manner is DBpedia [Lehmann et al., 2009]. While the combination of machine-driven extraction and crowdsourcing was a reasonable approach to produce a baseline version of a greatly useful resource, it was also the cause of a wide range of quality problems, in particular in the mappings between Wikipedia attributes and their corresponding DBpedia properties.

Our analysis of LD quality issues focuses on DBpedia as a representative data set for the broader Web of Data due to the diversity of the types of errors exhibited and the vast domain and scope of the data set. In our previous work [Zaveri et al., 2015], we compiled a list of data quality dimensions (criteria) applicable to Linked Data quality assessment. Afterwards, we mapped these dimensions to DBpedia [Zaveri et al., 2013a]. A sub-set of four criteria of the original framework were found particularly relevant in this setting: *Accuracy*, *Relevancy*, *Representational-Consistency* and *Interlinking*. To provide a comprehensive analysis of DBpedia quality, we further divided these four categories of problems into sub-categories. For the purpose of this study, from these categories we chose the following three triple-level quality issues.

Object incorrectly/incompletely extracted. Consider the triple: `dbpedia:Firewing dbpprop:isbn "978"^^xsd:integer`. This DBpedia resource is about the children's book 'Firewing', with the incomplete and incorrect value of the ISBN number. Instead of extracting the entire ISBN number from Wikipedia, `978-0-00-639194-4`, only the first three digits were extracted.

Data type incorrectly extracted. This category refers to triples with an incorrect data type for a typed literal. For example, in the DBpedia ontology, the range of the property
verb—activeYearsStartYear— is defined as `xsd:gYear`. Although the data type declaration is correct in the triple `dbpedia:Stephen_Fry dbpedia-owl:activeYears-StartYear "1981-01-01T00:00:00+02:00"^^xsd:gYear`, it is formatted as `xsd:dateTime`. The expected value is `"1981"^^xsd:gYear`.

Interlinking. In this case, links to external Web sites or other external data sources such as Wikimedia, Freebase, GeoSpecies or links generated via the Flickr wrapper are incorrect; that is, they do not show any related content pertaining to the resource.

The categories of quality problems just discussed occur pervasively in DBpedia. These problems might be present in other data sets, which are extracted in a similar fashion as DBpedia. Given the diversity of the situations in which they can be instantiated (broad range of data types and object values) and their sometimes deeply contextual character (interlinking), assessing them automatically is challenging. In the following we explain how crowdsourcing could support quality assessment processes.

[2] http://beta.bio2rdf.org/

5.2. Crowdsourcing

Our work on human-driven Linked Data quality assessment focuses on two forms of crowdsourcing: contests and paid microtasks. As discussed earlier, these crowdsourcing approaches exhibit different characteristics in terms of the types of tasks they can be applied to, the way the results are consolidated and exploited, and the audiences they target. Table 5.1 presents a summary of the two approaches as they have been used in this work for Linked Data quality assessment purposes.

Table 5.1.: Comparison between the proposed approaches to crowdsource LD quality assessment.

Characteristic	**Contest-based**	**Paid microtasks**
Participants	Controlled group: LD experts	Anonymous large group
Goal per task	Detecting and classifying LD quality issues	Confirming LD quality issues
Task size	Participants explore RDF resources and identify incorrect triples	Participants analyze human-readable information of given RDF triples
Task complexity	*Difficult:* the task requires knowledge on data quality issues	*Easy:* the task consists of validating pre-processed and classified triples
Time duration	Long (weeks)	Short (days)
Reward	A final prize	Micropayments
Reward mechanism	The winner gets the prize	Each participant receives a payment
Tool/platform	*TripleCheckMate*	Amazon Mechanical Turk (MTurk)

We applied the crowdsourcing pattern *Find-Fix-Verify* [Bernstein et al., 2010] to assess the quality of DBpedia. This pattern consists of a three-stage process, which is originally defined as follows. The *Find* stage asks the crowd to identify problematic elements within a data source. In the second stage, *Fix*, the crowd corrects the elements belonging to the outcome of the previous stage. The *Verify* stage corresponds to a final quality control iteration. Our approach (see Figure 5.1) leverages the expertise of LD experts in a contest to *find* and classify erroneous triples according to a pre-defined scheme [Zaveri et al., 2015]. The outcome of this stage – triples judged as 'incorrect' – is then *verified* by the MTurk workers, who are instructed to assess specific types of errors in the subset of triples. The implementation of the *fix* stage is out of the scope of this study, since the main goal of this work is identifying quality issues.

The Find-Fix-Verify pattern reduces the noise caused by low-quality participants, while the costs remain competitive with other crowdsourcing alternatives. In addition, this approach is efficient in terms of the number of questions asked to the paid microtask

crowd. In scenarios in which crowdsourcing is applied to enhance or validate the results of machine computation tasks, question filtering relies on specific thresholds or historical information about the likelihood that human input will significantly improve the results generated algorithmically. Find-Fix-Verify addresses scenarios, which can be hardly engineered, like in our case the discovery and classification of various types of errors in DBpedia. In these scenarios, in a first step one applies crowdsourcing not only to solve the task at hand, but also to define the specific questions, which need to be addressed. These steps can employ different types of crowds, as they require different skills and expertise [Bernstein et al., 2010]. In the following we elaborate on the specific processes carried out by each type of crowd in this work.

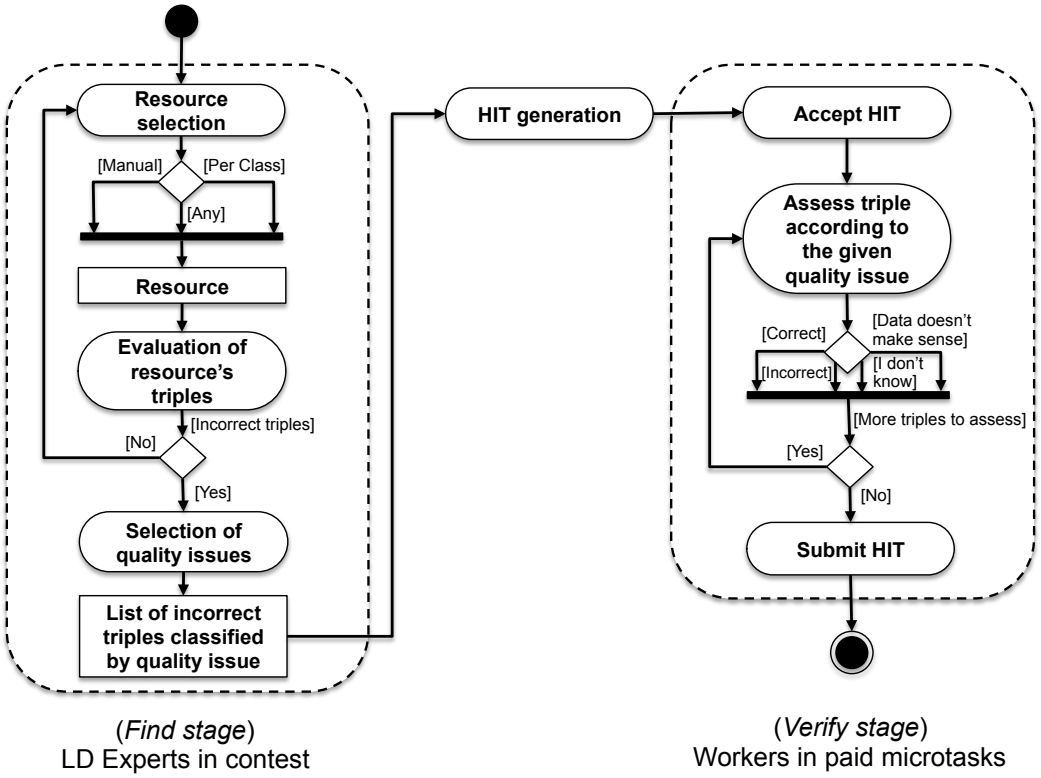


Figure 5.1.: Workflow of the applied Linked Data quality assessment methodology.

5.2.1. Contest-based Crowdsourcing

Contests as means to successfully involve experts in advancing science have a long-standing tradition in research, e.g. the Darpa challenges[3] and NetFlix[4]. In our case, we reached out to an expert crowd of researchers and Linked Data enthusiasts via a contest, in order to identify and classify specific types of Linked Data quality problems in

[3] http://www.darpa.mil/About/History/Archives.aspx
[4] http://www.netflixprize.com/

DBpedia. To collect the contributions from this crowd, in a previous work of ours [Zaveri et al., 2015], we developed a web-based tool, *TripleCheckMate*[5] (see Figure 5.2), which allows users to select resources, identify issues related to triples of the resource and classify these issues according to a pre-defined taxonomy of data quality problems. A prize was announced for the user submitting the highest number of (real) quality problems.

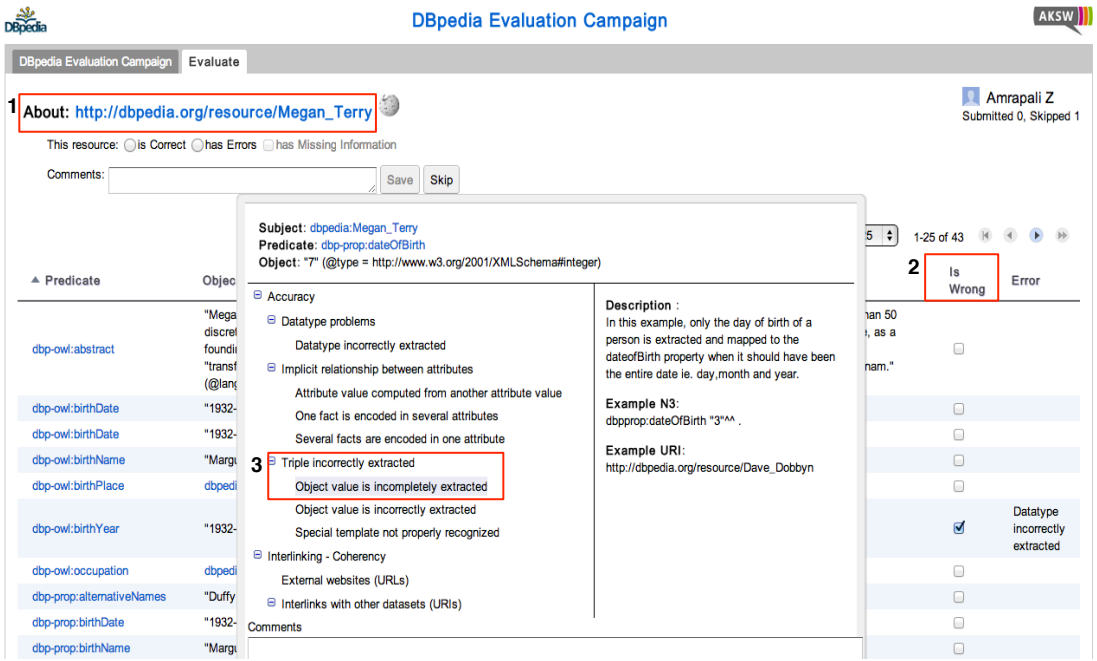


Figure 5.2.: Screenshot of the *TripleCheckMate* crowdsourcing data quality assessment tool.

As a basic means to avoid spam, each user first has to login using her Google Mail ID. Then, as shown in Figure 5.1, she is presented with three options to choose a resource from DBpedia: (i) *Any*, for random selection; (ii) *Per Class*, where she may choose a resource belonging to a particular class of her interest; and (iii) *Manual*, where she may provide a URI of a resource herself. Once a resource is selected following one of these alternatives, the user is presented with a table in which each row corresponds to an RDF triple of that resource. The next step is the actual quality assessment at triple level. The user is provided with the link to the corresponding Wikipedia page of the given resource in order to offer more context for the evaluation. If she detects a triple containing a problem, she checks the box 'Is Wrong'. Moreover, she can assign these troublesome triples to quality problems (according to the classification devised in [Zaveri et al., 2013a]), as shown in Figure 5.2. In this manner, the tool only records the triples that are identified as 'incorrect'. This is consistent with the *Find* stage from the Find-Fix-Verify pattern, where the crowd exclusively detects the problematic elements; while the remaining data is not taken into consideration.

[5]Available at http://github.com/AKSW/TripleCheckMate

The tool *TripleCheckMate* measures inter-rater agreements. This means that DBpedia resources are typically checked multiple times. This redundancy mechanism is extremely useful to analyze the performance of the users (as we compare their responses against each other), to identify quality problems which are likely to be real (as they are confirmed by more than one opinion) and to detect unwanted behavior (as users are not 'rewarded' unless their assessments are 'consensual').

The outcome of this contest corresponds to a set of triples judged as 'incorrect' by the experts and classified according to the detected quality issue.

5.2.2. Paid Microtasks

To fully unfold its benefits, this form of crowdsourcing needs to be applied to problems which can be broken down into smaller units of work (called 'microtasks' or 'Human Intelligence Tasks' – HITs) that can be undertaken in parallel by independent parties[6]. As noted earlier, the most common model implies small financial rewards for each worker taking on a microtask, whereas each microtask may be assigned to more than one worker in order to allow for techniques such as majority voting to automatically identify accurate responses.

We applied this crowdsourcing approach in order to *verify* quality issues in DBpedia RDF triples identified as problematic during the contest (see Figure 5.1). One of the challenges in this context is to develop useful human-understandable interfaces for HITs. In microtasks, optimal user interfaces reduce ambiguity as well as the probability to retrieve erroneous answers from the crowd due to a misinterpretation of the task. Further design criteria were related to spam detection and quality control; we used different mechanisms to discourage low-effort behavior, which leads to random answers and to identify accurate answers (see subsubsection 5.3.1.2).

Based on the classification of LD quality issues explained in Section 5.1, we created three different types of HITs. Each type of HIT contains the description of the procedure to be carried out to complete the task successfully. We provided the worker examples of incorrect and correct examples along with four options: (i) Correct; (ii) Incorrect; (iii) I cannot tell/I don't know; (iv) Data doesn't make sense. The third option was meant to allow the user to specify when the question or values were unclear. The fourth option referred to those cases in which the presented data was truly unintelligible. Furthermore, the workers were not aware that the presented triples were previously identified as 'incorrect' by experts and the questions were designed such that the worker could not foresee the right answer. The resulting HITs were submitted to Amazon Mechanical Turk using the MTurk SDK for Java[7]. We describe the particularities of each type of HIT in the following.

[6] More complex workflows, though theoretically feasible, require additional functionality to handle task dependencies.

[7] http://aws.amazon.com/code/695

(a) **Incorrect** object value in DBpedia (b) **Correct** object value in DBpedia

Figure 5.3.: Incorrect/incomplete object value: The crowd must compare the DBpedia and Wikipedia values and decide whether the DBpedia entry is correct or not for a given subject and predicate.

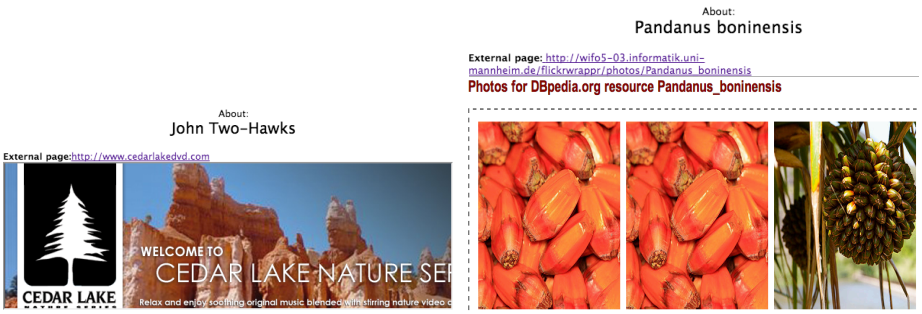
Incorrect/incomplete object value. In this type of microtask, we asked the workers to evaluate whether the value of a given RDF triple from DBpedia is correct or not. Instead of presenting the set of RDF triples to the crowd, we displayed human-readable information retrieved by dereferencing the Uniform Resource Identifier (URI) of the subject and predicate of the triple. In particular, we selected the values of the `foaf:name` or `rdfs:label` properties for each subject and predicate. Additionally, in order to provide contextual information, we implemented a wrapper which extracted the corresponding data encoded in the infobox of the Wikipedia article – specified as `foaf:isPrimaryTopicOf` of the subject. Figure 5.3 depicts the interface of the resulting tasks.

In the task presented in Figure 5.3, the worker must decide whether the date of birth of "Dave Dobbyn" is correct. According to the DBpedia triple, the value of this property is `3`, while the information extracted from Wikipedia suggests that the right value is `3 January 1957`. In addition, it is evident that the DBpedia value is erroneous as the value "3" is not appropriate for a date. Therefore, the right answer to this tasks is: the DBpedia data is incorrect.

An example of a DBpedia triple whose value is correct is depicted in Figure 5.3. In this case, the worker must analyze the date of birth of "Elvis Presley". According to the information extracted from Wikipedia, the date of birth of Elvis Presley is `January 8, 1935`, while the DBpedia value is `1935-01-08`. Despite the dates are represented in different formats, semantically the dates are indeed the same, thus the DBpedia value is correct.

Incorrect data type. This type of microtask consists of detecting those DBpedia triples whose data type – specified via `@type` – was not correctly assigned. The generation of the interfaces for these tasks was very straightforward, by dereferencing the URIs of the subject and predicate of each triple and displaying the values for the `foaf:name` or `rdfs:label`.

In the description of the task, we introduced the concept of data type of a value and provided two simple examples. The first example illustrates when the data type is incorrect while analyzing the entity "Torishima Izu Islands": `Given the property "name", is the value "鳥島" of type "English"?` A worker does not need to understand that the name of this island is written in "Japanese", since it is evident



(a) External link displaying **unrelated** content to the subject.

(b) Web page displaying **related** images to the subject.

Figure 5.4.: Incorrect link: The crowd must decide whether the content from an external web page is related to the subject.

that the language type "English" in this example is incorrect. In a similar fashion, we provided an example where the data type is assigned correctly by looking at the entity "Elvis Presley": `Given the property "name", is the value "Elvis Presley" of type "English"?`. According to the information from DBpedia, the value of the name is written in English and the type is correctly identified as English.

Incorrect links. In this type of microtask, we asked the workers to verify whether the content of the external page referenced from the Wikipedia article corresponds to the subject of the RDF triple. For the interface of the HITs, we provided the worker a preview of the Wikipedia article and the external page by implementing HTML `iframe` tags. In addition, we retrieved the `foaf:name` of the given subject and the link to the corresponding Wikipedia article using the predicate `foaf:isPrimaryTopicOf`.

Examples of this type of task are depicted in Figure 5.4. In the first example, the workers must decide whether the content in the given external web page is related to "John Two-Hawks". It is easy to observe that in this case the content is not directly associated to the person "John Two-Hawks". Therefore, the right answer is that the link is incorrect. On the other hand, we also exemplified the case when an interlink presents relevant content to the given subject. Consider the example in Figure 5.4, where the subject is the plant "Pandanus boninensis" and the external link is a web page generated by the DBpedia Flickr wrapper. The web page indeed shows pictures of the subject plant. Therefore, the correct answer is that the link is correct.

5.3. Evaluation

In our evaluation we investigated the following research questions: **(RQ2.3)** Is it possible to detect quality issues in LD data sets via crowdsourcing mechanisms? **(RQ2.4)** What type of crowd is most suitable for each type of quality issues? **(RQ2.5)** Which types of errors are made by lay users and experts?

5.3.1. Experimental Design

In the following we describe the settings of the crowdsourcing experiments and the creation of a gold standard to evaluate the results from the contest and microtasks.

5.3.1.1. Contest Settings

Participant expertise: We relied on the expertise of members of the Linked Open Data and the DBpedia communities who were willing to take part in the contest.
Task complexity: In the contest, each participant was assigned the full one-hop graph of a DBpedia resource. All triples belonging to that resource were displayed and the participants had to validate each triple individually for quality problems. Moreover, when a problem was detected, she had to map it to one of the problem types from a quality problem taxonomy.
Monetary reward: We awarded the participant who evaluated the highest number of resources a Samsung Galaxy Tab 2 worth 300 EU.
Assignments: Each resource was evaluated by at most two different participants.

5.3.1.2. Microtask Settings

Worker qualification: In MTurk, the requester can filter workers according to different qualification metrics. In this experiment, we recruited workers whose previous HIT acceptance rate is greater than 50%.
HIT granularity: In each HIT, we asked the workers to solve 5 different questions. Each question corresponds to an RDF triple and each HIT contains triples classified into one of the three quality issue categories discussed earlier.
Monetary reward: The micropayments were fixed to 4 US dollar cents. Considering the HIT granularity, we paid 0.04 US dollar per 5 triples.
Assignments: In MTurk, a requester can specify the number of different workers to be assigned to solve each HIT. This allows collecting multiple answers for each question, thus compensating the lack of LD-specific expertise of the workers. This mechanism is core to microtask crowdsourcing, which is primarily dedicated to 'routine' tasks that make no assumption about the knowledge or skills of the crowd besides basic human capabilities. The number of assignments was set up to 5 and the answer was selected applying majority voting. We additionally compared the quality achieved by a group of workers vs. the resulting quality of the worker who submitted the first answer.

5.3.1.3. Creation of Gold Standard

Two of the authors of this study (MA, AZ) generated the gold standard for all the triples obtained from the contest and submitted to MTurk. To generate the gold standard, each author independently evaluated the triples. After an individual assessment, they compared their results and resolved the conflicts via mutual agreement. The inter-rater agreement between them was 0.4523 for object values, 0.5554 for data types and 0.5666 for interlinks. The inter-rater agreement values were calculated using the

Cohen's kappa measure. Disagreement arose in the object value triples when one of the reviewers marked number values which are rounded up to the next round number as correct. For example, the length of the course of the "1949 Ulster Grand Prix" was 26.5Km in Wikipedia but rounded up to 27Km in DBpedia. In case of data types, most disagreements were considering the data type "number" of the value for the property "year" as correct. For the links, those containing unrelated content were marked as correct by one of the reviewers since the link existed in the original Wikipedia page.

