Methods for Matching of Linked Open Social Science Data

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
der Universität Mannheim

vorgelegt von
Benjamin Zapilko
aus Iserlohn

Mannheim, 2014

Dekan: Professor Dr. Heinz Jürgen Müller, Universität Mannheim
Referent: Professor Dr. York Sure-Vetter, Universität Mannheim
Korreferent: Professor Dr. Ansgar Scherp, Christian-Albrechts-Universität zu Kiel

Tag der mündlichen Prüfung: 30. Januar 2015

Abstract

In recent years, Semantic Web standards and technologies have matured. In particular, the concept of Linked Open Data (LOD), which describes methods for publishing, sharing and linking heterogeneous data according to established standards, has gained popularity and acceptance across various communities and domains. It has encouraged numerous research organizations, archives, libraries and governmental agencies to publish their data on the web. Science politics and organizations claimed that the potential of semantic technologies and data exposed in this manner may support and enhance research processes and infrastructures providing research information and services. With semantic technologies enabling a machine-interpretable processing of such semantically enriched data, researchers can be supported in their work.

In this thesis, we investigate whether these expectations can be met in the domain of the social sciences. In particular, we analyse and develop methods for matching social scientific data that is published as Linked Data, which we introduce as Linked Open Social Science Data. Based on expert interviews and a prototype application, we investigate the current consumption of LOD in the social sciences and its requirements. Following these insights, we first focus on the publication of Linked Open Social Science Data, since a complete publication is the basis for any further consumption, such as data matching. By extending and developing domain-specific ontologies for representing research communities, research data and thesauri, we achieve the necessary completeness and include all processes, data and structures of the social sciences. In the second part, methods for matching Linked Open Social Science Data are developed that address particular patterns and characteristics of the data. By developing assessment tests for statistical data, we enable gaining insight into whether a data set is technically and semantically suitable for a scientific analysis and whether two data sets can be matched for a combined analysis. We present two approaches for data matching considering two particular parts of Linked Data sets. The first approach focuses on datatype properties of Linked Data sets by utilizing regular expressions on their instance values. The second method matches object properties by considering the linked ontologies therein.

The results of this work contribute towards enabling a meaningful application of Linked Data in a scientific domain. Additionally, the developed ontologies and methods for data matching can be applied outside of the social sciences, since other domains and other data have comparable requirements for data publication and data matching.

Zusammenfassung

In den letzten Jahren sind Semantic Web Standards und Technologien weiter ausgereift. Besonders gewann das Konzept von Linked Open Open Data (LOD), das Methoden beschreibt, um heterogene Daten im Web nach offenen und etablierten Standards zu veröffentlichen und zu verlinken, an Popularität und wurde in verschiedenen Communities und Domänen positiv aufgenommen. Zahlreiche Forschungsorganisationen, Archive, Bibliotheken sowie staatliche Ämter sind seitdem dazu übergegangen, ihre Daten auf diese Weise im Web zu veröffentlichen. Wissenschaftspolitik und -organisationen sehen im Einsatz dieser Standards und Daten eine Unterstützung und Verbesserung von Forschungsprozessen sowie von Infrastrukturen, die Forschungsinformationen und darauf aufbauende Services anbieten. Mit Technologien des Semantic Web können semantisch angereicherte Daten und Informationen maschinell verarbeitet werden und für Wissenschaftler zugänglich und für ihre Arbeit nutzbar gemacht werden.

In der vorliegenden Arbeit wird untersucht, ob diese Erwartungshaltung am Beispiel der Sozialwissenschaften erfüllt werden kann. In einem für die Domäne typischen Anwendungsfall werden Verfahren analysiert und entwickelt, um sozialwissenschaftliche Daten, die als LOD veröffentlicht wurden, miteinander zu matchen. Hierfür wird der Begriff Linked Open Social Science Data eingeführt. Basierend auf Experteninterviews und eines technischen Prototypen wird die derzeitige Nutzung von LOD in den Sozialwissenschaften sowie die dafür nötigen Anforderungen untersucht. Diesen Erkenntnissen folgend wird im ersten Teil der Arbeit die vollständige Veröffentlichung von Linked Open Social Science Data ermöglicht, die eine notwendige Voraussetzung ist, diese Daten weiter zu nutzen und zu verarbeiten. Durch die Erweiterung und Entwicklung von domänenspezifischen Ontologien wird eine Repräsentation aller Prozesse, Daten und Strukturen ermöglicht, die in der sozialwissenschaftlichen Forschung relevant sind. Der zweite Teil widmet sich dem Matchen von Linked Open Social Science Data. Es werden Methoden entwickelt, die gezielt besondere Charakteristika dieser Daten berücksichtigen. Die Entwicklung von Assessment Tests ermöglicht, Einblick darüber zu gewinnen, ob ein oder mehrere Datensätze technisch und semantisch für das Matching geeignet sind. Außerdem werden zwei Matching Verfahren für Linked Data entwickelt. Das erste Verfahren nutzt die Instanzwerte in den Datatype Properties, um diese durch den Einsatz von regulären Ausdrücken zu matchen. Im zweiten Verfahren werden die Object Properties gematcht, indem der Overlap zwischen den in den Properties verlinkten Ontologien berechnet wird.

Die Ergebnisse dieser Arbeit tragen dazu bei, LOD und Semantic Web Technologien in einer wissenschaftlichen Domäne sinnvoll einzusetzen. Darüber hinaus können die entwickelten Ontologien und Matching Verfahren außerhalb der Sozialwissenschaften in ähnlichen Domänen mit vergleichbaren Anforderungen angewandt werden.