The tools used in our experiments and the results are available online, including the outcome of the contest,[8] the gold standard and microtask data (HITs and results).[9]

5.3.2. Results

The contest was open for a predefined period of time of three weeks. During this time, 58 LD experts analyzed 521 distinct DBpedia resources and, considering an average of 47.19 triples per resource in this data set [Zaveri et al., 2013a], we could say that the experts browsed around 24,560 triples. They detected a total of 1,512 triples as erroneous and classified them using the given taxonomy. After obtaining the results from the experts, we filtered out duplicates, triples whose objects were broken links and the external pages referring to the DBpedia Flickr Wrapper. In total, we submitted 1,073 triples to the crowd. A total of 80 distinct workers assessed all the RDF triples in four days. A summary of these observations are shown in Table 5.2.

We compared the common 1,073 triples assessed in each crowdsourcing approach against our gold standard and measured precision as well as inter-rater agreement values for each type of task (see Table 5.3). For the contest-based approach, the tool allowed two participants to evaluate a single resource. In total, there were 268 inter-evaluations for which we calculated the triple-based inter-agreement (adjusting the observed agreement with agreement by chance) to be 0.38. For the microtasks, we measured the inter-rater agreement values between a maximum of 5 workers for each type of task using Fleiss' kappa measure. While the inter-rater agreement between workers for the interlinking was high (0.7396), the ones for object values and data types was moderate to low with 0.5348 and 0.4960, respectively.

5.3.2.1. Incorrect/missing Values

As reported in Table 5.3, our crowdsourcing experiments reached a precision of 0.90 for MTurk workers (majority voting) and 0.72 for LD experts. Most of the missing or incomplete values that are extracted from Wikipedia occur with the predicates related to dates, for example: `(2005 Six Nations Championship, Date, 12)`. In these cases, the experts and workers presented a similar behavior, classifying 110 and 107 triples correctly, respectively, out of the 117 assessed triples for this class. The difference in precision between the two approaches can be explained as follows. There were 52 DBpedia triples whose values might seem erroneous, although they were correctly extracted

[8] http://nl.dbpedia.org:8080/TripleCheckMate/
[9] http://people.aifb.kit.edu/mac/DBpediaQualityAssessment/

Table 5.2.: Overall results in each type of crowdsourcing approach.

	Contest-based	**Paid microtasks**
Number of distinct participants	Total: 58	Values: 35 Data type: 31 Interlink: 31 Total: 80
Total time	3 weeks (predefined)	4 days
Total no. of triples evaluated	1,512	1,073
Object value	550	509
Data type	363	341
Interlinks	599	223

Table 5.3.: Inter-rater agreement and precision values achieved with the implemented approaches.

	Object values	**Data types**	**Interlinks**
	Inter-rater agreement		
LD experts	Calculated for all the triples: 0.38		
MTurk workers	0.5348	0.4960	0.7396
	(True positives, False positives)		
LD experts	(364, 145)	(282, 59)	(34, 189)
MTurk workers (first answer)	(257, 108)	(144, 138)	(21, 13)
MTurk workers (majority voting)	(307, 35)	(134, 148)	(32, 2)
Baseline	N/A	N/A	(33, 94)
	Achieved precision		
LD experts	0.7151	0.8270	0.1525
MTurk workers (first answer)	0.7041	0.5106	0.6176
MTurk workers (majority voting)	0.8977	0.4752	0.9412
Baseline	N/A	N/A	0.2598

from Wikipedia. One example of these triples is: `(English (programming language), Influenced by, ?)`. We found out that the LD experts classified all these triples as incorrect. In contrast, the workers successfully answered that 50 out of this 52 were correct, since they could easily compare the DBpedia and Wikipedia values in the HITs.

5.3.2.2. Incorrect Data Types

Table 5.3 exhibits that the experts are reliable (with 0.83 of precision) on *finding* this type of quality issue, while the precision of the crowd (0.51) on *verifying* these triples is relatively low. In particular, the first answers submitted by the crowd were slightly better than the results obtained with majority voting. A detailed study of these cases showed that 28 triples that were initially classified correctly, later were misclassified, and most of these triples refer to a language data type. The low performance of the MTurk workers compared to the experts is not surprising, since this particular task requires certain technical knowledge about data types and, moreover, the specification of values and types in LD.

In order to understand the previous results, we analyzed the performance of experts and workers at a more fine-grained level. We calculated the frequency of occurrences of data types in the assessed triples (see Table 5.4) and reported the number of true positives (TP) and false positives (FP) achieved by both crowdsourcing methods for each data type. Figure 5.6 depicts these results. The most notorious result in this task is the assessment performance for the data type "number". The experts effectively identified triples where the data type was incorrectly assigned as 'number"[10], for instance, in the triple `(Walter Flores, date of birth, 1933)` the value `1933` was number instead of date. These are the cases where the crowd was confused and determined that data type was correct, thus generating a large number of false positives. Nevertheless, it could be argued that the data type "number" in the previous example is not completely incorrect, when being unaware of the fact that there are more specific data types for representing time units. Under this assumption, the precision of the crowd would have been 0.8475 and 0.8211 for first answer and majority voting, respectively.

While looking at the typed literals in "English" (in RDF `@en`), Figure 5.6 shows that the experts perform very well when discerning whether a given value is an English text or not. The crowd was less successful in the following two situations: (i) the value corresponded to a number and the remaining data was specified in English, e.g. `(St. Louis School Hong Kong, founded, 1864)`; and (ii) the value was a text without special characters, but in a different language than English, for example German `(Woellersdorf-Steinabrueckl, Art, Marktgemeinde)`. The performance of both crowdsourcing approaches for the remaining data types were similar or not relevant due the low number of triples processed.

5.3.2.3. Incorrect Links

For this type of task, we additionally implemented a baseline approach to decide whether the linkage was correct. This automatic solution retrieves for each triple the external web page – which corresponds to the object of the triple – and searches for occurrences of the `foaf:name` of the subject within the page. If the number of occurrences is greater than 1, the algorithm interprets the external page as being related

[10] This error is very frequent when extracting dates from Wikipedia as some resources only contain partial data, e.g. only the year is available and not the whole date.

Table 5.4.: Frequency of data types in the crowdsourced triples.

Data type	Frequency		Data type	Frequency
Date	8		Number with decimals	19
English	127		Second	20
Millimetre	1		Volt	1
Nanometre	1		Year	15
Number	145		Not specified/URI	4

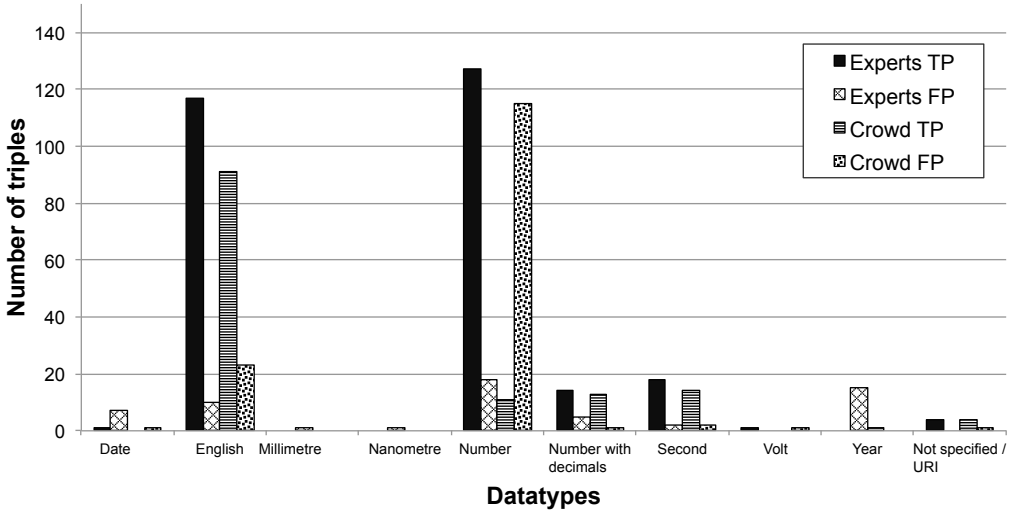


Figure 5.5.: True positives (TP) and false positives (FP) per data type in each crowdsourcing method.

Figure 5.6.: Analysis of true and false positives in "Incorrect data type" task.

to the resource. In this case the link is considered correct.

Table 5.3 displays the precision for each studied quality assessment mechanism. The implemented baseline approach achieved a precision of 0.26. It obviously failed in the cases where the external pages correspond to an image (which is the case of the 33% of the evaluated triples). On the other hand, the extremely low precision of 0.15 of the contest's participants was unexpected. We discarded the possibility that the experts have made these mistakes due to a malfunction of the *TripleCheckMate* tool used during the contest. We analyzed in details the 189 misclassifications of the experts:

- The 95 Freebase links[11] connected via `owl:sameAs` were marked as incorrect, although both the subject and the object were referring to same real-world entity,
- there were 77 triples whose objects were Wikipedia-upload entries; 74 of these triples were also classified incorrectly,

[11] http://www.freebase.com

- 20 links (blogs, web pages, etc.) referenced from the Wikipedia article of the subject were also misclassified, regardless of the language of the content in the web page.

The two settings of the MTurk workers outperformed the baseline approach. The 'first answer' setting reports a precision of 0.62, while the 'majority voting' achieved a precision of 0.94. The 6% of the links that were not properly classified by the crowd correspond to those web pages whose content is in a different language than English or, despite they are referenced from the Wikipedia article of the subject, their association to the subject is not straightforward. Examples of these cases are the following subjects and links: 'Frank Stanford' and `http://nw-ar.com/drakefield/`, 'Forever Green' and `http://www.stirrupcup.co.uk`f. We hypothesize that the design of the user interface of the HITs – displaying a preview of the web pages to analyze – helped the workers to easily identify those links containing related content to the triple subject.

5.4. Discussion

Referring back to the research questions formulated at the beginning of Section 5.3, our experiments let us understand the strengths and weaknesses of applying crowdsourcing mechanisms for data quality assessment, following the Find-Fix-Verify pattern. For instance, we were able to detect common cases in which none of the two forms of crowdsourcing we studied seem to be feasible **(RQ2.5)**. The most problematic task for the LD experts was the one about discerning whether a web page is related to a resource. Although the experimental data does not provide insights into this behavior, we are inclined to believe that this is due to the relatively higher effort required by this specific type of task, which involves checking an additional site outside the *TripleCheckMate* tool. In turn the MTurk workers did not perform so well on tasks about data types where they recurrently confused numerical data types with time units.

In each type of task, the LD experts and MTurk workers applied different skills and strategies to solve the assignments successfully **(RQ2.4)**. The data collected for each type of task suggests that the effort of LD experts must be applied on the *Find* stage of those tasks demanding specific-domain skills beyond common knowledge. On the other hand, the MTurk crowd was exceptionally good and efficient at performing comparisons between data entries, especially when some contextual information is provided. This result suggests that microtask crowdsourcing can be effectively applied on the *Verify* stage of these tasks and possibly on the *Find* stage of the 'incorrect links' task.

Regarding the accuracy achieved in both cases, we compared the outcomes produced by each of the two crowds against a manually defined gold standard and against an automatically computed baseline, clearly showing that both forms of crowdsourcing offer feasible solutions to enhance the quality of Linked Data data sets **(RQ2.3)**.

One of the goals of our work is to investigate how the contributions of the two crowdsourcing approaches can be integrated into LD curation processes, by evaluating the performance of the two crowds in a cost-efficient way. In order to do this, both

crowds must evaluate a common set of triples. The straightforward solution would be submitting to MTurk all the triples assessed by the LD experts, i.e. all the triples judged as 'incorrect' and 'correct' in the contest. As explained in Section 5.3, the experts browsed around $24,560$ triples in total. Considering our microtask settings, the cost of submitting all these triple to MTurk would add up to over US$ $1,000$. By contrast, our methodology aims at reducing the number of triples submitted to the microtask platform, while asking the workers to assess only the problematic triples found by the experts. By doing this, the cost of the experiments was reduced to only US$ 43.

The design of our methodology allowed us to exploit the strengths of both crowds: the LD experts detected and classified data quality problems, while the workers confirmed or disconfirmed the output of the experts in 'routine' tasks. In addition, an appropriate quality assurance methodology requires a quality control iteration, in this case performed by the MTurk workers. As can be seen in our experimental results (Table 5.3), it was not always the case that the triples judged as *incorrect* by the LD experts were indeed incorrect. In fact, the number of misjudged triples by the experts was 145 (out of 509) for incorrect/missing values, 59 (out of 341) for incorrect data type and 189 (out of 223) for incorrect interlinking. Therefore, always agreeing with the experts would deteriorate the overall output of the quality assurance process. In addition, the workers did not know that the data provided to them was previously classified as problematic. In consequence, the turkers could not have applied an strategy to guess the right answers.

5.5. Summary

In this chapter, we presented a methodology that adjusts the crowdsourcing pattern Find-Fix-Verify to exploit the strengths of experts and microtask workers. The *Find* stage was implemented using a contest-based format to engage with a community of LD experts in discovering and classifying quality issues of DBpedia resources. We selected a subset of the contributions obtained through the contest (referring to flawed object values, incorrect data types and missing links) and asked the MTurk crowd to *Verify* them. The evaluation showed that both types of approaches are successful but limited; in particular, the microtask experiments revealed that people with no expertise in Linked Data can be a useful resource to identify only very specific quality issues in an accurate and affordable manner, using the MTurk model. On the other hand, we showed that the assessment by LD experts was error-prone, costly and time consuming. We consider our methodology can be applied to RDF data sets, which are extracted from other sources and, hence, are likely to suffer from similar quality problems as DBpedia.

6. Semi-automated Quality Assessment of Linked Data

In this chapter, we describe the implementation of selected LD quality metrics that were identified as part of the survey we conducted (Chapter 3). This implementation was done by P. Westphal, C. Stadler, and J. Lehmann, who are colleagues from the AKSW group. This chapter is an excerpt from their technical report [Westphal et al., 2014], which implements several quality metrics – identified as part of the survey (as described in Chapter 3) – and provides a tool, namely R2RLint, to perform semi-automated quality assessment of LD. In particular, the technical details of the R2RLint[1] tool is described in this chapter, which builds upon the conceptual framework of the quality metrics to provide users different means of assessing the quality of their linked dataset. Details of 13 quality metrics belonging to seven quality dimensions (Section 6.1) are provided that can be used to assess the quality of Linked Data. The results of using this tool to assess the quality of four datasets that are part of the use case are provided in Chapter 7.

6.1. Data Quality Metrics

Data quality assessment involves the measurement of quality dimensions or criteria that are relevant to the consumer. A data quality assessment metric or measure is a procedure for measuring a data quality dimension [Zaveri et al., 2015]. Thus, to assess the quality of the four datasets, we selected a total of seven dimensions and 13 metrics, each of which we define in this section. For the actual metric definitions the following sets are introduced:

- $\mathcal{R}$ containing all RDF resources
- $\mathcal{B}$ containing all blank nodes, with $\mathcal{B} \subset \mathcal{R}$
- $\mathcal{L}$ containing all RDF literals, i.e. typed and plain literals
- $\mathcal{N}$ containing all RDF nodes, with $\mathcal{N} = \mathcal{R} \cup \mathcal{B} \cup \mathcal{L}$
- $\mathcal{T}$ containing all triples, with $\mathcal{T} = \mathcal{R} \times \mathcal{R} \setminus \mathcal{B} \times \mathcal{N}$

[1] https://github.com/AKSW/R2RLint

- $\mathcal{D}$ containing all datasets, with $\mathcal{D} = \mathbb{P}(\mathcal{T})$[2,3]

Besides this the functions s, p and o are defined to return the subject, predicate and object of an input triple. Moreover, if these functions get a *set* of triples as input, they yield the sets of all subjects, predicates and objects, respectively. The selected dimensions and metrics are defined as follows:

Availability

Metric 1 (A1: Dereferenceability of the URIs)**.** *Given a resource $r \in \mathcal{R} \setminus \mathcal{B}$, the quality score with regards to its* dereferenceability *is given as*

$$f(r) = \begin{cases} 1 & \text{if } \begin{array}{l}\textit{using } r\textit{'s URI as URL and}\\ \textit{requesting the corresponding}\\ \textit{Web resource via } \texttt{HTTP GET},\\ \textit{the returned HTTP response}\\ \textit{code is 200 after resolving}\\ \textit{any redirects}\end{array} \\ 0 & \textit{otherwise} \end{cases} \tag{6.1}$$

Completeness

Metric 2 (C1: Interlinking Completeness)**.** *Given the set R_{local} containing all local resources of a dataset $D \in \mathcal{D}$ as*

$$R_{local} = \bigcup_{t \in D} \left\{ r \;\middle|\; \begin{array}{l} r \in \begin{pmatrix} s(t) \cup \\ p(t) \cup \\ o(t) \end{pmatrix} \cap \mathcal{R} \wedge \\ \textit{the string representation of } r\textit{'s URI starts with a local prefix} \end{array} \right\} \cup \mathcal{B} \tag{6.2}$$

the cardinalities $|D|_{inst}$ counting the instances and $|D|_{ext}$ counting the interlinked external resources of a dataset D, are given as follows:

$$|D|_{inst} = \left| \left\{ s(t) \;\middle|\; \begin{array}{l} t \in D \wedge \\ s(t) \notin \begin{pmatrix} \texttt{rdfs:Class} \sqcup \\ \texttt{owl:Class} \end{pmatrix} \end{array} \right\} \cup \left\{ o(t) \;\middle|\; \begin{array}{l} t \in D \wedge \\ o(t) \notin \begin{pmatrix} \texttt{rdfs:Class} \sqcup \\ \texttt{owl:Class} \end{pmatrix} \\ o(t) \notin \mathcal{L} \wedge \\ p(t) \neq \texttt{owl:sameAs} \end{array} \right\} \right| \tag{6.3}$$

[2] $\mathbb{P}(S)$ here denotes the power set of a set S

[3] It has to be noted, that this dataset definition differs from the common definition of a dataset as a set of graphs, that consists of triples [Harris and Seaborne, 2013]. Even though the following metrics could also be introduced based on that definition, graphs are not considered here for brevity.

$$|D|_{ext} = \left| \begin{array}{l} \left\{ s(t) \left| \begin{array}{l} t \in D \ \wedge \ s(t) \notin R_{local} \wedge \\ o \in R_{local} \wedge \\ s(t) \notin \left(\begin{array}{l} \texttt{rdfs:Class} \sqcup \\ \texttt{owl:Class} \end{array} \right) \end{array} \right. \right\} \cup \\ \left\{ o(t) \left| \begin{array}{l} t \in D \ \wedge s(t) \in R_{local} \wedge \\ o(t) \notin R_{local} \ \wedge \ o(t) \notin \mathcal{L} \\ o(t) \notin \left(\begin{array}{l} \texttt{rdfs:Class} \sqcup \\ \texttt{owl:Class} \end{array} \right) \end{array} \right. \right\} \end{array} \right| \tag{6.4}$$

The interlinking completeness of a dataset D is then determined by the following function:

$$f(D) = \frac{|D|_{ext}}{|D|_{inst}} \tag{6.5}$$

Interlinking

Metric 3 (I1: External owl:sameAs Links)**.** *Given the cardinality $|D|_{ext\_same}$ counting the external same-as links of a dataset $D \in \mathcal{D}$ with*

$$|D|_{ext\_same} = \left| \begin{array}{l} \left\{ t \left| \begin{array}{l} s(t) \in R_{local} \wedge \\ p(t) = \texttt{owl:sameAs} \wedge \\ o(t) \notin R_{local} \end{array} \right. \right\} \cup \\ \left\{ t \left| \begin{array}{l} s(t) \notin R_{local} \wedge \\ p(t) = \texttt{owl:sameAs} \wedge \\ o(t) \in R_{local} \end{array} \right. \right\} \end{array} \right| \tag{6.6}$$

the quality score of a D with regards to its external `owl:sameAs` *links is defined as*

$$f(D) = \frac{|D|_{ext\_same}}{|D|} \tag{6.7}$$

Syntactic validity

Metric 4 (S1: Datatype-compatible Literals)**.** *Given a literal $l \in \mathcal{L}$, the quality score with regards to its datatype compatibility is defined as*

$$f(l) = \begin{cases} 0 & \text{if } l\text{'s value is not compatible with } l\text{'s datatype (e.g. according to [Peterson et al., 2012])} \\ 1 & \text{otherwise} \end{cases} \tag{6.8}$$

Metric 5 (S2: Valid Language Tag)**.** *Given a literal $l \in \mathcal{L}$, the quality score with regards to the validity of its language tag is defined as*

$$f(l) = \begin{cases} 0 & \text{if } l\text{'s language tag is not compliant with the BCP 47 standard [Phillips and Davis, 2009]} \\ 1 & \text{otherwise} \end{cases} \tag{6.9}$$

Consistency

Metric 6 (CO1: Basic Ontology Conformance)**.** *Given a dataset $D \in \mathcal{D}$ and the set $\mathcal{D}_{uvoc} \subset \mathcal{D}$ of vocabularies and ontologies used in D, the quality score with regards to its conformance to consistency aspects like*

- *CO1.1: Correct Datatype Property Value*
- *CO1.2: Correct Object Property Values*
- *CO1.3: Disjoint Classes Conformance*
- *CO1.4: Valid Range*

is determined by the function

$$f(D) = \begin{cases} 0 & \text{if } D \cup (\bigcup \mathcal{D}_{uvoc}) \text{ contains any contradictions} \\ 1 & \text{otherwise} \end{cases} \tag{6.10}$$

Metric 7 (CO2: Homogeneous Datatypes)**.** *Given the dataset $D \in \mathcal{D}$, the following set contains the occurrences of all properties in D and their value types:*

$$M = \bigcup_{t \in D} \left\{ (r, \theta) \;\middle|\; \begin{array}{l} r = p(t) \;\wedge\; o(t) \in \mathcal{L} \wedge \\ o(t) \text{ is of datatype } \theta \end{array} \right\} \tag{6.11}$$

The quality score with regards to the homogeneity of a given property $r \in \mathcal{R}$ is then defined as

$$f(r) = \begin{cases} 0 & \text{if } \left| \left\{ (r_M, \theta) \;\middle|\; \begin{array}{l} (r_M, \theta) \in M \wedge \\ r = r_M \end{array} \right\} \right| > 1 \\ 1 & \text{otherwise} \end{cases} \tag{6.12}$$

Metric 8 (CO3: Well-placed Classes)**.** *Considering the set $CLASSES$ containing the classes of a dataset $D \in \mathcal{D}$, the quality score of a resource $r \in CLASSES$ is defined by the following function:*

$$f(r) = \begin{cases} 0 & \text{if } r \in p(D) \\ 1 & \text{otherwise} \end{cases} \tag{6.13}$$

Metric 9 (CO4: No Ontology Hijacking)**.** *Given the set $\mathcal{D}_{avoc} \subset \mathcal{D}$ of known vocabularies, a triple $t \in D$ with $D \in \mathcal{D}$ is a violation with respect to the No Ontology Hijacking metric, if for any of the vocabularies $D_{avoc} \in \mathcal{D}_{avoc}$, $s(t) \in s(D_{avoc})$. In case, the URI of $s(t)$ does not share the local prefix(es) of D, but $s(t) \notin s(D_{avoc})$, t is considered as* bad smell*. The corresponding function to determine the quality score of a triple t is given as*

$$f(t) = \begin{cases} 0 & \text{if } \exists D_{avoc} \left(\begin{array}{l} D_{avoc} \in \mathcal{D}_{avoc} \wedge \\ s(t) \in s(D_{avoc}) \end{array} \right) \\ & \quad s(t) \notin R_{local} \\ 0.5 & \text{if } \nexists D_{avoc} \left(\begin{array}{l} D_{avoc} \in \mathcal{D}_{avoc} \wedge \\ s(t) \in s(D_{avoc}) \end{array} \right) \\ 1 & \text{otherwise} \end{cases} \tag{6.14}$$

Interpretability

Metric 10 (IN1: Typed Resources)**.** *The quality score of a local resource $r \in D$ with $D \in \mathcal{D}$ and $r \in R_{local}$ is determined by the function*

$$f(r) = \begin{cases} 1 & \text{if } (r, \texttt{rdf:type}, o) \in D \ \wedge \ o \in \mathcal{R} \\ 1 & \text{if } (r, \texttt{rdf:type}, \texttt{rdfs:Class}) \in D \\ 1 & \text{if } (r, \texttt{rdf:type}, \texttt{owl:Class}) \in D \\ 1 & \text{if } (r, \texttt{rdfs:subClassOf}, o) \in D \wedge o \in \mathcal{R} \\ 1 & \text{if } (r, \texttt{rdfs:subPropertyOf}, o) \in D \wedge o \in \mathcal{R} \\ 1 & \text{if } (r, \texttt{owl:equivalentClass}, o) \in D \wedge o \in \mathcal{R} \\ 1 & \text{if } (r, \texttt{owl:equivalentProperty}, o) \in D \wedge o \in \mathcal{R} \\ 0 & \text{otherwise} \end{cases} \tag{6.15}$$

Representational conciseness

Metric 11 (R1: Correct Collection Use)**.** *Given the dataset $D \in \mathcal{D}$ and a statement $t_i^{rest} \in D$ descibing an* `rdf:rest` *of a collection, the assessment of the correct collection use comprises the following checks:*

a) rest statement has `rdf:nil` subject*: check, if $s(t_i^{rest})$ =* `rdf:nil`

b) rest statement has literal object*: check, if $o(t_i^{rest})$ is a literal*

c) none or multiple first statements*: check, if there is none or more than one statement t_i^{first} with $s(t_i^{first}) = s(t_i^{rest})$ and $p(t_i^{first}) =$* `rdf:first`

d) first statement has literal object*: if there is a statement t_i^{first}, check if $o(t_i^{first})$ is a literal*

e) collection not terminated with `rdf:nil`*: check, if $o(t_i^{rest}) \neq$* `rdf:nil` *and there is no statement t_{i+1}^{rest} with $s(t_{i+1}^{rest}) = o(t_i^{rest})$*

f) multiple successors*: check, if there are multiple statements t_{i+1}^{rest} with $s(t_{i+1}^{rest}) = o(t_i^{rest})$*

g) multiple predecessors*: check, if there are multiple statements t_{i-1}^{rest} with $o(t_{i-1}^{rest}) = s(t_i^{rest})$*

The quality score of a collection rest statement $t_i^{rest} \in D$ is then defined by the following function:

$$f(t_i^{res}) = \begin{cases} 0 & \text{if any of the checks b) and d) is positive} \\ 0.5 & \text{if any of the checks a), c), e), f) and g) is positive} \\ 1 & \text{otherwise} \end{cases} \tag{6.16}$$

Metric 12 (R2: Correct Container Use)**.** *Given the dataset $D \in \mathcal{D}$ and a statement $t_i \in D$ having a container membership property on predicate position, the assessment of the correct container use comprises the following checks:*

a) container not typed: *if $p(t_i)$ = `rdf:_1`, check if neither `rdf:Bag`, `rdf:Seq`, nor `rdf:Alt` is assigned to $s(t_i)$ via `rdf:type`*

b) literal objects: *check, if $o(t_i)$ is a literal*

c) multiple entries for one container membership property: *check, if there is a statement $t_{i'}$ with $s(t_{i'}) = s(t_i)$, $p(t_{i'}) = p(t_i)$ and $o(t_{i'}) \neq o(t_i)$ (with $p(t_{i'}) \in$ `rdfs:ContainerMembershipProperty`)*

d) numbering gaps: *check, if*

- *there is a statement t_{i+2} with $s(t_{i+2}) = s(t_i)$ and the predicates of t_i and t_{i+2} are differing in two steps (with $p(t_{i+2})$ being the bigger one),*
- *but no statement t_{i+1} with $s(t_{i+1}) = s(t_i)$ and the predicates of t_i and t_{i+1} differing in one step (with $p(t_{i+1})$ being the bigger one)*

e) container starts at `rdf:_0`*: check if there is a statement t_0 with $p(t_0)$ = `rdf:_0`*

f) container membership properties with leading zeros: *check if there are statements t_i with $p(t_i)$ having leading zeros, e.g. `rdf:_023`*

The quality score of a container statement $t_i \in D$ is then defined by the following function:

$$f(t_i) = \begin{cases} 0 & \text{if any of the checks b), e) and f) is positive} \\ 0.5 & \text{if any of the checks a), c) and d) is positive} \\ 1 & \text{otherwise} \end{cases} \tag{6.17}$$

Metric 13 (R3: Correct Reification Use)**.** *Given the dataset $D \in \mathcal{D}$ and a statement $t_i \in D$ with either $p(t_i) \in \{$`rdf:subject` , `rdf:predicate` , `rdf:object`$\}$, or t_i being typed as `rdf:Statement`, the assessment of the correct reification use comprises the following checks:*

a) reification not typed properly: *check, if $s(t_i)$ is not typed as `rdf:Statement`*

b) none or multiple `rdf:subject` statements: *check, if there is none or more than one statement t_s with $s(t_s) = s(t_i)$ and $p(t_s)$ = `rdf:subject`*

c) literal value of `rdf:subject` property: *if t_s exists, check if $o(t_s)$ is a literal*

d) none or multiple `rdf:predicate` statements: *check, if there is none or more than one statement t_p with $s(t_p) = s(t_i)$ and $p(t_p)$ = `rdf:predicate`*

e) literal or blank node value of `rdf:predicate` property: *if t_p exists, check if $o(t_p)$ is a literal or a blank node*

f) none or multiple `rdf:object` statements: *check, if there is none or more than one statement t_o with $s(t_o) = s(t_i)$ and $p(t_o) =$ `rdf:object`*

6.2. Summary

In this chapter, we provided details of 13 quality metrics belonging to seven quality dimensions that can be used to assess a Linked Data source. In particular, for each dimension we provide different means of measuring it. These metrics are implemented as part of a tool, namely, R2RLint. This tool is employed to assess the quality of four datasets included in our use case and the results of this quality assessment are provided in Chapter 7.

7. Use Case Leveraging on Data Quality

7.1. Linked Data and Data Quality on the Web

With the recently LD paradigm [Heath and Bizer, 2011] emerging as a simple mechanism for employing the Web for data and knowledge integration, different communities are using Linked Data to provide and exchange information [Heath and Bizer, 2011]. LD, in fact, allows us to build mashups which go beyond the interlinking of data from different sources to uncover meaningful and impactful relationships. However, in all these efforts, one crippling problem is the underlying *data quality*. Incomplete, inconsistent or inaccurate data may strongly affect the results, leading to unreliable conclusions.

The objective of this chapter is to show the advantage of utilizing Linked Data in a particular use case i.e. to build the Health Economic Research (HER) Observatory, which takes into account the quality of the integrated datasets. In particular, the observatory aims to assess the impact of Research and Development (R&D) on countries' economic performance and healthcare for which we use the Structural Equation Modeling (SEM) methodology. This chapter is based on [Zaveri et al., 2014b], where I have contributed in (i) converting and extracting data and the respective variables and (ii) analyzing and interpreting the results from the data quality assessment and the SEM. The main contributions of this chapter are threefold:

- Show the usefulness of Linked Data to build the HER Observatory,
- Perform semi-automated quality assessment of the four datasets integrated into the observatory by using the R2RLint tool (described in Chapter 6) and show the importance of data quality,
- Apply Structural Equation Modeling on the datasets of the observatory to assess the impact of R&D on economic performance and healthcare.

The chapter is structured as follows. In Section 7.2 we discuss previous efforts undertaken for analyzing societal progress indicators, which are the foundation of this study, and formulate the main research question. In Section 7.3, we explain the SEM methodology and provide details on the datasets, variables and data extraction performed. In Section 7.4,the results of the data quality and of the SEM assessment are reported. Section 7.5 concludes with directions to future work.

7.2. Background and Research Question

In this section, we first describe the previous works, which use LD for analyzing several societal progress indicators (Section 7.2.1). Thereafter, we discuss in detail the limitations of these previous studies, proposing feasible solutions to overcome them (Section 7.2.2). We finally formulate the research questions we intend to answer in this chapter (Section 7.2.3).

7.2.1. Previous Efforts

ReDD-Observatory. The ReDD-Observatory is a project to evaluate the disparity between active areas of biomedical research and the global burden of disease using LD and data-driven discovery [Zaveri et al., 2011]. In particular, data from three datasets is used: (i) ClinicalTrials.gov, (ii) PubMed[1] and (iii) Global Health Observatory (GHO)[2] are obtained from their linked data sources (in case of ClinicalTrials.gov and PubMed), converted to linked data (in case of GHO) and interlinked to form an integrated dataset. This integrated dataset is then queried to answer questions in terms of how good or bad investments in terms of clinical trials or research have proved successful in curing illnesses. This is done by evaluating the disparity between the amount of research and burden of disease i.e. by querying the integrated datasets. The final aim of this effort is to provide policy makers with a set of well-informed reports. The idea is motivated by the fact that policy makers, in particular in emerging regions, are unable to access such information due to high costs of obtaining it. Therefore, they are unable to appropriately allocate resources, which in turn can negatively affect the quality of life of significant parts of the population.

Structural Equation Modeling. In a previous project [Zaveri et al., 2013d], the Structural Equation Modeling methodology was applied to publicly available Linked Data in an attempt to study the correlation between R&D activity and (i) economic, (ii) educational and (iii) healthcare performances in European countries. In particular, data from EuroStat and the World Bank (which were already available as LD) was extracted. Specific variables were selected from both datasets and following substantial analysis, a model was identified that provided the best fit for our hypothesized model to reach the best possible adequacy and theoretical reasoning. As a result, it was determined that investments in R&D positively influences educational status, but curiously, does not directly influence economic performance. This project also supported the idea that LD can facilitate these types of studies backed by robust statistical analysis.

Despite achieving meaningful results for these particular societal progress indicators, in both previous efforts, we encountered a number of limitations, which are discussed in the following section.

[1] http://www.ncbi.nlm.nih.gov/pubmed
[2] http://apps.who.int/gho/data/node.main

7.2.2. Limitations

The main limitations that were encountered in the previous two studies are (i) *data quality* problems such as unavailability, incompleteness, reliability, syntactic validity; (ii) *inadequacy of the research indices* to calculate specific societal progress indicators; (iii) insufficient *coverage* of the data to analyze results over several years, insufficient variables and outdated data; all the three of them considerably affected the results. We thus discuss each problem in detail along with the respective solution, to build a reliable and up-to-date observatory of societal progress.

L1: Data Quality. In both the previous efforts to analyzing the societal progress indicators, the results were severely affected due to several data quality problems. Problems such as unavailability and/or incompleteness of the data, which not only includes data being unavailable to be queried but also not available for a particular country or year were major hindrances in acquiring reliable results. The unavailability of data may be due to the server being inaccessible at the time of querying. Incompleteness may be due to the fact that certain developing countries do not have the standard means to collect health-related information or they just do not have access to healthcare facilities.

Another data quality issue is the inadequate number of interlinks between the datasets. The number of interlinks, which could be automatically discovered, were limited as the datasets did not contain standardized identifiers for naming diseases, countries etc. For example, "AIDS" in LinkedCT[3] could not automatically be matched with "Acquired Immunodeficiency Syndrome" in PubMed using basic string similarity. Although we attempted to address this limitation through the use of Unified Medical Language System (UMLS)[4] (in particular MeSH) for interlinking diseases, the links were still not complete. Moreover, there were a smaller number of diseases present in the GHO dataset as compared to LinkedCT and PubMed (due to different levels of specificity) indicating lower coverage or more general entries, thus further complicating the automatic linking. In case of missing values, those particular countries with incomplete data had to be excluded in the SEM methodology.

Additionally, the syntactic validity, inconsistency of the dataset, problems in interpretability of the dataset as well as representational-conciseness were several other problems which hindered the achievement of reliable results. Moreover, trustworthiness of the datasets was an issue, which led to claims of the disparity as well as impact of R&D on the economic and healthcare performances being debatable due to the questionable reliability and provenance of the underlying datasets. Thus, in this chapter, we assess the quality of the four datasets used in the analysis and fix the problems identified before utilizing them in our usage scenario. In particular, we use the R2RLint tool Chapter 6 to perform the quality assessment. We choose this methodology over the other crowdsourcing quality assessments discussed earlier (Chapter 4 and Chapter 5) as those methodologies showed that the assessment was not only time-consuming and costly but also error prone. In this semi-automatic assessment, the users are also involved

[3] http://linkedct.org
[4] http://www.nlm.nih.gov/research/umls/

(as she can choose the metrics that apply to the use case) and is also provided with the triples causing the errors so the results are interpretable. Moreover, the assessment process is less error-prone and inexpensive to perform quality assessment over four datasets that are part of this use case.

L2: Methodology. The indices used in the ReDD-Observatory to calculate the disparity suffered from lack of accuracy in evaluating the disparity between amount of research performed on a particular disease and the burden of the disease. Moreover, it was difficult to interpret the results directly based on the output since the indices used to measure the disparity did not directly represent the imbalance. That is, first of all, the number of publications on a particular disease was not representative of the value of the information in the publications. Then, the direct comparison of the number of publications with the burden of disease was not entirely reliable to calculate the disparity. Therefore, we use the SEM methodology supported by Exploratory Factor Analysis (EFA) to calculate the societal progress indicators. This output is thus much more reliable since it relies on statistical correlations between different variables and thus is easily interpretable.

L3: Coverage. Despite the exponential growth of data available on the Web, coverage is still a major issue. When integrating data and performing analysis on the integrated dataset, the coverage of the base data is important. In our case, the GHO data included reports for death and Disability Adjusted Life Years (DALY) measures only till the year 2004, thus limiting the overall coverage of our integrated substrate. Therefore, we now utilize the data from the WorldBank dataset, which reports the values for the total number of deaths per country per year for all diseases. Additionally, the LinkedCT dataset that we used in the ReDD-Observatory was from the 2010 RDF dump[5]. In this next iteration, we thus use the 2013 dump[6]. This dump contains up-to-date data about any new trials started as well as status of the current trials i.e. if they are completed along with the results.

Moreover, we include several other indicators from different datasets such as United States Patent and Trademark Office (USPTO) patent data (as an indicator for innovation), country level publication statistics from the Scimago dataset (as an indicator of scientific performance of countries) and WorldBank for several other indicators (discussed in detail in Section 7.3.2) to expand the coverage of the data analyzed. Additionally, in the previous study, we only included European countries, whereas now we analyze data for all countries, thus expanding the coverage of the base data used for the analysis.

[5] http://www.cs.toronto.edu/~oktie/linkedct/linkedct-dump-2010-02-10.tar.gz

[6] http://www.cs.toronto.edu/~oktie/linkedct/linkedct-dump-2013-10-01.nt.bz2/

7.2.3. Research Question

Besides healthcare, *basic research and development* (R&D) is a crucial important driver for innovation, economic progress and social welfare [Adams, 1990, Henderson et al., 1998]. Scientific production concerns especially basic research, but the results, which are generated are not only long-term ones but produce spillovers that have short and medium term effects on industrial innovation [Mansfield, 1995]. This innovation-economic growth nexus has been a focal point of academic researchers and policy makers due to the evidence of increase in economic growth in those countries that increase the level of patenting activities [Hasan and Tucci, 2010]. For example, Europe as a whole has a high impact on the global biomedical literature, having contributed with a growing number of articles ($210,433$ publications in public health research [Larsen and von Ins, 2010]) and a significant citation impact [Lab, 2012]. The impact of Europe on broader healthcare and social welfare issues, however, is poorly understood. In other words, we know little on whether the biomedical research currently produced in any country translates into better economic and healthcare conditions to the local population. Although the credit goes to the university research for economic impact, there is no consensus on how to measure it [Bessette, 2003]. Measuring this impact poses a challenging endeavor, which involves the identification, gathering and analyzing of diverse data. Thus, the research question that we aim to address is:

- Can we link structured information to evaluate the impact of R&D on economic and healthcare performance in a country?

7.3. Methodology and Datasets

In this section, we first explain the methodology (Section 7.3.1) and then provide details of the datasets and variables used along with the data extraction procedure (Section 7.3.2).

7.3.1. Methodology

In order to address the problems identified regarding the methodology used in previous studies (L2 Methodology – Section 7.2.2), we use the SEM methodology to calculate the impact of R&D on health and economic performance of countries. Thus, in this section we describe the various aspects related to SEM along with the steps applied.

7.3.1.1. Exploratory Data Analysis

Performing exploratory data analysis is essential to analyze the feasibility of the data to detect problematic variables, missing values, outliers and other descriptive information about the data to be included in the analysis. All the information from each of the 17 variables (described in Section 7.3.2) about all the 196 countries was retrieved and analyzed.

In particular, when there was a pattern of missing data (lacking the last three years, for instance) the country was excluded from the sample. Otherwise, if no pattern was found a multiple imputation method was applied to deal with the incompleteness [Schafer, 2008]. Normality distribution was assessed through the Anderson-Darling normality test [Gross, 2012] to detect oscillations in the Gaussian distribution in order to adjust the analytical methods to the appropriate distribution. We used the Mahalonobis distance [Stats Package, 2013] to identify univariate and multivariate outliers and the Mardia coefficient and multivariate kurtosis to identify multivariate normality [Korkmaz, 2013]. Either univariate or multivariate analysis of the outliers or normality distribution is determinant to define which underlying methods (for e.g. extraction) will be applied in the factor analytical process.

7.3.1.2. Theoretical Framework of the Model

The model was initially conceived in order to assess the predictor role of the latent variable[7] Research and Development (R&D) on countries' economy and healthcare. Specifically, we hypothesized that *R&D would have a direct effect over the Economy (GDP) and General Health indicators (birth rate, death rate, death rate, and immunization efforts).* The relation between these variables has been separately reported in a number of studies [Bessette, 2003, Daraio et al., 2011, Hanushek and Woessmann, 2010, Kilpeläinen et al., 2012]. Therefore, we gathered a core set of variables (as described in Section 7.3.2) related to each of these factors, which represented the situation affecting all countries. However, for our initial model we only kept the data displayed at the country level, excluding other (although interesting) information at regional level.

7.3.1.3. Structural Equation Modeling

Structural Equation Modeling [Hox, 1998, Kline, 2011] is a method that has been used in health sciences [Hays et al., 2005], economic research [Hair et al., 2012] to model causal relations among latent and observed variables. This method evaluates the relation between latent variables. For example, in this study we argue that the general concept of economic performance is only possible to explain through a latent variable specified by other observed variables such as Gross Domestic Product (GDP) etc. Hence, we used SEM to test the outlined hypothesis of a conceptual model based on the effect of R&D on the economic and healthcare situation of countries all over the world.

Our SEM was tested by the jigsaw method [Bollen, 2000]. This procedure expects the adequacy of the measurement variables (latent variables that will enter the model) into isolated confirmatory factor analysis models. By doing this measurement before the structural equations, we are able to define the model's identification with the latent variables before testing. Therefore a two step strategy is defined to design a SEM [Kline, 2011]:

[7] Latent variables are those that cannot be measured directly, but is an underlying concept involving other observed variables (variables measured directly).

Step 1. The first step in an EFA analysis is to specify the latent variables through a sequence of EFA and Confirmatory Factor Analysis (CFA) in a way that an EFA is performed to detect latent factors and CFA to confirm its structure. If the latent structure does not show adequate indicators then re-evaluate the EFA by making a sequence of EFA-CFA-EFA-CFA until an adequate measurement model is obtained. Then one must define the extraction[8] and rotation methods[9]. Once the non-normal distribution of the data is detected, EFA is performed with Principal Axis extraction method, which fits this data distribution better. A Promax (Oblique) rotation was performed because we believed that the latent variables would be correlated [Costello and Osborne, 2005]. The obtained factor loading values[10] above 0.30 were considered acceptable. Models developed by EFA were then tested through CFA sequentially until an adequate model was obtained.

CFA procedure evaluated the model adjustment and adequacy through fitness indicators, factor loadings and individual item reliability. Weighted Least Square was the estimation method used, due to non-normal multivariate normality that was obtained. The indicators used to assess the fitness of the model were (cf. Table 7.4): (i) X2/Df (P-valor): chi-square (ii) Root Mean Square Error of Approximation (RMSEA): values inferior to 0.08 are considered as acceptable fit and 0.05 as a adequate fit; (iii) Tucker-Lewis Index (TLI): acceptable fit with values superior to 0.90; (iv) Comparative Fit Index (CFI): values superior to 0.90 are accepted as adequate fit and 0.95 as good fit); (v) Standardized Root Mean Square Residual (SRMR): Standardized Root Mean Square Residual

Step 2. SEM was applied to test the hypothetical model using the same indicators as described in the measurement model evaluation (Step 1) as well as factor loadings and individual item reliability. The path coefficients were interpreted as: small effect for loadings <0.10, medium effect for loading until 0.30 and high effect for loadings >0.50 [Kline, 2011]. Data analysis was performed through R Language Statistical Software version 3.0 [Hornik, 2008], with the specific SEM analysis developed with the SEM package [Fox, 2006].

7.3.1.4. Geographical Information System

After modeling, we anticipated that data analysis might be affected by the geographical location. Therefore, we used Exploratory Spatial Data Analysis (ESDA) and the software package GeoDa version 0.9.5-i (Spatial Analysis Laboratory, University of Illinois, Urbana-Champaign, IL, USA) to determine measures of global spatial autocorrelation and local spatial autocorrelation [Anselin et al., 2010]. To evaluate the existence of spatial autocorrelation, we defined a spatial weight matrix - W. This matrix allows for

[8] Extraction method is the statistical approach applied to extract the amount of variance of the data that is shared by the variables revealing the latent constructs.

[9] Rotation technique is used to clarify which variables load into each latent construct.

[10] Factor loading is a metric that indicates the amount of contribution of that specific factor to explain the variance in the observed variable.

the measurement of non-random associations between the value of a variable observed in a given geographical unit with the value of variables observed in the neighboring units. Furthermore, we used the binary matrix-type Queen, which attributes a value of one for neighbors in any spatial location within the analyzed region [Anselin, 2010].

Additionally, we calculated spatial autocorrelation evaluating the effect of hindex in each country in association with GDP and Health Outcomes indicators for each country using the Global Morań index (I) for univariate and bivariate analysis [Anselin, 2010, Perobelli and Haddad, 2006]. This index measures both the spatial autocorrelation and the weighted neighborhood matrix, indicating that the mortality rates of a given region might be similar to those of neighboring regions. Values of Morań's I vary between -1 and +1. Values greater or smaller than the expected value of Morań's I [E (I) = -1/(N - 1)] indicate a positive or negative autocorrelation, respectively. If the value of Morań's I is 0 (zero), the region is considered to have spatial independence [Anselin, 2010, Perobelli and Haddad, 2006].

Morań's I values between 0 and +1 indicate positive spatial association (direct). This indicates that regions with high Morań's I values for the variable in question are surrounded by regions which also have high variable values (high/high). Similarly, regions with low variable values are surrounded by neighbors which also have low variable values (low/low). Negative values of Morań's I (from 0 to -1) represent negative spatial association (reverse). Therefore, regions with high Morań's I values are surrounded by regions with low variable values, while regions with low Morań's I variable values are surrounded by neighbors with high variable values [Anselin, 2010, Druck et al., 2004, Perobelli and Haddad, 2006].

To identify patterns of spatial association that were significant and specific to each analyzed area, we used Local Indicators of Spatial Association (LISA). LISA allowed us to identify the existence of spatial clusters, or sites with high or low values for the analyzed variables, ultimately determining regions that can contribute to spatial autocorrelation [Perobelli and Haddad, 2006].

7.3.2. Datasets, Variables and Data Extraction

For the HER Observatory, we choose four relevant datasets (to address L3 Coverage – Section 7.2.2), namely (i) LinkedCT, (ii) USPTO Linked Patents, (iii) Scimago and (iv) World Bank. We describe each dataset along with the details of the total of 17 variables chosen from them. For data extraction, we use the SPARQL package for R [R Core Team, 2014], which allows us to directly run SPARQL queries against a SPARQL endpoint within R and retrieve results [van Hage et al., 2014]. The code to extract data using R and SPARQL is available online[11].

LinkedCT. LinkedCT[12] is the Linked Data version of ClinicalTrials.gov. ClinicalTrials.gov is a database of statistical studies providing evidence for the effectiveness of

[11] https://github.com/amrapalijz/R-LOD-SEM/blob/master/RSPARQL
[12] http://linkedct.org

a treatment option (most often a drug/medication for a particular disease). LinkedCT contains information about $62,000$ governmentally as well as privately funded clinical trials conducted around the world, amounting to about 10 million triples.

The main information we use from LinkedCT is the total number of trials per country for each disease, which indicates the amount of R&D performed for each disease. Listing 7.1 shows the SPARQL query to retrieve the number of trials per country for the year 2000.

```
#Endpoint: http://db0.aksw.org:8895/sparql
PREFIX linkedct:<http://data.linkedct.org/vocab/resource/>
SELECT DISTINCT ?countryname ?conditionname count(distinct(?trial)) AS ?NoOfTrials
FROM <http://data.linkedct.org>
WHERE {
?trial     a                          linkedct:trial .
?trial     linkedct:trial_condition ?condition .
?condition rdfs:label                 ?conditionname .
?trial     linkedct:trial_location_countries ?country .
?country   rdfs:label                 ?countryname .
?trial     linkedct:completion_date ?date .
FILTER regex(?date, '2000')}
GROUP BY ?countryname ?conditionname
ORDER BY DESC(count(distinct(?trial)))
```

Listing 7.1: Extraction of data from LinkedCT

USPTO Linked Patents. The USPTO[13] is part of the US department of Commerce and grants patents to businesses and inventors for their inventions in addition to registration of products and intellectual property identification. As of December 2011, more than 8.7 million patents have been issued and 16 million applications have been received. A total of 7 million patents dated from 1790 onwards are available. Additionally, Google has also made all the patents available for download in XML format[14]. We converted this bulk of data (spanning 10 years) from XML to RDF conforming to the Linked Data principles[15].

```
#Endpoint: http://us.patents.aksw.org/sparql
PREFIX patent:<http://us.patents.aksw.org/schema/>
SELECT COUNT(DISTINCT (?s)) AS ?NoOfPatents ?countryname
FROM <http://uspatents.aksw.org/>
WHERE
{?s       patent:country    ?country .
?country  rdfs:label        ?countryname .
?s        dcterms:date      ?year .
FILTER regex(?year, '2000-01-01'^^xsd:date)}
```

Listing 7.2: Extraction of number of patents per country for the year 2000 from the USPTO Linked Patents dataset.

The information we extracted from the patents data is the number of patents per country per year, which is an indicator of R&D in terms of innovation in each country. Listing 7.2 shows the SPARQL query to retrieve the data from the USPTO Linked Patents dataset.

[13] http://www.uspto.gov/

[14] http://www.google.com/googlebooks/uspto-patents-grants-text.html

[15] Available at http://us.patents.aksw.org.

Scimago. The SCImago Journal & Country Rank[16] is a portal that reports on the journals and country scientific indicators calculated from the information obtained from the Scopus database. In particular, the Global SCImago Institutions Rankings (SIR) report takes into account those organizations from any country, which has at least 100 documents published each year.

```
#Endpoint: http://db0.aksw.org:8895/sparql
PREFIX scimago:<http://scimago.org/>
SELECT DISTINCT ?countryname ?year ?noOfDocs ?hindex
FROM <http://scimago.org>
WHERE {
?s       rdf:type        scimago:journalRanking .
?s       scimago:country ?country .
?country rdfs:label      ?countryname .
?s       scimago:year    ?year .
?s       scimago:docs    ?noOfDocs .
?s       scimago:hIndex  ?hindex .
FILTER (?year=<http://reference.data.gov.uk/id/year/2000>) }
```

Listing 7.3: Extraction of the number of documents and hindex per country for the year 2000 from the Scimago dataset.

The variables we utilize from this dataset are:

- Total number of published documents per country per year (noD)
- hindex - country's number of articles (h) that have received at least h citations (hindex)

Since this dataset was not already available as Linked Data, we converted the selected information from CSV to RDF using LODRefine[17]. Listing 7.3 shows the SPARQL query to retrieve the number of documents and hindex per country per year from the Scimago dataset.

WorldBank. The World Bank[18] is an international financial institution that collects and processes large amounts of data on the basis of economic models and makes them openly available[19]. The available data covers a wide variety of topics such as Agriculture and Rural Development, Education, Health, Public and Private Sector, Science and Technology etc. The World Bank data has been converted and published as LD and is available at http://worldbank.270a.info. In particular, the World Development Indicators, which present the most current and accurate global development data accessible, are available as RDF[20]. The variables we choose are:

- Adolescent Fertility Rate (AFR), which reports the number of births per $1,000$ women aged $15–19$ (code: SP.ADO.TFRT).

[16] http://www.scimagojr.com/countryrank.php
[17] http://code.zemanta.com/sparkica/
[18] http://www.worldbank.org/
[19] http://data.worldbank.org/
[20] http://worldbank.270a.info/classification/indicator.html

- Birth Rate (BR), which indicates the crude birth rate i.e. the number of live births occurring during the year per $1,000$ population. This indicator is estimated at the middle of the year (code: SP.DYN.CBRT.IN)[21]
- Death Rate (DR), which is the number of crude deaths occurring during the year per $1,000$ population also estimated at the middle of the year (code: SP.DYN. CDRT.IN).
- GDP, which is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the product. Data is CD).
- Health Expenditure Public (HEP) (% of total expenditure) reports the recurrent and capital spending from government (both central as well as local) budgets, external borrowings and grants and social (or compulsory) health insurance funds. The total health expenditure is the sum of public and private health expenditure, which covers the provision of health services as mentioned in the indicator "Health expenditure per capita" (code: SH.XPD.PUBL).
- High-technology Export (HET) (% of manufactured exports) are the products with high R&D intensity, such as in aerospace, computers, pharmaceuticals, scientific instruments, and electrical machinery (code:TX.VAL.TECH.MF.ZS).
- Immunization DPT (IDPT) (% of children ages $12–23$ months) measures the percentage of children between the ages of $12–23$ months who have received vaccinations before 12 months or at any time before the survey. After receiving three doses of vaccine, a child is considered adequately immunized against diphtheria, pertussis (or whooping cough), and tetanus (DPT) (code: SH.IMM.IDPT).
- Immunization Measles (IM) (% of children ages 12 - 24 months) measures the percentage of children between the ages of $12–23$ months who have received vaccinations before 12 months or at any time before the survey. After receiving one dose of vaccine, a child is considered adequately immunized against measles (code: SH.IMM.MEAS).
- Incidence of Tuberculosis (ITB) (per $100,000$ people) is the estimated number of new pulmonary, smear positive and extra-pulmonary tuberculosis cases, which also includes patients with HIV (code: SH.TBS.INCD).
- Mortality Rate (MR), infant (per $1,000$ live births) is the number of infants dying before reaching one year of age, per $1,000$ live births in a given year (code: SP.DYN.IMRT.IN).

[21] The code is to be used when querying for the particular variables by replacing the value for the "indicator" property in the SPARQL query Listing 7.4.

- Public Spending on Education (PSE), total (% of govt expenditure) reports the total public education expenditure (current and capital) expressed as a percentage of total government expenditure for all sectors in a given financial year. This public expenditure on education includes the government spending on educational institutions, both public and private, education administration and subsidies for private entities such as students or households etc (code:SE.XPD.TOTL.GB.ZS).
- Research and Development Expenditure (RGDP) (% of GDP) reports the expenditures for research and development, which are the current and capital expenditures i.e. both public and private on creative work undertaken systematically to increase knowledge. This work includes knowledge of humanity, culture, society as well as the use of knowledge for new applications. R&D covers basic, applied and experimental research and development (code: GB.XPD.RSDV.GD.ZS).
- Researchers in R&D (RRP) (per million) reports the number of professionals engaged in the conception or creation of new knowledge, products, processes, methods, or systems and in the management of the projects (code: SP.POP.SCIE.RD.P6).

Listing 7.4 shows the SPARQL query to retrieve the variables from the World Bank dataset.

```
#Endpoint: http://worldbank.270a.info/sparql
SELECT ?label ?obsValue ?year
WHERE { ?observationURI
     property:indicator       indicator:SP.ADO.TFRT;
     sdmx-dimension:refArea   ?refAreaURI;
     sdmx-dimension:refPeriod ?year;
     sdmx-measure:obsValue    ?obsValue .
?refAreaURI a                 dbo:Country .
?refAreaURI skos:prefLabel    ?label .
FILTER (?year=<http://reference.data.gov.uk/id/year/2000>)}
```

Listing 7.4: Extraction of the Adolescent Fertility Rate variable from the World Bank dataset for the year 2000.

7.4. Results

In this section, we report the results of the data quality assessment performed on the four datasets (Section 7.4.1) and also results of applying the SEM methodology to the four datasets (Section 7.4.2).

7.4.1. Data Quality Assessment

The data quality assessment of the four datasets was performed using the R2RLint tool (described in Chapter 6) with the 13 metrics (described in Section 6.1). The results are available online[22] containing quality scores for the introduced metrics and references to the actual data that caused quality problems, thus enabling the user to locate and

[22] http://pwestphal.aksw.org/dq/

Table 7.1.: Results of assessing the quality of the LinkedCT dataset for the 2010 and 2013 versions. * These values are errors reported per 100,000 triples.

Abbr.	Dimension	Metric	v2010	v2013
A1	Availability	Dereferenceability of the URIs	342.78*	1,040.18*
C1	Completeness	Interlinking Completeness	0.1606	0.0772
I1	Interlinking	External Same-as Links	0.000248289	0.0
CO3	Consistency	Well-placed Classes	8,472.90*	866.61*

fix these issues immediately. It is important to note that each result is associated with a time stamp, as to when the assessment was performed. Thus, certain quality issues such as dereferenceability of URIs are subjective to the availability of the SPARQL endpoint at that particular time. Also, it is to be noted that the quality assessment was performed only on a random sample for the USPTO Patents and WorldBank datasets whereas LinkedCT and Scimago were assessed entirely.

7.4.1.1. LinkedCT

The quality assessment of the LinkedCT dataset was performed in two stages. First, the dataset used for the ReDD-Observatory, that is version 2010[23] was assessed. Then, the data from 2013[24] was assessed, which is used in the current project. The results of both the assessments are reported in Table 7.1 and discussed here:

The data quality assessment results for the 2010 LinkedCT dataset are:

- A1: Dereferenceability of the URIs – There were 33,608 errors reported, which were mainly due to the unavailability of external resources such as those from DailyMed[25] or DrugBank[26] datasets. However, as mentioned earlier, the results are provided with a timestamp thus suggesting that these external datasets might not have been available at that particular time.
- C1: Interlinking Completeness – The interlinking completeness score returned was 0.160606, which is relatively high, showing a good interlinking to external datasets.
- I1: External `owl:sameAs` links – The external same as links assessment produced a score of 0.000248289, which is low due to the presence of very few `owl:sameAs` links to external datasets.
- CO3: Well-placed Classes – There were 830,738 erroneous occurrences reported. The resources that occurred with this type of error were:

[23] http://www.cs.toronto.edu/~oktie/linkedct/linkedct-dump-2010-02-10.tar.gz
[24] www.cs.toronto.edu/~oktie/linkedct/linkedct-dump-2013-10-01.nt.bz2
[25] http://datahub.io/dataset/fu-berlin-dailymed
[26] http://datahub.io/dataset/fu-berlin-drugbank

  - `linkedct:collaborator_agency`
  - `linkedct:trials`
  - `linkedct:intervention`
  - `linkedct:location`
  - `linkedct:overall_official`
  - `linkedct:condition`
  - `linkedct:primary_outcomes`
  - `linkedct:reference`
  - `linkedct:results_reference`
  - `linkedct:secondary_outcomes`

The results from assessing the quality of the 2013 LinkedCT dataset are:

- A1: Dereferenceability of the URIs – There were $5,257$ errors reported applying the dereferenceability metric to a random sample of the LinkedCT dataset of about $500,000$ triples. Again, these errors were due to the unavailability of the external resources from DailyMed and DrugBank at the time of assessment.
- I1: Interlinking Completeness – The interlinking completeness score was 0.077257.
- CO3: Well-placed Classes – The number of violations regarding the Well-placed Classes metric decreased to $297,827$. The resources that caused these errors were:
  - `linkedct:condition_browse`
  - `linkedct:eligibility`
  - `linkedct:intervention_browse`
  - `linkedct:oversight_info`
  - `linkedct:responsible_party`
  - `linkedct:sponsor_group`

Table 7.2.: Results of assessing the quality of USPTO Linked Patents, Scimago and WorldBank datasets.

Abbr.	Dimension	Metric	USPTO Patents	Scimago	WorldBank
C1	Completeness	Interlinking Completeness	0.000161	0.0	0.000476
I1	Interlinking	External sameAs Links	0.0607	0.0	0.0
CO3	Consistency	Well-placed Classes	0	1	0

7.4.1.2. USPTO Linked Patents

Results of assessing the USPTO Linked Patents dataset are:

- C1: Interlinking completeness score returned was 0.000161, which suggests that the dataset is interlinked with only a low number of external datasets.
- I1: External `owl:sameAs` links returned a score of 0.0607. Even though its actual value is low, the score is considered to be medium (with regards to the definition of the corresponding metric) thus indicating the presence of a considerable number of *same as* interlinks to external datasets.

7.4.1.3. Scimago

As a result of assessing the quality of the Scimago dataset, the only error that was returned was of defining the resource `<http://scimago.org/country>` as a class (CO3). This was fixed before using this dataset in calculating the Observatory results.

7.4.1.4. WorldBank

The result of assessing the WorldBank dataset is:

- C1: Interlinking Completeness – The score returned was 0.0005, which is relatively low and suggests that the dataset is not well interlinked with external datasets.

The results of assessing the quality of the USPTO Linked Patents, Scimago and WorldBank datasets are reported in Table 7.2. After analyzing the table, it can be seen that the major issues lie in the interlinking completeness of the datasets whereas there were no significant issues reported for the other quality metrics. In particular, the interlinking completeness for World Bank was higher than the USPTO Patents dataset, even though the interlinking completeness score was low. However, there were a significantly large number of `owl:sameAs` links.

7.4.2. HER Observatory

In this section, we first describe the results of constructing the parts of the model, in particular, determining the latent and observed variables. Then, we describe the process of choosing the best fit for the model to reach the best possible adequacy and theoretical reasoning. The R script for this process is available at https://github.com/amrapalijz/R-LOD-SEM/blob/master/sem_script.R.

The first task was to integrate all the variables and as a result, from the original 196 countries that entered the analysis, several were excluded due to data incompleteness. Thus, only 20 countries constituted the final sample. Also, after excluding the variables that contained data quality problems, a total of 11 variables (from the initial 17) were

Table 7.3.: Descriptions and abbreviations used for each of the 11 observed variables belonging to each of the latent variables of the SEM.

Latent variables	Observed variables	Abbreviation
General Health Outcomes	Adolescent fertility rate (births per $1,000$ women ages 15–19)	AFR
	Birth rate, crude (per $1,000$ people)	BR
	Death rate, crude (per $1,000$ people)	DR
	Health expenditure public (% of total health expenditure)	HEP
	Immunization DPT (% of children ages 12 – 23 months)	IDPT
	Immunization measles (% of children ages 12 – 24 months)	IM
	Mortality rate, infant (per $1,000$ live births)	MR
Research and Development (R&D)	Number of articles (h) that have received at least h citations	hindex
	Total number of published documents per country per year	nOD
	High-technology export (% of manufactured exports)	HTE
Economic performance	GDP per capita (current US$)	GDP

included in the analysis as listed in Table 7.3. As a consequence, out of the four datasets, only two datasets were used in the analysis i.e. Scimago and WorldBank. LinkedCT and USPTO Linked Patents were excluded because of incomplete and inconsistent data. This significantly limited the results as further potential research questions could not be answered (discussed in Section 7.5). The time range consisted of 12 years i.e. from 1999 to 2010.

The first step in constructing the SEM was to choose the latent variables that are relevant for the formulated hypothesis. All the variables extracted from the datasets were conceptualized as parts of the construct of the model's theoretical framework. However, in order to develop a latent variable we must assess how the variances of each variable relate to the existence of an underlying latent factor. Therefore, we applied a set of EFAs and CFAs to reach the best possible factor structure to apply to the model [Kline, 1994].

Correlations between the variables were analyzed to assess the pattern of relations and possible clusters amongst the variables in the model. It is noteworthy that GDP was moderately to highly related to all variables, while hindex and HET had strong correlations (0.57) as depicted in Figure 7.1. However, the number of documents was

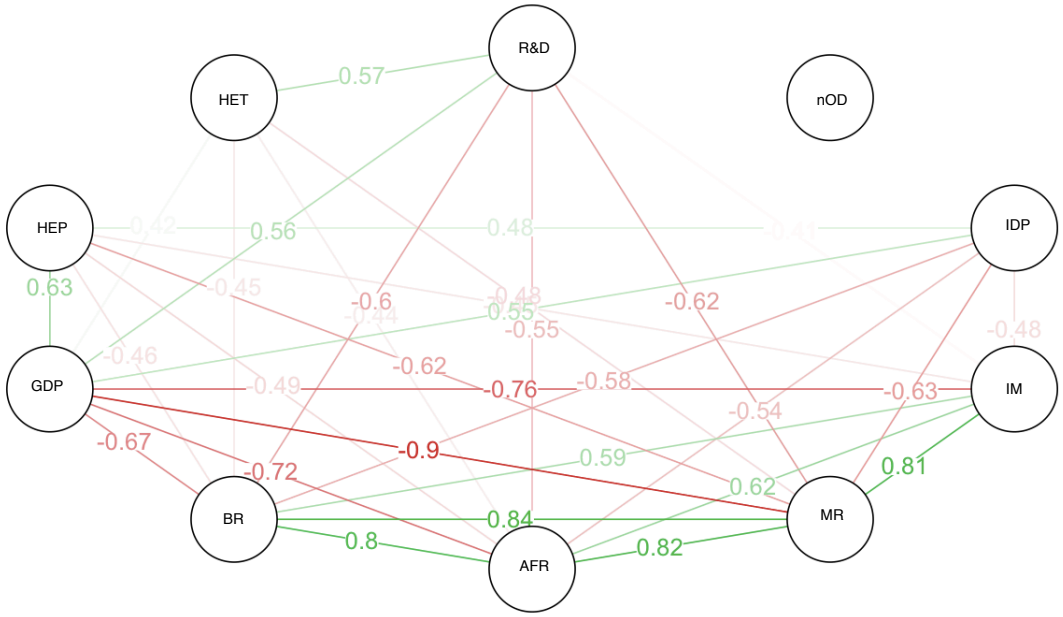


Figure 7.1.: Network display of correlations between the variables added to data analysis. Proximity of the nodes (circles) and edges thickness (lines) indicate correlation level, with closer nodes and thicker edges been stronger correlations.

not related to the other R&D variables, not providing enough strength to a latent model. General Health Outcomes were all moderate to strongly correlated among them (R <0.40), suggesting the presence of a latent model.

To confirm the existence of the latent models, EFA models were conducted to explore R&D and General Health Outcomes variables. Although the correlation analysis showed that only two variables (hindex and HET) would compose the model, we tried a one-factor model with all three indicators for R&D, but results indicated a poor fit and did not provide evidences to the model.

Eigen values and screenplot analysis indicated the possibility of one or two latent factors, to the General Health Outcomes latent variable. Therefore, different EFAs were applied to test for one, two, and three factors structures. One factor model solution showed better indicators explaining 60% of a variance of the variables in the dataset. However, factor loadings and commonalities indicated problems with the variables DR (Factor Loading (FL) <0.30 and Communality (H) <0.50), which meant that these variables were not contributing enough to the latent factor structure specification. After excluding these variables from the model, we decided to test how the factor structures would fit in the CFA models with and without both variables.

CFA models were developed for General Health Outcome latent construct, all variables had adequate factor loadings and fit indicators, suggesting a good adequacy to the model. In summary, during the model specification phase (Step 1) we noticed that one of the observed variables was not adjusting to latent variable model, which might have influenced the final SEM. Thus from the initial 7 observed variables of the General

Health Outcomes we ended up with a model constituted by 6 observed variables.

Initially we developed a model (Model A, Table 7.4) with only one exogenous variable (R&D as a latent variable- causing the effect) and two endogenous variables (GDP and General Health Outcomes receiving the effect). Exogenous variables are those that originate an effect (path-arrow) to other variables in the model, while endogenous variables are those that receive the effect (path). However, this model showed poor fitness indicators (Table 7.4) and several highly correlated residuals. In order to improve this model's fit, we needed to fix the covariance of errors between variables that would make the model loose its meaning.

Analyzing the residuals behavior and based on the Step 1 analysis, we opted to test a model without a latent variable for R&D and insert each hindex, nOD, HET as observed variables in the model. Thus, the second model (Model B, Table 7.4) investigated a direct path from each hypothesized R&D observed variable. This model showed some problems in its fit indicators (Table 7.4).

In order to improve the model specification and fit indicators, we assessed the modification indices, which are indications of the extent of the model's fit results that will be improved by adding an additional path to the model. Modification indices suggested the presence of a covariance between the observed variables that constituted General Health Outcomes, therefore affecting the models performance. Then, a third model (Model C, Table 7.4) was tested (Figure 7.2), drawing a correlational (double headed arrow) path between the observed variables BR and AFR and fixing the covariance errors.

Table 7.4.: CFA Fit Indicators and their respective measurements for all the four models. Model C is adopted in this study. The measurements are: (i) X2/Df (P-valor): chi-square (ii) Root Mean Square Error of Approximation: values inferior to 0.08 are considered as acceptable fit and 0.05 as a adequate fit; (iii) TLI: acceptable fit with values superior to 0.90; (iv) Comparative Fit Index: values superior to 0.90 are accepted as adequate fit and 0.95 as good fit and (v) Standardized Root Mean Square Residual

	Model A	Model B	Model C	Model D
X2/Df (P-valor)	501.703/34 (0.001)	694.090/35 (0.001)	480.867/33 (0.001)	629.222/35 (0.001)
RMSEA (CI 95%)	.13 (.12;.14)	.15 (.14;.16)	.13 (.10;.14)	.14 (.13;.15)
TLI	.86	.81	.90	.82
CFI	.89	.85	.90	.87
SRMR	.05	.11	.08	.08

Model C was the one with *best fitness indicators* except for the RMSEA, which was above the proposed cutoff point [Schermelleh-Engel et al., 2003, Hox, 1998]. This model was able to explain 26.1% of the variance in General Health Outcomes and 43.8%

of the variance in GDP. Path coefficients showed that only hindex had moderate to high effect on GDP (0.66) and General Health Outcomes (−0.42). HET had small effect on GDP (0.31). These values might be understood in the same meaning as a regression coefficient (although they are not the same), thus the value varies (generally) from 0 to 1, indicating the size of the effect for that specific path. Positive and negative signs indicate the reciprocity of the relation, thus positive values show proportional modulation while negative values indicate inverse relations.

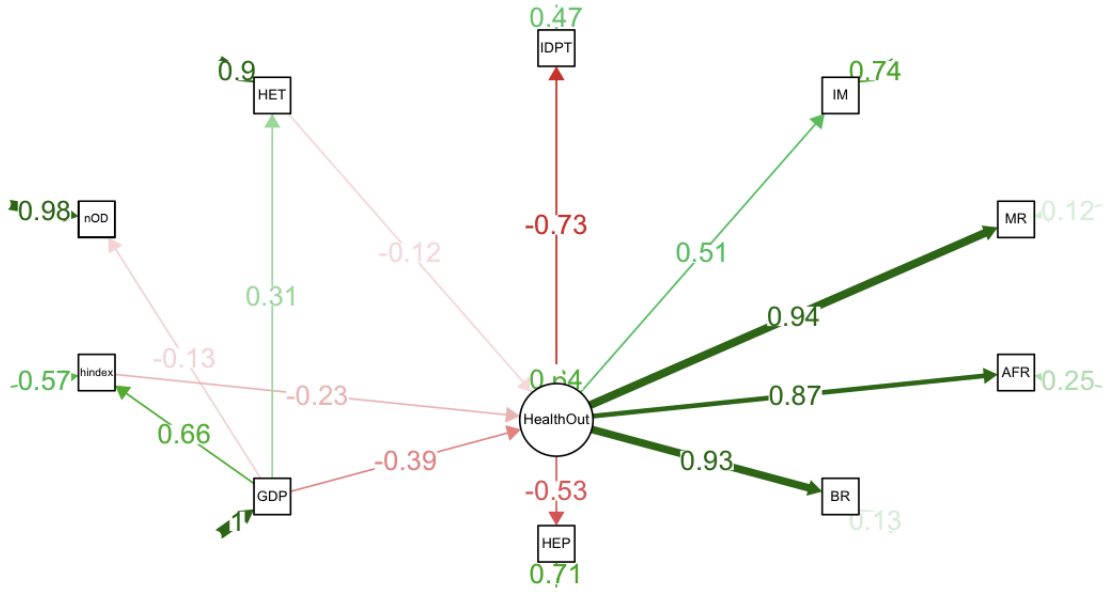


Figure 7.2.: Structural equation model of the influence of R&D on economic and healthcare performance. Values on the arrows connecting latent variables are the path coefficients and indicate the effects weight. Positive and negative signs indicate the reciprocity of the relation, thus positive values show proportional modulation while negative values indicate inverse relations.

Finally a fourth model (Model D, Table 7.4) was tested, mainly for validation purposes, in order to show that our model had a better chance of explaining the relations among the latent variables. This model had GDP as the main predictor (as an exogenous variable), R&D variables as the mediators (endogenous and exogenous) and General Health Outcomes as the outcomes. The rationale here is that GDP is the main predictor of the outcomes and this effect can be mediated by R&D variables. However this model did not show a good fitness indicator and the modifications needed to improve its specification could not be accepted because they did not demonstrate theoretical coherence. Therefore, we decided that Model C was the best possible solution to the relations between the latent variables we developed.

After adjusting and modifying the model to find the best possible fit, and also comparing with a different predictive model possibility, we verified that R&D positively influences the economical and healthcare systems in the countries. This result supports our initial hypothesized model and suggest that the quality in R&D positively influ-

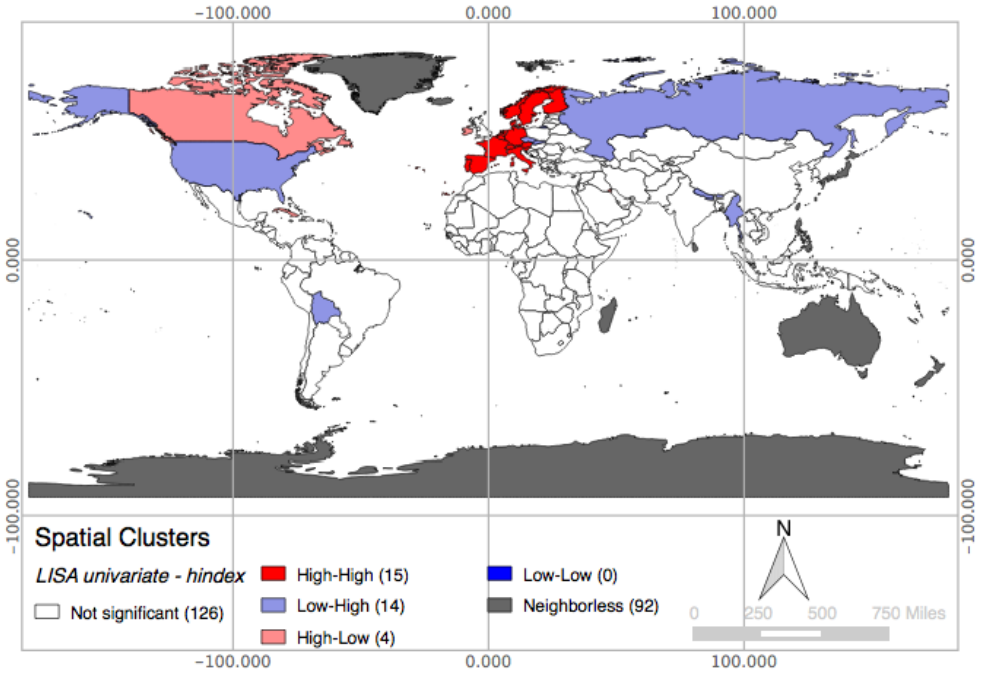


Figure 7.3.: Exploratory spatial analysis by country through local indicators of spatial association (LISA) univariate analysis: cluster formation according to hindex (R&D) rate. Countries with high Morań's I values are surrounded by countries which also have high variable values (high/high), in this case 15 countries. Similarly, regions with low variable values are surrounded by neighbors, which also have low variable values (low/low), in this case 0.

ences the economic status in terms of GDP, and also negatively influences the birth rate, immunization and adolescent fertility. Contradicting our hypothesis, the amount of R&D (number of documents) did not directly influence the countries' economic performance neither health system indicators. However this points to the possibility of other covariants inflicting this model's relations enhancing or impairing the effect of R&DD over economic development.

We found a positive spatial autocorrelation regarding hindex i.e. R&D (I = 0.3250, p = 0.001)[27]. This shows that countries with a high level of hindex are surrounded by countries with high hindex rates, indicating that countries with high hindex are surrounded with other high hindex countries as shown in Figure 7.3.

There was a significant positive association (see Figure 7.4) between each of three indicators and hindex: HEP (I = 0.1455, P = 0.002), HET (I = 0.1109, P = 0.005) and GDP (I = 0.1290, P = 0.002). In addition, six indicators presented significant negative associations (see Figure 7.4) with hindex: AFR (I = −0.1430, P = 0.001), BR (I = −0.1072, P = 0.001), DR (I = −0.0432, P = 0.0230), IDPT (I = −0.1042, P = 0.001), IM (I = −0.1303, P = 0.001) and MR (I = −0.1571, P = 0.001). Although significant,

[27] I is Global Morań index, see subsubsection 7.3.1.4 for details

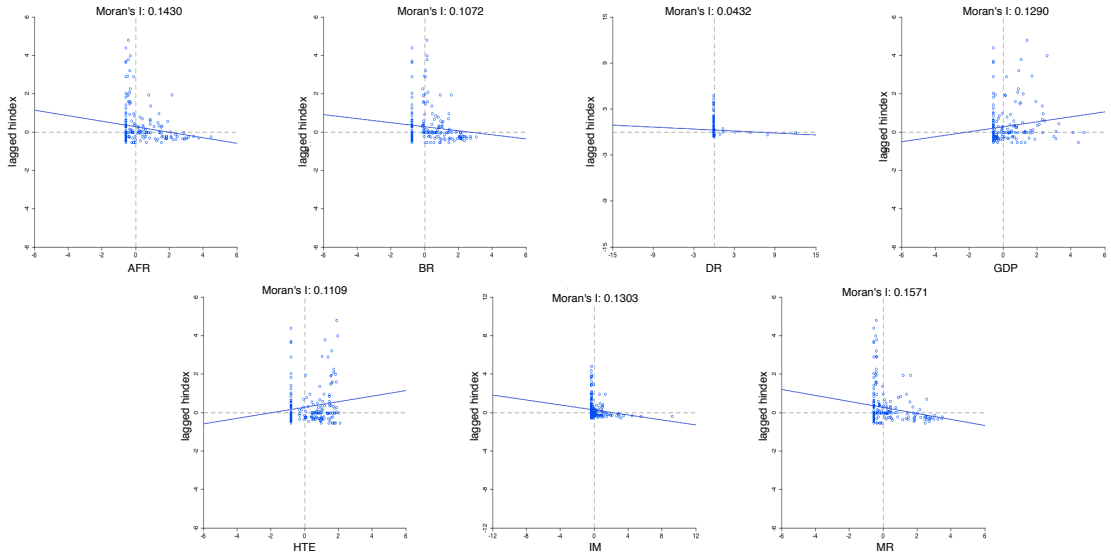


Figure 7.4.: Morań's diagram of dispersion (bivariate analysis). Analysis of Economic and General Health Outcome variables of each country (X axis) with the weighted average hindex of the neighbor countries (Y axis).

autocorrelations were weak in terms of intensity, suggesting that other confounders might be able to improve this spatial association.

It is noteworthy that these results indicate a discrepancy between countries when looking into the spatial association of hindex and other variables. Although autocorrelation values are low, we understand that countries with high hindex have less Health Outcome indicators and higher GDP, but are surrounded by countries with different characteristics, depicting the discrepancy in development. Also, these low values of association indicate that other variables might be affecting these associations, such as socioeconomic status of the continent. Europe, for instance had a positive spatial association between hindex in the countries, while other continents did not find the same pattern of clustering.

7.5. Summary, Impact, Limitations and Future Work

Summary. In this chapter, we showcased the usefulness of Linked Data to evaluate the impact of research and development on economic and healthcare performance. The analyses reported in this chapter have been based on four datasets integrated into the HER Observatory. We showed that the assessment of data quality is very important for any data analysis. Moreover, we showed that in comparison to the previous crowdsourcing quality assessment methodologies (Chapter 4 and Chapter 5), which are not only error-prone and time-consuming, this semi-automated methodology is feasible to perform quality assessment of datasets. Additionally, the user is involved in the assessment as she

can choose which metrics apply to her use case and also is provided with interpretable results. By using Structural Equation Modeling, we showed that we can link structured information to evaluate the impact of R&D on economic and healthcare performance at the country level. Our results show that R&D positively influences countries' economic and healthcare systems.

Impact. Measuring the economic and healthcare performance of countries permits policy makers to determine whether research strategies developed by countries are aligned with their respective healthcare needs. This information is critical for guiding and shaping policies that facilitate the destruction of inequalities, planning health systems, improve healthcare delivery, promote and sustain population welfare, allocating budgets for R&D (set spending priorities), monitor progress and evaluate what works and what does not [Murray et al., 2004, Schlotthauer et al., 2008]. This analysis, in turn, helps improve the technology and methods for societal progress and motivates governments to collect and analyze useful data and compare assessments of inputs, service delivery and achievements for economic and health outcomes. Revisiting our user scenario introduced in Section 1.2, Ms. Sharma can analyze and interpret the results from our use case and measure the healthcare performance of India. This allows her to determine whether research strategies developed by India are aligned with the respective healthcare needs.

Limitations. We encountered several issues in our current approach. Firstly, there was a minimal amount of post-processing required to unify the labels of the data, for example, the dates are sometimes only provided as URIs. Secondly, the quality of the datasets was a major hindrance, in particular we had to exclude several variables from our analysis, for incompleteness and inconsistency issues. As a result, two (out of the four) datasets were not used for calculating the correlations among the indicators, leading to the loss of valuable information. In particular, due to the exclusion of the LinkedCT and USPTO Linked Patent datasets, interesting research questions, such as the impact of R&D on the healthcare (in terms of clinical trials) and innovation (in terms of patents), have not been answered. Also, with the huge amount of variables involved in the analysis, the analysis was computationally exhaustive.

8. Related Work

This chapter contains related work for four different areas that are part of this thesis: (i) Data Quality Dimensions, (ii) Data Quality Assessment Efforts, (iii) Data Quality Assessment Tools and (iv) Calculation of Societal Progress Indicators

8.1. Data Quality Dimensions

There are a number of studies, which have identified, defined and grouped data quality dimensions into different classifications [Wang and Strong, 1996, Wand and Wang, 1996, Redman, 1997, Naumann, 2002, Batini and Scannapieco, 2006, Jarke et al., 2010]. Recently, there are a number of data quality dimensions that have been identified relevant to LD, namely, accuracy, timeliness, completeness, relevancy, conciseness, consistency [Bizer and Cyganiak, 2009]. Further quality criteria such as uniformity, versatility, comprehensibility, amount of data, validity, licensing, accessibility and performance are also introduced as means of assessing the quality of LD [Flemming, 2011]. The novel data quality aspects original to LD include, for example, coherence via links to external datasets, data representation quality or consistency with regard to implicit information. Furthermore, inference mechanisms for knowledge representation formalisms on the Web, such as OWL, usually follow an open world assumption, whereas databases usually adopt closed world semantics.

The research community is also still debating on the exact meaning of each dimension. Means to measure each dimension i.e. metrics are also not clearly defined for each dimension. Thus, in our study [Zaveri et al., 2015], we conducted a systematic literature review to gather all the relevant articles on data quality assessments specifically on LD, We qualitatively analyze the 30 identified studies and unify and formalize 18 data quality dimensions along with an example and identify 69 metrics to assess the quality of LD. Also, an example alongside each dimension helps to gain a clear picture of the dimension. Moreover, we provide several metrics from each dimension identified in the existing articles and also classify them into being either qualitatively (QL) or quantitatively (QN) assessed.

8.2. Data Quality Assessment Efforts

Web Data quality assessment frameworks. There are several efforts in developing data quality assessment frameworks in order to assess the data quality of LOD. These efforts are semi-automated [Flemming, 2011], automated [Guéret et al., 2012b] or

manual [Bizer and Cyganiak, 2009, Mendes et al., 2012b]. Other researchers analyzed the quality of Web [Cafarella et al., 2008] and RDF [Hogan et al., 2010] data. The second study focuses on errors occurred during the publication of Linked Data sets. Recently, a survey [Hogan et al., 2012] looked into four million RDF/XML documents to analyse Linked Data conformance. Even though these frameworks introduce useful methodologies to assess the quality of a dataset, either the results are difficult to interpret, do not allow a user to choose the input dataset or require a considerable amount of user involvement. In our experiment [Acosta et al., 2013], we used crowdsourcing to perform the evaluation because (i) none of the frameworks provided the granularity of quality criteria that we identified to be quality problems in DBpedia resources and (ii) we were interested in whether it was possible to use crowdsourcing to assess and thus improve the quality of a dataset.

An effort to assess the quality of web data was undertaken in 2008 [Cafarella et al., 2008], where 14.1 billion HTML tables from Google's general-purpose web crawl were analyzed in order to retrieve those tables that have high-quality relations. Additionally, there have been studies focused on assessing the quality of RDF data [Hogan et al., 2010] to report the errors occurring while publishing RDF data and the effects and means to improve the quality of structured data on the web. As part of an empirical study [Hogan et al., 2012] 4 million RDF/XML documents were analyzed, which provided insights into the level of conformance in these documents with respect to the LD guidelines. Even though these studies accessed a vast amount of web or RDF/XML data, most of the analysis was performed automatically and therefore the problems arising due to contextual discrepancies were overlooked. Another study aimed to develop a framework for the DBpedia quality assessment [Kreis, 2011]. In this study, particular problems of the DBpedia extraction framework were taken into account and integrated in the framework. However, only a small sample (75 resources) was assessed in this case and an older DBpedia version (2010) was analyzed. Considering that the DBpedia extraction framework was considerably enhanced since then, our efforts [Zaveri et al., 2013a, Acosta et al., 2013] shed light on the recent problems that hinder the quality of DBpedia.

Crowdsourcing Linked Data management tasks. There are already a number of efforts which use crowdsourcing focused on a specific type of task. For example, crowdsourcing is used for entity linking or resolution [Demartini et al., 2012], quality assurance and resource management [Wang et al., 2012] or for enhancement of ontology alignments [Sarasua et al., 2012] especially in Linked Data. However, in our case, we did not submit tasks to the popular internet marketplaces such as Amazon Mechanical Turk or CrowdFlower[1]. Instead, we used the intelligence of a large number of researchers who were particularly conversant with RDF to help assess the quality of one of the important and most linked dataset, DBpedia.

Several important Linked Data publication initiatives like DBpedia [Lehmann et al.,

[1] http://crowdflower.com/

2009] and contests have been organized, including challenges[2] to the European Data Innovator Award[3]. At a technical level, specific LD management tasks have been subject to human computation, including games with a purpose [Markotschi and Völker, 2010, Thaler et al., 2011] and microtasks. For instance, microtasks have been used for entity linking [Demartini et al., 2012] quality assurance, resource management [Wang et al., 2012] and ontology alignment [Sarasua et al., 2012].

8.3. Data Quality Assessment Tools

Out of the 30 core articles identified in our survey, 12 provide tools for data quality assessment. Thus, in this section, we compare these 12 tools based on eight different attributes (see Table 8.1 and Table 8.2).

Accessibility/Availability. In the tables, only the tools marked with a ✔are available to be used for quality assessment. The other tools are either available only as a demo or screencast (Trellis, ProLOD) or not available at all (TrustBot, WIQA, DaCura).

Licensing. Most of the tools are available using a particular software license, which specifies the restrictions with which they can be redistributed. The Trellis and LinkQA tools are open-source and as such by default they are protected by copyright, which is *All Rights Reserved*. Also, WIQA, Sieve, RDFUnit and TripleCheckMate are all available with open-source license: the Apache Version 2.0[4] and Apache licenses. tSPARQL is distributed under the GPL v3 license[5]. However, no licensing information is available for TrustBot, ProLOD, Flemming's tool, DaCura and LiQuate.

Automation. The automation of a system is the ability to automatically perform its intended tasks thereby reducing the need for human intervention. In this context, we classify the 12 tools into semi-automated and automated approaches. As seen in Table 8.1 and Table 8.2, all the tools are semi-automated except for LinkQA, which is completely automated, as there is no user involvement. LinkQA automatically selects a set of resources, information from the Web of Data (i.e. SPARQL endpoints and/or dereferenceable resources) and a set of triples as input and generates the respective quality assessment reports.

On the other hand, the WIQA, Sieve and RDFUnit require a high degree of user involvement. Specifically in Sieve, the definition of metrics has to be done by creating an XML file, which contains specific configurations for a quality assessment task. In case of RDFUnit, the user has to define SPARQL queries as constraints based on SPARQL query templates, which are instantiated into concrete quality test queries. Although it gives the users the flexibility of tweaking the tool to match their needs, it requires much time for understanding the required XML file structure and specification as well as the SPARQL language.

[2]For example: Semantic Web Challenge http://challenge.semanticweb.org/
[3]http://2013.data-forum.eu/tags/european-data-innovator-award
[4]http://www.apache.org/licenses/LICENSE-2.0
[5]http://www.gnu.org/licenses/gpl-3.0.html

Table 8.1.: Comparison of quality assessment tools according to several attributes.

	Trellis, Gil et al., 2002 [Gil and Ratnakar, 2002]	**TrustBot, Golbeck et al., 2003 [Golbeck et al., 2003]**	**tSPARQL, Hartig, 2008 [Hartig, 2008]**	**WIQA, Bizer et al., 2009 [Bizer and Cyganiak, 2009]**	**ProLOD, Böhm et al., 2010 [Böhm et al., 2010]**	**Flemming, 2010 [Flemming, 2011]**
Accessibility	–	–	✔	–	–	✔
Licensing	Open-source	–	GPL v3	Apache v2	–	–
Automation	Semi-automated	Semi-automated	Semi-automated	Semi-automated	Semi-automated	Semi-automated
Collaboration	Yes	No	No	No	No	No
Customizability	✔	✔	✔	✔	✔	✔
Scalability	–	No	Yes	–	–	No
Usability	2	4	4	2	2	3
Maintenance (Last updated)	2005	2003	2012	2006	2010	2010

Table 8.2.: Comparison of quality assessment tools according to several attributes.

	LinkQA, Gueret et al., 2012 [Guéret et al., 2012a]	**Sieve, Mendes et al., 2012 [Mendes et al., 2012b]**	**RDFUnit, Kontokostas et al.,2014 [Kontokostas et al., 2014]**	**DaCura, Feeney et al., 2014 [Feeney et al., 2014]**	**Triple CheckMate, Zaveri et al., 2013 [Zaveri et al., 2013a]**	**LiQuate, Ruckhaus et al., 2014 [Ruckhaus et al., 2014]**
Accessibility	✔	✔	✔	–	✔	✔
Licensing	Open-source	Apache	Apache	–	Apache	–
Automation	Automated	Semi-automated	Semi-automated	Semi-automated	Semi-automated	Semi-automated
Collaboration	No	No	No	Yes	Yes	No
Customizability	No	✔	✔	✔	✔	No
Scalability	Yes	Yes	Yes	No	Yes	No
Usability	2	4	3	1	5	1
Maintenance (Last updated)	2011	2012	2014	2013	2013	2013

The other semi-automated tools, Trellis, TrurstBot, tSPARQL, ProLOD, Flemming's tool, DaCura, TripleCheckMate and LiQuate require a minimum amount of user involvement. TripleCheckMate provides evaluators with triples from each resource and they are required to mark the triples, which are incorrect as well as map it to one of the pre-defined quality problem. Even though the user involvement here is higher than the other tools, the user-friendly interface allows a user to evaluate the triples and map them to corresponding problems efficiently.

For example, Flemming's Data Quality Assessment Tool requires the user to answer a few questions regarding the dataset (e.g. existence of a human-readable license) or they have to assign weights to each of the pre-defined data quality metrics via a form-based interface.

Collaboration. Collaboration is the ability of a system to support co-operation between different users of the system. From all the tools, Trellis, DaCura and TripleCheckMate support collaboration between different users of the tool. The Trellis user interface allows several users to express their trust value for a data source. The tool allows the

users to add and store their observations and conclusions. Decisions made by users on a particular source are stored as annotations, which can be used to analyze conflicting information or handle incomplete information.

In case of DaCura, the data-architect, domain expert, data harvester and consumer collaborate together to maintain a high-quality dataset. TripleCheckMate, allows multiple users to assess the same Linked Data resource and therefore allowing to calculate the inter-rater agreement to attain a final quality judgement.

Customizability. Customizability is the ability of a system to be configured according to the users' needs and preferences. In this case, we measure the customizability of a tool based on whether the tool can be used with any dataset that the user is interested in. Only LinkQA and LiQuate cannot be customized since the user cannot add any dataset of her choice. The other ten tools can be customized according to the use case. For example, in TrustBot, which is an IRC bot that makes trust recommendations to users (based on the trust network it builds), the users have the flexibility to submit their own URIs to the bot at any time while incorporating the data into a graph. Similarly, Trellis, tSPARQL, WIQA, ProLOD, Flemming's tool, Sieve, RDFUnit, DaCura and TripleCheckmate can be used with any dataset.

Scalability. Scalability is the ability of a system, network, or process to handle a growing amount of work or its ability to be enlarged to accommodate that growth. Out of the 12 tools only five, the tSPARQL, LinkQA, Sieve, RDFUnit and TripleCheckMate tools are scalable, that is, they can be used with large datasets. Flemming's tool and TrustBot are reportedly not scalable for large datasets [Flemming, 2011, Golbeck et al., 2003]. Flemming's tool, on the one hand, performs analysis based on a sample of three entities whereas TrustBot takes as input two email addresses to calculate the weighted average trust value. Trellis, WIQA, ProLOD, DaCura and LiQuate do not provide any information on the scalability.

Usability/Documentation. Usability is the ease of use and learnability of a human-made object, in this case the quality assessment tool. We assess the usability of the tools based on the ease of use as well as the complete and precise documentation available for each of them thus enabling users to find help easily. We score them based on a scale from 1 (low usability) to 5 (high usability). TripleCheckMate is the easiest tool with a user-friendly interface and a screencast explaining its usage. Thereafter, TrustBot, tSPARQL and Sieve score high in terms of usability and documentation followed by Flemming's tool and RDFUnit. Trellis, WIQA, ProLOD and LinkQA rank lower in terms of ease of use since they do not contain useful documentation of how to use the tool. DaCura and LiQuate do not provide any documentation except for a description in the paper.

Maintenance/Last updated. With regards to the current status of the tools, while TrustBot, Trellis and WIQA have not been updated since they were first introduced in 2003, 2005 and 2006 respectively, ProLOD and Flemming's tool have been updated in 2010. The recently updated tools are LinkQA (2011), tSRARQL and Sieve (2012), DaCura, TripleCheckMate, LiQuate (2013) and RDFUnit (2014) and are currently being maintained.

8.4. Calculation of Societal Progress Indicators

There are several organizations which focus on calculating specific societal progress indicators for specific countries or globally. For example, the World Bank provides reports on several different areas of research such as Health, Nutrition and Population, Science and Technology Development, Education etc.[6] for the 188 member countries. However, this data is not always up-to-date and needs to be manually analyzed thus being error-prone and time consuming.

Another organization is the Agency for Healthcare Research and Quality (AHRQ), which produces annual reports to measure trends in effectiveness of care, patient safety, timeliness and efficiency of care i.e. the latest available findings on quality of and access to health care. However, the organization only focuses on the US member states[7]. Additionally, a study devised 13 carefully calibrated performance indicators to compose the World University Rankings[8], thus only focusing towards one of the societal progress indicators. On the other hand, a KOF Index of Globalization[9] was proposed to measure the economic, social and political dimensions of globalization. However, the data is only available until 2011.

Particularly for calculating the research-disease disparity, among the many possibilities is the generation of cross-sectional studies comparing estimates of disease-specific research productivity with different indices measuring the burden of disease [Cary P. Gross and Powe, 1999, Gillum et al., 2011]. Other methods include the use of suitable statistical measures on samples of data to quantify the disparity [Bonito et al., 2005]. These methods to calculate the disparity are not only cumbersome and time consuming but also are limited in that they use a limited sample of the data for analysis as opposed to using entire datasets.

Although previous efforts have highlighted disparity issues between disease burden and research efforts in a given country that are beginning to be addressed, we see major pending problems that we intended to address through this project. First, since all information has to be manually collected by experts, current methods to generate reports that evaluate Research-Disease Disparity are burdensome and expensive. This problem is particularly pervasive in countries that need these evaluations the most, namely developing countries where the cost of such evaluations is prohibitive. As a direct consequence of this cost and expertise issue, current reports are not published as often, thus decreasing the ability of policy makers to obtain a current perspective on the magnitude of these problems. Additionally, since reports are scarce, comparison with other countries are not possible, thus making it difficult to search for successful policy models that could be used to level and decrease the disparity levels across nations.

[6] https://openknowledge.worldbank.org/
[7] http://www.ahrq.gov/research/findings/nhqrdr/index.html
[8] http://www.timeshighereducation.co.uk/world-university-rankings/2013-14/subject-ranking/subject/life-sciences/methodology
[9] http://globalization.kof.ethz.ch/

9. Conclusions and Future Work

This chapter provides an overview on the main contributions of this thesis along with the solutions to the research questions introduced in Chapter 1 (Section 1.4). It then discusses the future directions in which we intend to move further to extend and broaden the research conducted in the specific areas.

9.1. Summary of Contributions

In this section, we revisit each research question and provide a summary of the solution and the contributions provided by this thesis.

9.1.1. Descriptions of data quality dimensions and metrics

The research question we aimed to answer is:

- RQ1: What are the existing approaches to assess the quality of Linked Data employing a conceptual framework integrating prior approaches?

We further divided this RQ into the following:

- RQ1.1: What are the data quality problems that each approach assesses?
- RQ1.2: Which are the data quality dimensions and metrics supported by the proposed approaches?

To address this question, we conducted a systematic literature review and identified 30 different approaches that propose a data quality assessment methodology, specifically for LD. We first identified the problems that each of the 30 approaches addressed (RQ1.1) and then mapped these problems to a particular data quality dimension. We then unified the definitions that each approach provides and formalized them (RQ1.2) in Chapter 3 for 18 identified dimensions. We explained each dimension with the help of an example. Additionally, we provided a total of 69 metrics for these dimensions (RQ1.2). Furthermore, we classified each metric into being qualitatively or quantitatively assessed. These dimensions and metrics formed the core of this thesis as they are used in formulating the quality problem taxonomy (Chapter 4), which in turn is used to select the types of quality issues that are presented to the MTurk workers (Chapter 5). Also, specific metrics identified as a result of this survey are implemented as part of a tool and used to assess the quality of four datasets that are part of our use case (Chapter 7).

9.1.2. User-driven data quality assessment methodologies

The research question we aimed to answer is:

- RQ2: How can we address the quality of Linked Data using a user-driven methodology?

We further divided this RQ into the following:

- RQ2.1 How feasible is it to employ LD experts to assess the quality issues of LD?
- RQ2.2 How feasible is it to use a combination of user-driven and semi-automated methodology to assess the quality of LD?
- RQ2.3 Is it possible to detect quality issues in LD datasets via crowdsourcing mechanisms?
- RQ2.4 What type of crowd is most suitable for each type of quality issues?
- RQ2.5 Which types of assessment errors are made by lay users and experts?
- RQ2.6 How can we semi-automatically assess the quality of datasets and provide meaningful results to the user?

In order to address these research questions, we presented three different data quality assessment methodologies, which are user-driven and/or sensitive to a use case.

Firstly, we presented a user-driven methodology for assessing the quality of LD sources comprising of a manual and a semi-automatic process. In the *manual* process, the first phase includes the detection of common quality problems and their representation in a quality problem taxonomy. The second phase comprises of the evaluation of a large number of individual resources, according to the quality problem taxonomy, performed by users. This process is accompanied by a tool, namely *TripleCheckMate*, wherein a user assesses an individual resource and evaluates each fact for correctness. In this case, the user is a LD expert who is conversant with RDF. The second methodology is a *semi-automatic* process, in which the first phase involves the detection of common quality problems by the automatic creation of an extended schema for DBpedia. The second phase involves the generation and manual verification of schema axioms.

As a result of our study, more than 500 resources of the DBpedia dataset were analyzed. We found that a substantial number of problems exist and the overall quality, with a 11.93% error rate, is moderate. Moreover, the semi-automatic analysis revealed more than $200,000$ violations of property characteristics. In addition to the quality analysis of DBpedia, we devised a generic methodology for LD quality analysis, derived a comprehensive taxonomy of extraction quality problems and developed a tool, which can assist in the evaluation. All these contributions can be reused for analyzing any other extracted dataset (by domain experts). In sum, we showed that the employment of Linked Data experts to assess the quality issues of LD is feasible to a certain extent as it can be time-consuming and costly. We illustrated that a combination of user-driven and

semi-automated methodology to perform LD quality assessment is a feasible means of performing quality assessment.

Another means we employed for assessing the quality of linked data sets was via *crowdsourcing*. We utilized the wisdom of the crowd, i.e. workers from the online crowdsourcing platform MTurk, to assess the quality of DBpedia. In particular, we used the results from the previous user-driven assessment (performed by LD experts) and fed them to MTurk. We then compared the two methodologies in order to determine the cost and time feasibility of the approaches. We reported the results obtained by applying both these methodologies to DBpedia in Chapter 5. As a result, we presented a methodology that adjusts the crowdsourcing pattern Find-Fix-Verify to exploit the strengths of experts and microtask workers. The *Find* stage was implemented using a contest-based format to engage with a community of LD experts in discovering and classifying quality issues of DBpedia resources. We selected a subset of the contributions obtained through this contest (referring to flawed object values, incorrect data types and missing links) and asked the MTurk crowd to *Verify* them. The evaluation showed that both types of approaches are feasible but limited; in particular, the microtask experiments revealed that people with no expertise in Linked Data can be a useful resource to identify only very specific quality issues in an accurate and affordable manner, using the MTurk model. We consider our methodology to be applicable to RDF data sets, which are extracted from other sources and, hence, are likely to suffer from similar quality problems as DBpedia.

The third assessment methodology we proposed is that which implements the data quality metrics identified in our survey to provide a tool, namely *R2RLint*, to assess the quality of LD. This methodology is a combination of a semi-automated methodology and user-driven quality assessment. The user is not only provided the results of the assessment but also specific entities that cause the errors, which help users understand the quality issues and thus can fix them. We provided the specific dimensions along with detailed explanations of the implementation of the metrics in Chapter 6. The results of the quality assessment of the four datasets, namely LinkedCT, USPTO, Scimago and WorldBank, using the R2RLint tool are reported in Chapter 7.

9.1.3. Consumption of Linked Data leveraging on data quality

The research question we aimed to answer is:

- RQ3: How can we exploit Linked Data for a particular use case and ensure good data quality?

In response to this question, we designed a use case employing Linked Data to build the HER Observatory of societal progress indicators. We chose four linked datasets, namely LinkedCT, USPTO, Scimago and WorldBank and integrated them to determine the impact of research and technology on health and economic performance of countries per year. We performed this analysis using SEM and EFA methods, which produces reliable and interpretable results. In order to ensure *good* data quality of the datasets,

we performed semi-automated quality assessment on all the four datasets involved in the use case. We employed the semi-automated methodology, using the R2RLint tool to perform this assessment. Moreover, we showed that in comparison to the previous crowdsourcing quality assessment methodologies (Chapter 4 and Chapter 5), which are not only error-prone and time-consuming, this semi-automated methodology is feasible to perform quality assessment of datasets. Additionally, the user is involved in the assessment as she can choose which metrics apply to her use case and also is provided with interpretable results.

Thus, this RQ brought together both the challenges that this thesis addresses, that is, utilization of LD for a specific use case enhanced with the assessment of the quality of the datasets involved. We showed the importance of the role of data quality assessment and improvement in such a use case since two, out of the four, datasets were not utilized in calculating the results due to data quality problems. We provided details of the use case, results of the data quality assessment and results of the use case in Chapter 7. We also described the advantages of using a combination of semi-automated and user-driven quality assessment performed in this use case.

Revisiting our user scenario introduced in Section 1.2, Ms. Sharma can analyze and interpret the results from our use case and measure the healthcare performance of India. This allows her to determine whether research strategies developed by India are aligned with the respective healthcare needs for MDR-TB. This information is critical for guiding and shaping policies that facilitate the destruction of inequalities, planning health systems, improving healthcare delivery, promote and sustain population welfare, allocating budgets for R&D (set spending priorities), monitor progress and evaluate what works and what does not [Murray et al., 2004, Schlotthauer et al., 2008]. This analysis, in turn, helps improve the technology and methods for societal progress and motivates governments to collect and analyze useful data and compare assessments of inputs, service delivery and achievements for economic and health outcomes.

9.2. Limitations and Future Work

In this section, we describe the limitations and future work with regards to the main contributions of this thesis.

9.2.1. Quality Assessment Methodology for Linked Data

We introduced several methodologies that can be utilized to assess the quality of a LD source. However, we showed the quality assessment methodology applicable only to the use case, thus we aim to evaluate the methodology in its application to assess the quality of other datasets in different domains. We identified that one important component missing from these methodologies is the improvement of the datasets after quality assessment. In case of performing the quality assessment of DBpedia using LD experts, we aim to adopt an agile methodology to improve the quality in future versions by regularly providing feedback to the DBpedia maintainers to fix the problems

identified. Moreover, we aim to perform quality analysis in regular intervals in order to demonstrate possible improvements.

In case of using the crowd as a means to assess the quality, future work will first focus on conducting new experiments to test the value of the crowd for further different types of quality problems as well as for different LD sets from other knowledge domains. In the longer term, our work will also look into how to optimally integrate crowd contributions – by implementing the *Fix* stage – into curation processes and tools, in particular with respect to the trade-offs of costs and quality between manual and automatic approaches. Moreover, we plan to device a generic methodology for assessment of LD quality that not only takes the use case into account throughout the analysis but also includes the improvement phase.

9.2.2. Quality Assessment Tools for Linked Data

In order to assess the quality of LD sources, there should be user-friendly tools available so that even lay users are able to assess and interpret the results. Thus, we aim to improve the R2RLint tool to provide a user interface for assessment of quality and representation of results in an accessible way. We also aim to further implement more metrics that can be used to assess the quality of LD sources. Moreover, we plan to use the tool to assess the quality of other datasets so as to find ways to improve its usability and assess its feasibility in all cases.

9.2.3. Consumption of Linked Data leveraging on Data Quality

In our use case, we encountered major information loss due to data quality problems of the datasets involved. Thus, we aim to improve the quality of the included datasets so that they can be used in the next calculations. Also as future work, we plan to streamline the process of acquiring, converting, quality assessment, integrating and analyzing data relevant to the societal progress indicators. We intend to keep the HER Observatory up-to-date by integrating new data as and when they will be available. We also intend to add more relevant LD sources so as to answer even more relevant research questions. Our ultimate goal will be to offer a reliable system for policy makers to make informed decisions on these indicators by ensuring good quality, which would benefit society as a whole in the long run.

A. Curriculum Vitae

Amrapali Zaveri

Email: zaveri@informatik.uni-leipzig.de
Website: http://aksw.org/AmrapaliZaveri

Personal Data

Name: Amrapali Zaveri
Birth date: November 9th, 1984
Birth place: Mumbai, India
Nationality: Indian

Education

2010 – Present
University of Leipzig, Germany
Ph.D., Faculty of Mathematics and Computer Science, Department of Computer Science
Thesis title: **Linked Data Quality Assessment and its Application to Societal Progress Measurement**

2007 – 2009
Sikkim Manipal University of Health, Medical and Technological Sciences, India
M.Sc in Bioinformatics
Grade A, Score: 177/200, 82.5%

2005 – 2007
Fergusson College, University of Pune, India
B.Sc. in Zoology.

Grade: First class with distinction, 76.78%

2002 – 2005
S.M. Choksey High School and Junior College, Pune, India
Higher Second School Certificate, Science stream.
Grade: 1, 74.5%

2000
St.Mary's School, Pune, India
Secondary School Certiciate.
Grade: 1 distinction, 84.16%

Work experience

15 June, 2010 - 15 June 2011
AKSW Research group, Universität Leipzig, Germany
Position: Research Internship
Responsibility: Preliminary work on ReDD-Observatory project
http://redd.aksw.org

2 Feb, 2009 - 15 May 2010
National Neuroscience Institute, Singapore
Position: Senior Research Assistant
Responsibility: Working on development of ontologies to standardize research reporting and creating a center for excellence in research reporting in neurosurgery
http://www.nni.com.sg/

2 Feb, 2009 - 15 May 2010
Duke-NUS Graduate Medical School, Singapore
Position: Senior Research Assistant
Responsibility: Working with the "Research on Research" group, co-ordinating people and projects involving use of ontologies and semantic web technology, also performing meta-analysis
http://researchonresearch.duhs.duke.edu/site/

23 July, 2008 - 4 Aug 2008
Tathapi, India
Responsibility: Created an annotated bibliography on Human Resource issues of Health Workers in India
http://www.tathapi.org/

May, 2008 - July 2008
DNS E-business Consultancy
Responsibility: Created an ASP page to display records from a MS SQL back-end

database

15 Sept, 2006 - 30 March 2008
InSilico Consulting, Pune, India
Position: Technical Associate
Responsibility: Written technical documentation, help files, manuals and training material for bioinformatics/cheminformatics data mining software called InforSense KDE
www.inforsense.com

Feb - May 2007
M.Sc. dissertation
Project: PubChem Datamining using a Chemical Ontology

April - June 2006
M.Sc. first year project
Project: Secondary structure prediction from a given amino acid sequence

Research Interests

- Knowledge interlinking and fusion
- Data quality
- Biomedical data publishing
- Healthcare research
- Bioinformatics

Selected Publications

1. *Crowdsourcing Linked Data quality assessment*, In Proceedings of 12th International Semantic Web Conference, 21-25 October 2013, Sydney, Australia (ISWC 2013) [Acosta et al., 2013]
2. *Using Linked Data to evaluate the impact of Research and Development in Europe: a Structural Equation Model*, In Proceedings of 12th International Semantic Web Conference, 21-25 October 2013, Sydney, Australia (ISWC 2013) [Zaveri et al., 2013d]
3. *User-driven Quality Evaluation of DBpedia*, In Proceedings of 9th International Conference on Semantic Systems, I-SEMANTICS '13, Graz, Austria, September 4-6, (I-Semantics 2013) [Zaveri et al., 2013a]

4. *ReDD-Observatory: Using the Web of Data for Evaluating the Research-Disease Disparity*, In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI 2011) [Zaveri et al., 2011]

5. *Using Linked Data to build an Observatory of societal progress indicators*, Journal of Web Semantics (2014) [Zaveri et al., 2014b]

6. *Publishing and Interlinking the USPTO Patent Data*, Semantic Web Journal (2014) [Zaveri et al., 2014a]

7. *Quality assessment methodologies for Linked Data: A Survey*, Semantic Web Journal (2014) [Zaveri et al., 2015]

8. *Publishing and Interlinking the Global Health Observatory Dataset*, Semantic Web Journal (2013) [Zaveri et al., 2013b]

9. *Towards Biomedical Data Integration for Analyzing the Evolution of Cognition*, In Proceedings of Ontology and Data in Life Sciences Workshop (ODLS 2013) [Zaveri et al., 2013c]

10. *TripleCheckMate: A Tool for Crowdsourcing the Quality Assessment of Linked Data*, In Proceedings of the 4th conference on Knowledge Engineering and Semantic Web, October 7 - 9 (KESW 2013) [Kontokostas et al., 2013]

11. *Evaluating the disparity between active areas of biomedical research and the global burden of disease employing Linked Data and data-driven discovery*, In Proceedings of Ontologies in Biomedicine and Life Sciences Workshop (OBML 2010) [Zaveri et al., 2010b]

12. *Achieving high research reporting quality through the use of computational ontologies*, Neuroinformatics Journal (2010) [Zaveri et al., 2010a]

13. *Center for Excellence in Research Reporting in Neurosurgery (CERR-N) - Achieving high quality through the use of computational ontologies*, Journal of Neurotrauma (2009) [Zaveri et al., 2009]

Technical and Programming Skills

- Semantic Web: RDF, SPARQL, OntoWiki
- Databases: Oracle 9i, MySQL
- OS: Linux (Red Hat Certified)
- Misc: Technical writing, Scientific writing

Projects

- **ReDD-Observatory**:
  Project to evaluate the disparity between active areas of biomedical research and the global burden of disease using Linked Data and data-driven discovery.
  Available at http://aksw.org/Projects/ReDDObservatory.html.
- **User-driven quality evaluation of DBpedia**:
  Evaluating the quality of the resources present in DBpedia employing Linked Data experts.
  Available at http://aksw.org/Projects/DBpediaDQ.html.
- **Crowdsourcing DBpedia Quality Assessment**:
  Evaluating the quality of DBpedia using crowdsourcing platforms.
  Available at http://aksw.org/Projects/DBpediaDQCrowd.html.
- **Publishing and Interlinking the Global Health Observatory Dataset**:
  Converting, interlinking and publishing the Global Health Observatory data, provided by WHO, as Linked Data.
  Available at http://aksw.org/Projects/GHO.html.
- **Publishing and Interlinking the USPTO Patent Data**:
  Converting, interlinking and publishing the USPTO Patent data, as Linked Data.
  Available at http://aksw.org/Projects/USPatents.html.

Research Community Service

- **Guest co-editor** for International Journal on Semantic Web and Information Systems' (IJSWIS) Special Issue on Web Data Quality
  http://www.ijswis.org/?q=node/45
- **Program Committee/Reviewer** for:
  - *Conferences*: International Semantic Web Conference (ISWC), International Conference on Web Engineering (ICWE), Knowledge Engineering and Semantic Web (KESW)

- *Workshops*: Linked Data on the Web (LDOW), Ontologies in Biomedicine and Life Sciences (OBML), International Workshop on Linked Science (LISC), Crowdsourcing the Semantic Web (CrowdSem), Human-Semantic Web Interaction (HSWI), Linked Data Quality (LDQ)
- *Journals*: International Journal on Semantic Web and Information Systems (IJSWIS), Semantic Web Journal (SWJ), International Journal of Knowledge and Learning (IJKL)

Presentations

- 2 SEPT 2014, LDQ Presentation on "Methodology for Assessment of Linked Data Quality: A Framework" http://www.slideshare.net/amrapalijz/ldq2014-dq-methodology
- 24 OCT 2013, ISWC Presentation on "Using Linked Data to evaluate the impact of Research and Development in Europe: a Structural Equation Model" http://www.slideshare.net/amrapalijz/iswc2013-az
- 8 OCT 2013, KESW CONFERENCE Presentation on "TripleCheckMate: A tool for Crowdsourcing the Quality Assessment of Linked Data" http://www.slideshare.net/amrapalijz/triplecheckmate
- 5 OCT 2013, RUSSIAN SCHOOL OF OPEN DATA Presentation on Linked Data http://opendataschool.ru/2013/09/third-lesson/
- 2 OCT 2013, KESW SCHOOL Lecture on basic and advanced SPARQL http://slidewiki.org/deck/750_semantic-data-web-lecture-series
- 16 SEPT 2013, ODLS Presentation on "Towards Biomedical Data Integration for Analyzing the Evolution of Cognition" http://www.slideshare.net/amrapalijz/cogevo-odls-presentation
- 5 SEPT 2013, I-SEMANTICS Presentation on "User-driven Quality Evaluation of DBpedia" http://www.slideshare.net/amrapalijz/i-semantics-d-bpediadq
- 3 SEPT 2012, W3C HEALTH CARE LIFE SCIENCE GROUP Presentation on ReDD-Observatory http://goo.gl/prSs9
- 27 AUG 2012, W3C HEALTH CARE LIFE SCIENCE GROUP Presentation on Global Health Observatory http://goo.gl/UUlJc
- 18 - 22 OCT 2010, BRAZILIAN CONGRESS ON HEALTH INFORMATICS Tutorial on Computational Ontologies and Linked Data http://researchonresearch.duhs.duke.edu/site/?page_id=4722

- 9 - 10 Sept 2010, Workshop on Ontologies in Biomedicine and Life Sciences Evaluating the Disparity between Active Areas of Biomedical Research and the Global Burden of Disease Employing Linked Data and Data-driven Discovery https://wiki.imise.uni-leipzig.de/Gruppen/OBML/Workshops/2010en
- 1 Sept 2010, W3C's Linked Open Drug Data Group Conversion of GHO dataset to RDF using SCOVO vocabulary http://esw.w3.org/images/8/89/Amrapali_Zaveri_PPT.pdf
- Sept 2009, Joint Symposium of International and National Neurotrauma Societies Poster titled "Center for Excellence in Research Reporting in Neurosurgery (CERR-N) – achieving high quality through the use of computational ontologies" http://www.neurotrauma.org/2009/
- Dec 2006, International Conference on Bioinformatics (InCoB) Poster titled "TB Gateway – a comprehensive database on TB – achieving high quality through the use of computational ontologies" http://www.incob2006.in

Language Skills

- **Gujrathi**: Native
- **English**: Advanced
- **German**: Intermediate (Niveau stufe C1 Certified)
- Familiar with **Hindi** and **Marathi**.

List of Abbreviations

AFR Adolescent Fertility Rate, pp. 98, 100, 106, 108

MTurk Amazon Mechanical Turk, pp. IV, 7, 9, 21, 51, 66, 67, 69, 72, 112, 117, 119

BR Birth Rate, pp. 99, 106, 108

CFA Confirmatory Factor Analysis, pp. 95, 104–106, 132

CFI Comparative Fit Index, pp. 95, 106, 132

DALY Disability Adjusted Life Years, p. 92

DBMS Relational Database Management System, p. 20

DR Death Rate, pp. 99, 105, 108

EFA Exploratory Factor Analysis, pp. 92, 95, 104, 105, 119

ESDA Exploratory Spatial Data Analysis, p. 95

GDP Gross Domestic Product, pp. 94, 96, 99, 104, 106–108

GHO Global Health Observatory, pp. 90–92

HEP Health Expenditure Public, pp. 99, 108

HER Health Economic Research, pp. IV, 6, 8, 10, 89, 96, 109, 119

HET High-technology Export, pp. 99, 104–108

HIT Human Intelligent Tasks, pp. 51, 52, 67, 72, 74–77, 80

IDPT Immunization DPT, pp. 99, 108

IM Immunization Measles, pp. 99, 108

ITB Incidence of Tuberculosis, p. 99

LD Linked Data, pp. III, IV, VII–IX, 1–10, 21, 49, 51, 52, 54–56, 58–60, 62, 64–70, 72, 74–78, 80–82, 84, 86, 88–90, 96, 98, 109, 111–113, 117–121, 131, 133

LISA Local Indicators of Spatial Association, p. 96

LOD Linked Open Data, pp. 61, 75, 111

MDR-TB Multi-drug-resistant tuberculosis, pp. 3, 120

MR Mortality Rate, pp. 99, 108

OWL Web Ontology Language, pp. 12, 18, 19

PSE Public Spending on Education, p. 100

RDF Resource Description Framework, pp. VII, 7, 8, 11–16, 18–20, 52, 55, 60, 66, 68, 69, 72–74, 81, 92, 97, 98, 112, 119, 131, 133

RDFS Resource Description Framework Schema, pp. 18, 19

RGDP Research and Development Expenditure, p. 100

RMSEA Root Mean Square Error of Approximation, pp. 95, 106, 132

R&D Research and Development, pp. 89–91, 93, 94, 97, 105–108, 110, 134

RRP Researchers in R&D, p. 100

SEM Structural Equation Modeling, pp. 89–95, 100, 104, 110, 119

SIR SCImago Institutions Rankings, p. 98

SPARQL SPARQL Protocol and RDF Query Language, pp. 7, 8, 11, 19, 20, 53, 61, 96–98, 101

SRMR Standardized Root Mean Square Residual, pp. 95, 106, 132

TLI Tucker-Lewis Index, pp. 95, 106, 132

UMLS Unified Medical Language System, p. 91

URI Uniform Resource Identifier, pp. 11, 12, 14, 15, 20, 73, 101, 110

URL Uniform Resource Locator, pp. 11, 12

USPTO United States Patent and Trademark Office, pp. 92, 96, 97, 101–104, 110, 119, 132

W3C World Wide Web consortium, pp. 11, 18, 19

WHO World Health Organization, p. 3

WWW World Wide Web, p. 1

List of Tables

List of Figures

Bibliography

[Acosta et al., 2013] Acosta, M., Zaveri, A., Simperl, E., Kontokostas, D., Auer, S., and Lehmann, J. (2013). Crowdsourcing linked data quality assessment. In Alani, H., Kagal, L., Fokoue, A., Groth, P., Biemann, C., Parreira, J. X., Aroyo, L., Noy, N., Welty, C., and Janowicz, K., editors, *Proceedings of the 12th International Semantic Web Conference, (ISWC)*, volume 8219 of *Lecture Notes in Computer Science*, pages 260–276. Springer Berlin Heidelberg.

[Adams, 1990] Adams, J. D. (1990). Fundamental stocks of knowledge and productivity growth. *Journal of Political Economy*, 4:673–702.

[Albertoni and Perez, 2013] Albertoni, R. and Perez, A. G. (2013). Assessing linkset quality for complementing third-party datasets. In Guerrini, G., editor, *Proceedings of the joint International Conference on Extending Database Technology and International Conference on Database Theory (EDBT/ICDT)*, pages 52–59. ACM.

[Anselin, 2010] Anselin, L. (2010). *Geographical information systems: principles, techniques, management and applications - Chapter 17. Interactive techniques and exploratory spatial data analysis*. Wiley: New York.

[Anselin et al., 2010] Anselin, L., Syabri, I., and Kho, Y. (2010). Geoda: An introduction to spatial data analysis. In Fischer, M. M. and Getis, A., editors, *Handbook of Applied Spatial Analysis*, pages 73–89. Springer Berlin Heidelberg.

[Auer and Lehmann, 2007] Auer, S. and Lehmann, J. (2007). What Have Innsbruck and Leipzig in Common? Extracting Semantics from Wiki Content. In Franconi, E., Kifer, M., and May, W., editors, *Proceedings of the 4th European conference on The Semantic Web: Research and Applications (ESWC)*, volume 4519 of *Lecture Notes in Computer Science*, pages 503–517. Springer Berlin Heidelberg.

[Auer et al., 2010] Auer, S., Weidl, M., Lehmann, J., Zaveri, A., and Choi, K.-S. (2010). I18n of Semantic Web Applications. In *The Semantic Web – ISWC 2010 (ISWC)*, volume 6497 of *Lecture Notes in Computer Science*, pages 1–16. Springer Berlin Heidelberg.

[Baader et al., 2003] Baader, F., Diageo, C., McGuinness, D., Nardi, D., and Patel-Schneider, P., editors (2003). *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press New York, NY, USA.

[Batini and Scannapieco, 2006] Batini, C. and Scannapieco, M. (2006). *Data Quality: Concepts, Methodologies and Techniques*. Springer-Verlag New York, Inc.

[Bechhofer et al., 2004] Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F., and Stein, L. A. (2004). OWL Web Ontology Language Reference. Technical report, World Wide Web Consortium. http://www.w3.org/TR/owl-ref/.

[Beckett, 2004] Beckett, D. (2004). RDF/XML syntax specification (revised). W3C recommendation, World Wide Web Consortium. http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/.

[Berners-Lee, 2006] Berners-Lee, T. (2006). Linked data - design issues. http://www.w3.org/DesignIssues/LinkedData.html.

[Berners-Lee and Connolly, 2011] Berners-Lee, T. and Connolly, D. (2011). Notation3 (N3): A readable RDF syntax. Technical report, World Wide Web Consortium. http://www.w3.org/TeamSubmission/n3/.

[Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5):34–43.

[Bernstein et al., 2010] Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., and Panovich, K. (2010). Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology (UIST)*, pages 313–322, New York, NY, USA. ACM.

[Bessette, 2003] Bessette, R. W. (2003). Measuring the economic impact of university-based research. *Journal of Technology Transfer*, 28(3-4):355–361.

[Bishop et al., 2011] Bishop, B., Kiryakov, A., Ognyanoff, D., Peikov, I., Tashev, Z., and Velkov, R. (2011). OWLIM: A family of scalable semantic repositories. *Semantic Web*, 2(1):1–10.

[Bizer, 2007] Bizer, C. (2007). *Quality-Driven Information Filtering in the Context of Web-Based Information Systems*. PhD thesis, Freie Universität Berlin.

[Bizer and Cyganiak, 2009] Bizer, C. and Cyganiak, R. (2009). Quality-driven information filtering using the WIQA policy framework. *Journal of Web Semantics*, 7(1):1–10.

[Bleiholder and Naumann, 2008] Bleiholder, J. and Naumann, F. (2008). Data fusion. *ACM Computing Surveys (CSUR)*, 41(1):1.

[Böhm et al., 2010] Böhm, C., Naumann, F., Abedjan, Z., Fenz, D., Grütze, T., Hefenbrock, D., Pohl, M., and Sonnabend, D. (2010). Profiling linked open data with ProLOD. In *IEEE International Conference on Data Engineering (ICDE)*, pages 175–178. IEEE.

[Bollen, 2000] Bollen, K. A. (2000). *Structural Equations with Latent Variables*. Wiley Series in Probability and Mathematical Statistics.

[Bonatti et al., 2011] Bonatti, P. A., Hogan, A., Polleres, A., and Sauro, L. (2011). Robust and scalable linked data reasoning incorporating provenance and trust annotations. *Journal of Web Semantics*, 9(2):165–201.

[Bonito et al., 2005] Bonito, A. J., Eicheldinger, C. R., and Lenfestey, N. F. (2005). Health disparities: Measuring health care use and access for racial/ethnic populations. Technical report, RTI International.

[Boyce et al., 2014] Boyce, R. D., Ryan, P. B., Norén, G. N., Schuemie, M. J., Reich, C., Duke, J., Nicholas, and Tatonetti, P. (2014). Bridging islands of information to establish an integrated knowledge base of drugs and health outcomes of interest. *Drug Safety*, 37(7).

[Brickley and Guha, 2004] Brickley, D. and Guha, R. V. (2004). RDF vocabulary description language 1.0: RDF schema. Technical report, World Wide Web Consortium. http://www.w3.org/TR/2004/REC-rdf-schema-20040210/.

[Broekstra et al., 2002] Broekstra, J., Kampman, A., and van Harmelen, F. (2002). Sesame: A generic architecture for storing and querying RDF and RDF schema. In Horrocks, I. and Hendler, J., editors, *The Semantic Web — ISWC 2002 (ISWC)*, volume 2342 of *Lecture Notes in Computer Science*, pages 54–68. Springer Berlin Heidelberg.

[Bühmann and Lehmann, 2012] Bühmann, L. and Lehmann, J. (2012). Universal OWL axiom enrichment for large knowledge bases. In ten Teije, A., Völker, J., Handschuh, S., Stuckenschmidt, H., d'Acquin, M., Nikolov, A., Aussenac-Gilles, N., and Hernandez, N., editors, *Knowledge Engineering and Knowledge Management (EKAW)*, volume 7603 of *Lecture Notes in Computer Science*, pages 57–71. Springer Berlin Heidelberg.

[Cafarella et al., 2008] Cafarella, M. J., Halevy, A. Y., Wang, D. Z., Wu, E., and Zhang, Y. (2008). WebTables: exploring the power of tables on the web. *Proceedings of the VLDB Endowment*, 1(1):538–549.

[Carroll, 2003] Carroll, J. J. (2003). Signing RDF graphs. In Fensel, D., Sycara, K., and Mylopoulos, J., editors, *The Semantic Web - ISWC 2003 (ISWC)*, volume 2870 of *Lecture Notes in Computer Science*, pages 369–384. Springer Berlin Heidelberg.

[Cary P. Gross and Powe, 1999] Cary P. Gross, G. F. A. and Powe, N. R. (1999). The Relation between Funding by the Ntaional Institute of Health and the Burden of Diseases. *The New England Journal of Medicine*, 340(1881-1887).

[Chen and Garcia, 2010] Chen, P. and Garcia, W. (2010). Hypothesis generation and data quality assessment through association mining. In *9th IEEE International Conference on Cognitive Informatics (ICCI)*, pages 659–666. IEEE.

[Clark et al., 2008] Clark, K. G., Feigenbaum, L., and Torres, E. (2008). SPARQL Protocol for RDF. World Wide Web Consortium, Recommendation REC-rdf-sparql-protocol-20080115. http://www.w3.org/TR/rdf-sparql-protocol/.

[Costello and Osborne, 2005] Costello, A. B. and Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment Research and Evaluation*, 10(7).

[Daraio et al., 2011] Daraio, C., Bonaccorsi, A., Geuna, A., Lepori, B., Bach, L., Bogetoft, P., Cardoso, M. F., Castro-Martinez, E., Crespi, G., de Lucio, I. F., Fried, H., Garcia-Aracil, A., Inzelt, A., Jongbloed, B., Kempkes, G., Llerena, P., Matt, M., Olivares, M., Pohl, C., Raty, T., Rosa, M. J., Sarrico, C. S., Simar, L., Slipersaeter, S., Teixeira, P. N., and Eeckaut, P. V. (2011). The european university landscape: A micro characterization based on evidence from the Aquameth project. *Research Policy*, 40(1):148 – 164.

[Dave and Berners-Lee, 2011] Dave, D. and Berners-Lee, T. (2011). Turtle - Terse RDF Triple Language. Technical report, World Wide Web Consortium. http://www.w3.org/TeamSubmission/turtle/.

[Demartini et al., 2012] Demartini, G., Difallah, D., and Cudré-Mauroux, P. (2012). Zencrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *21st International Conference on World Wide Web (WWW)*, pages 469–478. ACM.

[Dezani-Ciancaglini et al., 2012] Dezani-Ciancaglini, M., Horne, R., and Sassone, V. (2012). Tracing where and who provenance in linked data: A calculus. *Theoretical Computer Science*, 464:113–129.

[Ding and Finin, 2006] Ding, L. and Finin, T. (2006). Characterizing the semantic web on the web. In Cruz, I., Decker, S., Allemang, D., Priest, C., Schwabe, D., Mika, P., Uschold, M., and Aroyo, L., editors, *The Semantic Web - ISWC 2006 (ISWC)*, volume 4273 of *Lecture Notes in Computer Science*, pages 242–257. Springer Berlin Heidelberg.

[Druck et al., 2004] Druck, S., Carvalho, M. S., Câmara, G., and Monteiro, A. M. V. (2004). *Spatial analysis of geographic data*. EMBRAPA.

[Erling and Mikhailov, 2009] Erling, O. and Mikhailov, I. (2009). RDF support in the virtuoso DBMS. In Pellegrini, T., Auer, S., Tochtermann, K., and Schaffert, S., editors, *Networked Knowledge - Networked Media*, volume 221 of *Studies in Computational Intelligence*, pages 7–24. Springer Berlin Heidelberg.

[Feeney et al., 2014] Feeney, K. C., O'Sullivan, D., Tai, W., and Brennan, R. (2014). Improving curated web-data quality with structured harvesting and assessment. *International Journal on Semantic Web and Information Systems*, 10(2):35–62.

[Flemming, 2011] Flemming, A. (2011). Qualitätsmerkmale von linked data-veröffentlichenden datenquellen. Master's thesis, Humboldt-Universität zu Berlin, Institut für Informatik.

[Fox, 2006] Fox, J. (2006). Structural Equation Models. Technical report, R CRAN Packages.

[Fürber and Hepp, 2011] Fürber, C. and Hepp, M. (2011). SWIQA - a semantic web information quality assessment framework. In Tuunainen, V. K., Rossi, M., and Nandhakumar, J., editors, *Proceedings of the 19th European Conference on Information Systems (ECIS)*, volume 15, pages 19–30. IEEE Computer Society.

[Gamble and Goble, 2011] Gamble, M. and Goble, C. (2011). Quality, trust, and utility of scientific data on the web: Towards a joint model. In *Proceedings of the 3rd International Web Science Conference (WebSci)*, pages 1–8. ACM.

[Gil and Artz, 2007] Gil, Y. and Artz, D. (2007). Towards content trust of web resources. *Web Semantics*, 5(4):227 – 239.

[Gil and Ratnakar, 2002] Gil, Y. and Ratnakar, V. (2002). Trusting information sources one citizen at a time. In *Proceedings of the First International Semantic Web Conference on The Semantic Web (ISWC)*, Lecture Notes in Computer Science, pages 162–176. Springer Berlin Heidelberg.

[Gillum et al., 2011] Gillum, L. A., Gouveia, C., Dorsey, E. R., Pletcher, M., Mathers, C. D., McCulloch, C. E., and Johnston, S. C. (2011). NIH Disease Funding Levels and Burden of Disease. *PLoS One*.

[Golbeck, 2006] Golbeck, J. (2006). Using trust and provenance for content filtering on the semantic web. In Finin, T., Kagal, L., and Olmedilla, D., editors, *Proceedings of the Workshop on Models of Trust on the Web, at the 15th World Wide Web conference (WWW)*.

[Golbeck et al., 2003] Golbeck, J., Parsia, B., and Hendler, J. (2003). Trust networks on the semantic web (cia). In Klusch, M., Omicini, A., Ossowski, S., and Laamanen, H., editors, *Cooperative Information Agents VII*, volume 2782 of *Lecture Notes in Computer Science*, pages 238–249. Springer Berlin Heidelberg.

[Grant and Beckett, 2004] Grant, J. and Beckett, D. (2004). RDF test cases. W3C recommendation, World Wide Web Consortium. http://www.w3.org/TR/rdf-testcases/.

[Gross, 2012] Gross, J. (2012). nortest: Tests for normality. Technical report, R CRAN Packages.

[Guéret et al., 2012a] Guéret, C., Groth, P., Stadler, C., and Lehmann, J. (2012a). Assessing linked data mappings using network measures. In Simperl, E., Cimiano, P.,

Polleres, A., Corcho, O., and Presutti, V., editors, *The Semantic Web: Research and Applications (ESWC)*, volume 7295 of *Lecture Notes in Computer Science*, pages 87–102. Springer Berlin Heidelberg.

[Guéret et al., 2012b] Guéret, C., Groth, P. T., Stadler, C., and Lehmann, J. (2012b). Assessing linked data mappings using network measures. In Simperl, E., Cimiano, P., Polleres, A., Corcho, O., and Presutti, V., editors, *The Semantic Web: Research and Applications (ESWC)*, volume 7295 of *Lecture Notes in Computer Science*, pages 87–102. Springer Berlin Heidelberg.

[Hair et al., 2012] Hair, J. F., Sarstedt, M., Ringle, C. M., and Mena, J. A. (2012). An assessment of the use of partial least squares structural equation modeling in marketing research. *Journal of the Academy of Marketing Science*, 40(3):414–433.

[Halpin et al., 2010] Halpin, H., Hayes, P., McCusker, J. P., McGuinness, D., and Thompson, H. S. (2010). When owl:sameAs isn't the Same: An Analysis of Identity in Linked Data. In Patel-Schneider, P. F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J. Z., Horrocks, I., and Glimm, B., editors, *The Semantic Web – ISWC 2010 (ISWC)*, volume 6496 of *Lecture Notes in Computer Science*, pages 305–320. Springer Berlin Heidelberg.

[Hanushek and Woessmann, 2010] Hanushek, E. and Woessmann, L. (2010). *Education and economic growth*. Elsevier.

[Harris and Seaborne, 2013] Harris, S. and Seaborne, A., editors (2013). *SPARQL 1.1 Query Language*. World Wide Web Consortium. Available at http://www.w3.org/TR/2013/REC-sparql11-query-20130321/.

[Hartig, 2008] Hartig, O. (2008). Trustworthiness of data on the web. In *STI Berlin and CSW PhD Workshop, Berlin, Germany*.

[Hartig and Zhao, 2009] Hartig, O. and Zhao, J. (2009). Using web data provenance for quality assessment. In Freire, J., Missier, P., and Sahoo, S. S., editors, *Proceedings of the 1st International Workshop on the Role of Semantic Web in Provenance Management (SWPM) at ISWC*, volume 526 of *CEUR Workshop Proceedings*.

[Hasan and Tucci, 2010] Hasan, I. and Tucci, C. (2010). The innovation-economic growth nexus: Global evidence. *Research Policy*.

[Hays et al., 2005] Hays, R. D., Revicki, D., and Coyne, K. S. (2005). Application of Structural Equation Modeling to Health Outcomes Research. *Eval Health Prof*, 28(3):295–309.

[Heath and Bizer, 2011] Heath, T. and Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space (1st edition)*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool.

[Heflin, 2004] Heflin, J. (2004). OWL Web Ontology Language Use Cases and Requirements. Technical report, World Wide Web Consortium. http://www.w3.org/TR/webont-req/.

[Henderson et al., 1998] Henderson, R., Jaffe, A., and Trajtenberg, M. (1998). Universities as a source of commercial technology: a detailed analysis of university patenting, 1965-1988. *Review of Economics and Statistics*, pages 119–127.

[Hogan et al., 2010] Hogan, A., Harth, A., Passant, A., Decker, S., and Polleres, A. (2010). Weaving the Pedantic Web. In Bizer, C., Heath, T., Berners-Lee, T., and Hausenblas, M., editors, *3rd Linked Data on the Web (LDOW) Workshop at WWW*, volume 628, Raleigh, North Carolina, USA. CEUR Workshop Proceedings.

[Hogan et al., 2012] Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., and Decker, S. (2012). An empirical survey of linked data conformance. *Journal of Web Semantics*, 14:14–44.

[Hornik, 2008] Hornik, K. (2008). The R project: A language and environment for statistical computing. http://www.r-project.org/.

[Howe, 2006] Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine*, 14(6).

[Hox, 1998] Hox, J. (1998). An intorduction to Structureal Equation Modeling. *Family Science Review*, 11:354–373.

[Jacobi et al., 2011] Jacobi, I., Kagal, L., and Khandelwal, A. (2011). Rule-based trust assessment on the semantic web. In *Proceedings of the 5th international conference on Rule-based reasoning, programming, and applications (RuleML)*, pages 227–241. Springer Berlin Heidelberg.

[Jarke et al., 2010] Jarke, M., Lenzerini, M., Vassiliou, Y., and Vassiliadis, P. (2010). *Fundamentals of Data Warehouses*. Springer Publishing Company, 2nd edition.

[Jentzsch et al., 2009] Jentzsch, A., Hassanzadeh, O., Bizer, C., Andersson, B., and Stephens, S. (2009). Enabling tailored therapeutics with linked data. In *Proceedings of the WWW workshop on Linked Data on the Web (LDOW)*.

[Juran, 1974] Juran, J. (1974). *The Quality Control Handbook*. McGraw-Hill, New York.

[Kilpeläinen et al., 2012] Kilpeläinen, K., Tuomi-Nikula, A., Thelen, J., Gissler, M., Sihvonen, A.-P., kramers, P., and Aromaa, A. (2012). Health indicators in europe: availability and data needs. *The European Journal of Public Health*.

[Kitchenham, 2004] Kitchenham, B. (2004). Procedures for performing systematic reviews. Technical report, Joint Technical Report Keele University Technical Report TR/SE-0401 and NICTA Technical Report 0400011T.1.

[Kline, 1994] Kline, P. (1994). *An Easy Guide to Factor Analysis*. Routledge, London.

[Kline, 2011] Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling*. The Guilford Press, New York.

[Knuth et al., 2012] Knuth, M., Hercher, J., and Sack, H. (2012). Collaboratively patching linked data. In *Proceedings of the 2nd International Workshop on Usage Analysis and the Web of Data (USEWOD)*.

[Kontokostas et al., 2014] Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., and Zaveri, A. (2014). Test-driven Evaluation of Linked Data Quality. In *Proceedings of the 23rd International Conference on World Wide Web (WWW)*, pages 747–758. ACM.

[Kontokostas et al., 2013] Kontokostas, D., Zaveri, A., Auer, S., and Lehmann, J. (2013). TripleCheckMate: A Tool for Crowdsourcing the Quality Assessment of Linked Data. In Klinov, P. and Mouromtsev, D., editors, *Knowledge Engineering and the Semantic Web (KESW)*, volume 394 of *Communications in Computer and Information Science*, pages 265–272. Springer Berlin Heidelberg.

[Korkmaz, 2013] Korkmaz, S. (2013). Multivariate normality tests. Technical report, R CRAN Packages.

[Kreis, 2011] Kreis, P. (2011). Design of a Quality Assessment Framework for the DBpedia Knowledge Base. Master's thesis, Freie Universität Berlin.

[Krejcie and Morgan, 1970] Krejcie and Morgan (1970). Determining sample size for research activities. *Educational and Psycholoigcal Measurement*, 30:607–610.

[Lab, 2012] Lab, S. (2012). SCImago institutions rankings. Technical report, SCImago Research Group.

[Labra Gayo et al., 2012] Labra Gayo, J. E., Kontokostas, D., and Auer, S. (2012). Multilingual linked open data patterns. *Semantic Web Journal*.

[Larsen and von Ins, 2010] Larsen, P. O. and von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, 84(3):575–603.

[Lehmann, 2009] Lehmann, J. (2009). DL-Learner: learning concepts in description logics. *Journal of Machine Learning Research*, 10:2639–2642.

[Lehmann et al., 2009] Lehmann, J., Bizer, C., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). DBpedia - a crystallization point for the web of data. *Journal of Web Semantics*, 7(3):154–165.

[Lehmann and Bühmann, 2010] Lehmann, J. and Bühmann, L. (2010). ORE - A Tool for Repairing and Enriching Knowledge Bases. In Patel-Schneider, P. F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J. Z., Horrocks, I., and Glimm, B., editors, *The Semantic Web – ISWC 2010 (ISWC)*, volume 6497 of *Lecture Notes in Computer Science*, pages 177–193. Springer Berlin Heidelberg.

[Lehmann and Hitzler, 2010] Lehmann, J. and Hitzler, P. (2010). Concept learning in description logics using refinement operators. *Machine Learning journal*, 78(1-2):203–250.

[Lehmann et al., 2014] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C. (2014). DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*.

[Lei et al., 2007a] Lei, Y., Nikolov, A., Uren, V., and Motta, E. (2007a). Detecting Quality Problems in Semantic Metadata without the Presence of a Gold Standard. In *Workshop on "Evaluation of Ontologies for the Web" (EON) at the WWW*, volume 329 of *CEUR Workshop Proceedings*, pages 51–60.

[Lei et al., 2007b] Lei, Y., Uren, V., and Motta, E. (2007b). A framework for evaluating semantic metadata. In *4th International Conference on Knowledge Capture (KCAP)*, number 8 in K-CAP '07, pages 135 – 142. ACM.

[Leo Pipino and Rybold, 2005] Leo Pipino, Ricahrd Wang, D. K. and Rybold, W. (2005). *Developing Measurement Scales for Data-Quality Dimensions*, volume 1. M.E. Sharpe, New York.

[Mansfield, 1995] Mansfield, E. (1995). Academic research underlying industrial innovations: sources, characteristics, and financing. *Review of Economics and Statistics*.

[Markotschi and Völker, 2010] Markotschi, T. and Völker, J. (2010). GuessWhat?! - Human Intelligence for Mining Linked Data. In *Proceedings of the Workshop on Knowledge Injection into and Extraction from Linked Data (KIELD) at EKAW*.

[Maynard et al., 2006] Maynard, D., Peters, W., and Li, Y. (2006). Metrics for evaluation of ontology-based information extraction. In *Workshop on "Evaluation of Ontologies for the Web" (EON) at WWW*.

[Mendes et al., 2012a] Mendes, P., Bizer, C., Miklos, Z., Calbimonte, J.-P., Moraru, A., and Flouris, G. (2012a). D2.1: Conceptual model and best practices for high-quality metadata publishing. Technical report, PlanetData Deliverable.

[Mendes et al., 2012b] Mendes, P., Mühleisen, H., and Bizer, C. (2012b). Sieve: Linked data quality assessment and fusion. In *Proceedings of the joint International Conference on Extending Database Technology and International Conference on Database Theory (EDBT/ICDT) workshops*, pages 116–123. ACM.

[Moher et al., 2009] Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS medicine*, 6(7).

[Morsey et al., 2012] Morsey, M., Lehmann, J., Auer, S., Stadler, C., and Hellmann, S. (2012). DBpedia and the Live Extraction of Structured Data from Wikipedia. *Program: electronic library and information systems*, 46:27.

[Mostafavi et al., 2004] Mostafavi, M., G., E., and Jeansoulin, R. (2004). Ontology-based method for quality assessment of spatial data bases. In *International Symposium on Spatial Data Quality (ISSDQ)*, volume 4, pages 49–66.

[Murray et al., 2004] Murray, C. J. L., Lopez, A. D., and Wibulpolprasert, S. (2004). Monitoring global health: time for new solutions. *BMJ*.

[Naumann, 2002] Naumann, F. (2002). *Quality-Driven Query Answering for Integrated Information Systems*, volume 2261 of *Lecture Notes in Computer Science*. Springer-Verlag.

[Orr, 1998] Orr, K. (1998). Data quality and systems theory. *Communications of the ACM*, 41(2):66–71.

[Paulheim and Bizer, 2014] Paulheim, H. and Bizer, C. (2014). Improving the quality of linked data using statistical distributions. *International Journal on Semantic Web and Information Systems*.

[Perobelli and Haddad, 2006] Perobelli, F. S. and Haddad, E. A. (2006). Inter-state commerce patterns in brazil, 1985 to 1997. *Rev Econ Contempo*.

[Peterson et al., 2012] Peterson, D., Gao, S. S., Malhotra, A., Sperberg-McQueen, C. M., and Thompson, H. S., editors (2012). *W3C XML Schema Definition Language (XSD) 1.1 Part 2: Datatypes*. W3C Recommendation. World Wide Web Consortium. http://www.w3.org/TR/xmlschema11-2/.

[Phillips and Davis, 2009] Phillips, A. and Davis, M., editors (2009). *Tags for Identifying Languages*. Number 5646 in Request for Comments. Internet Engineering Task Force. Available at http://tools.ietf.org/rfc/bcp/bcp47.txt.

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

[Prud'hommeaux and Seaborne, 2008] Prud'hommeaux, E. and Seaborne, A. (2008). SPARQL query language for RDF. W3C recommendation, World Wide Web Consortium. http://www.w3.org/TR/rdf-sparql-query/.

[R Core Team, 2014] R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Redman, 1997] Redman, T. C. (1997). *Data Quality for the Information Age*. Artech House, 1st edition.

[Ruckhaus et al., 2014] Ruckhaus, E., Baldizán, O., and Vidal, M.-E. (2014). Analyzing Linked Data Quality with LiQuate. In *On the Move to Meaningful Internet Systems: (OTM) 2013 Workshops*, volume 8186, pages 629–638. Lecture Notes in Computer Science.

[Rula et al., 2012] Rula, A., Palmonari, M., and Maurino, A. (2012). Capturing the Age of Linked Open Data: Towards a Dataset-independent Framework. In *IEEE Sixth International Conference on Semantic Computing (ICSC)*, pages 218–225. IEEE Computer Society.

[Sarasua et al., 2012] Sarasua, C., Simperl, E., and Noy, N. (2012). CrowdMap: Crowdsourcing Ontology Alignment with Microtasks. In Cudré-Mauroux, P., Heflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Parreira, J. X., Hendler, J., Schreiber, G., Bernsteina, A., and Blomqvist, E., editors, *The Semantic Web – ISWC 2012 (ISWC)*, volume 7649 of *Lecture Notes in Computer Science*, pages 525–541. Springer Berlin Heidelberg.

[Schafer, 2008] Schafer, J. L. (2008). Norm package for R, version 3. Technical report, The Methodology Center, The Pennsylvania State University.

[Schermelleh-Engel et al., 2003] Schermelleh-Engel, K., Moosbrugger, H., and Muller, H. (2003). Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8(2):23–74.

[Schlotthauer et al., 2008] Schlotthauer, A. E., Badler, A., Cook, S. C., Pérez, D. J., and Chin, M. H. (2008). Evaluating interventions to reduce health care disparities: an RWJF program. *Health Affairs*, 27:568–73.

[Shekarpour and Katebi, 2010] Shekarpour, S. and Katebi, S. (2010). Modeling and evaluation of trust with an extension in semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(1):26 – 36.

[Stats Package, 2013] Stats Package, T. R. (2013). Mahalanobis distance. Technical report, R Documentation.

[Thaler et al., 2011] Thaler, S., Siorpaes, K., and Simperl, E. (2011). SpotTheLink: A Game for Ontology Alignment. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, pages 1711–1712. ACM.

[Tramp et al., 2010] Tramp, S., Heino, N., Auer, S., and Frischmuth, P. (2010). RDFauthor: Employing RDFa for Collaborative Knowledge Engineering. In Cimiano, P. and Pinto, H., editors, *Knowledge Engineering and Management by the Masses (EKAW)*, volume 6317 of *Lecture Notes in Computer Science*, pages 90–104. Springer Berlin Heidelberg.

[van Hage et al., 2014] van Hage, W. R., Kauppinen, T., Graeler, B., Davis, C., Hoeksema, J., Ruttenberg, A., and Bahls, D. (2014). *SPARQL Package, v1.6*. R Foundation for Statistical Computing.

[Wand and Wang, 1996] Wand, Y. and Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11):86–95.

[Wang et al., 2012] Wang, J., Kraska, T., Franklin, M. J., and Feng, J. (2012). CrowdER: crowdsourcing entity resolution. *Proceedings of the VLDB Endowment*, 5(11):1483–1494.

[Wang and Strong, 1996] Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–33.

[Westphal et al., 2014] Westphal, P., Stadler, C., and Lehmann, J. (2014). Quality assurance of RDB2RDF mappings. Technical report, Department of Computer Science, University of Leipzig. Available at http://svn.aksw.org/papers/2014/report_QA_RDB2RDF/public.pdf.

[Wienand and Paulheim, 2014] Wienand, D. and Paulheim, H. (2014). Detecting Incorrect Numerical Data in DBpedia. In Presutti, V., d'Amato, C., Gandon, F., d'Acquin, M., Staab, S., and Tordai, A., editors, *The Semantic Web: Trends and Challenges (ISWC)*, volume 8465 of *Lecture Notes in Computer Science*, pages 504–518. Springer Berlin Heidelberg.

[Wikipedia, 2013] Wikipedia (2013). SPARQL — Wikipedia, The Free Encyclopedia. [Online; accessed 31-March-2013].

[Williams et al., 2012] Williams, A. J., Harland, L., Groth, P., Pettifer, S., Chichester, C., Willighagen, E. L., Evelo, C. T., Blomberg, N., Ecker, G., CaroleGoble, and Mons, B. (2012). Open PHACTS: semantic interoperability for drug discovery. *Drug Discovery Today*, 17(21-22):1188–1198.

[World Wide Web Consortium, 2004] World Wide Web Consortium (2004). Resource Description Framework (RDF). http://www.w3.org/RDF/.

[World Wide Web Consortium, 2009] World Wide Web Consortium (2009). W3C Semantic Web Activity. http://www.w3.org/2001/sw/.

[Yu, 2007] Yu, L. (2007). *Introduction to Semantic Web and Semantic Web services*. Chapman & Hall/CRC, Boca Raton, FL.

[Zaveri et al., 2010a] Zaveri, A., Cofiel, L., Shah, J., Pradhan, S., Chan, E., Dameron, O., Pietrobon, R., and Ang, B. T. (2010a). Achieving high research reporting quality through the use of computational ontologies. *Neuroinformatics*, 8(4):261–271.

[Zaveri et al., 2014a] Zaveri, A., Hassan, M. M., Yousef, T., Auer, S., and Lehmann, J. (2014a). Publishing and Interlinking the USPTO Patent Data (Under review). Semantic Web Journal, http://www.semantic-web-journal.net/content/publishing-and-interlinking-uspto-patent-data.

[Zaveri et al., 2013a] Zaveri, A., Kontokostas, D., Sherif, M. A., Bühmann, L., Morsey, M., Auer, S., and Lehmann, J. (2013a). User-driven quality evaluation of DBpedia. In Sabou, M., Blomqvist, E., Noia, T. D., Sack, H., and Pellegrini, T., editors, *Proceedings of the 9th International Conference on Semantic Systems (ICSS)*, pages 97–104. ACM.

[Zaveri et al., 2013b] Zaveri, A., Lehmann, J., Auer, S., Hassan, M. M., Sherif, M. A., and Martin, M. (2013b). Publishing and Interlinking the Global Health Observatory Dataset. *Semantic Web Journal*, Special Call for Linked Dataset descriptions(3):315–322.

[Zaveri et al., 2013c] Zaveri, A., Nowick, K., and Lehmann, J. (2013c). Towards Biomedical Data Integration for Analyzing the Evolution of Cognition. In *GI-Jahrestagung'13*, pages 1900–1907.

[Zaveri et al., 2011] Zaveri, A., Pietrobon, R., Auer, S., Lehmann, J., Martin, M., and Ermilov, T. (2011). ReDD-Observatory: Using the Web of Data for Evaluating the Research-Disease Disparity. In Boissier, O., Benatallah, B., Papazoglou, M. P., Ras, Z. W., and Hacid, M., editors, *Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence (ICWI)*, pages 178–185. IEEE Computer Society.

[Zaveri et al., 2010b] Zaveri, A., Pietrobon, R., Ermilov, T., Martin, M., Heino, N., and Auer, S. (2010b). Evaluating the disparity between active areas of biomedical research and the global burden of disease employing linked data and data-driven discovery. In H.Herre, R.Hoehndorf, J.Kelso, and S.Schulz, editors, *Ontologien in Biomedizin und Lebenswissenschaften (OBML) 2010 Workshop Proceedings*, Mannheim.

[Zaveri et al., 2015] Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., and Auer, S. (2015). Quality Assessment Methodologies for Linked Data: A Survey. *Semantic Web Journal*.

[Zaveri et al., 2009] Zaveri, A., Shah, J., Pradhan, S., Ang, B. T., and Pietrobon, R. (2009). Center for Excellence in Research Reporting in Neurosurgery (CERR-N) - Achieving high quality through the use of computational ontologies. *Journal of Neurotrauma*, 26(8).

[Zaveri et al., 2013d] Zaveri, A., Vissoci, J. R. N., Daraio, C., and Pietrobon, R. (2013d). Using Linked Data to Evaluate the Impact of Research and Development in Europe: A Structural Equation Model. In Alani, H., Kagal, L., Fokoue, A., Groth, P., Biemann, C., Parreira, J. X., Aroyo, L., Noy, N., Welty, C., and Janowicz, K., editors, *12th International Semantic Web Conference (ISWC)*, volume 8219 of *Lecture Notes in Computer Science*, pages 244–259. Springer Berlin Heidelberg.

[Zaveri et al., 2014b] Zaveri, A., Vissoci, J. R. N., Westphal, P., Junior, J. R. N., de Andrade, L., Daraio, C., and Lehmann, J. (2014b). Using Linked Data to Build an Observatory of Societal Progress indicators Leveraging on Data Quality (Under Review). *Journal of Web Semantics*.

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemä aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, den 11.5.2015

Amrapali Zaveri