Contents

1 Introduction **1**

1.1 Motivation . . . 3

1.2 Objectives of This Work . . . 5

1.3 Research Questions . . . 7

1.4 Structure of the Thesis . . . 9

1.5 Publications . . . 11

2 Background & Related Work **13**

2.1 Linked Open Data . . . 13

2.1.1 Design Considerations and Modelling Patterns . . . 14

2.1.2 Publishing Linked Data . . . 19

2.1.3 Consuming Linked Data . . . 21

2.2 Statistical Linked Data . . . 23

2.3 Schema Matching . . . 28

2.3.1 The Matching Problem . . . 28

2.3.2 Classification of Matching Techniques . . . 30

2.3.3 Overview on Matching Systems . . . 32

2.3.4 Evaluation Methods of Matching Approaches . . . 33

3 Linked Data Consumption in the Social Sciences **39**

3.1 Qualitative Interviews . . . 39

3.1.1 Methodology . . . 40

3.1.2 Preparation and Conduction of Interviews . . . 42

3.1.3 Results . . . 43

3.1.4 Discussion and Limitations . . . 47

3.2 Prototypical Semantic Data Library for the Social Sciences . . . 47

3.2.1 Theoretical Background . . . 48

3.2.2 Application Scenario . . . 49

3.2.3 Framework for a Semantic Data Library . . . 51

3.2.4 Prototype Implementation . . . 57

3.2.5 Discussion and Limitations . . . 61

3.3 Summary . . . 63

4 Publication of Linked Open Social Science Data **65**

4.1 Levels of Publishing Linked Open Social Science Data . . . 65

4.2 Publishing the Research Processes of the Social Sciences 67
4.2.1 Analysis of Existing Approaches 68
4.2.2 Design Considerations for an Extension 71
4.2.3 Extending the Semantic Web for Research Communities (SWRC) Ontology 72
4.3 Publishing Person-level Research Data Using Data Documentation Initiative RDF (DDI-RDF) 75
4.3.1 Metadata Formats and Vocabularies for Representing Research Data 76
4.3.2 Data Documentation Initiative (DDI) Metadata Model 80
4.3.3 DDI as Linked Data 82
4.3.4 Model of the DDI-RDF Discovery Vocabulary (Disco) 83
4.3.5 Conversion Process of DDI-XML to DDI-RDF 85
4.3.6 Relationships to Other RDF Vocabularies 87
4.3.7 Discussion and Limitations 89
4.4 Publishing a Domain-specific Thesaurus with Linked Data 89
4.4.1 SKOS – The Simple Knowledge Organization System 91
4.4.2 Modelling the Thesaurus for the Social Sciences (TheSoz) Using SKOS 92
4.4.3 Establishing Links to Other Thesauri Using Semi-automatically Link Detection 101
4.4.4 Discussion and Limitations 113
4.5 Summary 121

5 Data Matching for Published Linked Open Social Science Data 123
5.1 Matching Statistical Linked Data 124
5.2 Assessment Tests for Statistical Linked Data 125
5.2.1 Basic Requirements for Valid Statistical Data 127
5.2.2 Rules for Assessing Statistical Data 128
5.2.3 Definition of Assessment Tests for Statistical Linked Data 130
5.2.4 Technical Implementation 133
5.2.5 Evaluation Setup and Results 137
5.2.6 Discussion and Limitations 140
5.3 Instance-based Schema Matching 142
5.3.1 Challenges of Schema Matching with Statistical Linked Data . . . 143
5.3.2 Schema Matching Using Regular Expressions 144
5.3.3 Algorithm and Implementation 147
5.3.4 Evaluation Setup and Results 149
5.3.5 Discussion and Limitations 156
5.4 Object Property Matching Utilizing the Overlap between Imported Ontologies158
5.4.1 Challenges in Object Property Matching 159
5.4.2 Use Case: Statistical Linked Data 160
5.4.3 Analysis of Patterns of Statistical Linked Data 161
5.4.4 Approach and Similarity Measures 163
5.4.5 Evaluation Setup 166

5.4.6 Benchmark for Statistical Linked Data 167
5.4.7 Results 169
5.4.8 Discussion and Limitations 171
5.5 Summary 172

6 Conclusion and Future Work 173
6.1 Summary of Contributions 173
6.2 Future Work 178

Bibliography 181

Appendix 207

A Guideline for the Qualitative Interviews 209

B Extension of the SWRC Ontology 213

C DDI-RDF Discovery Vocabulary 217

D Benchmark for Statistical Data 229

List of Figures

1.1 The LOD cloud diagram 2014 [SBJC14]. 2
1.2 Structure of the thesis. 9

2.1 Linked Data publishing recipes and workflows [HB11]. 19
2.2 Statistical Data organized in a Star Schema. 24
2.3 Example of Statistical Linked Data. 26
2.4 The matching process [ES07]. 29
2.5 Classification of matching approaches [ES07]. 31

3.1 Framework for a Semantic Data Library. 52
3.2 Visualization of a combined query on different data sets [ZHM11]. 55
3.3 Filtered Visualization with one party and answer type. 56
3.4 Implemented prototype. 59
3.5 Result output of the prototype. 61

4.1 DDI Data Life Cycle [BCWZ12]. 81
4.2 Overview of DDI-RDF model. 84
4.3 Excerpt from the RDFS- and OWL-based representation of DDI-RDF [BCWZ12]. 86
4.4 Example of a USE/USEDFOR relationship. 93
4.5 Example of a USE COMBINATION relationship. 95
4.6 Example of an Alternative Non-Descriptor (AD). 95
4.7 Example of a USE/USED FOR relation using SKOS extensions. 99
4.8 Example for a USE COMBINATION relation using SKOS extensions. . . 100
4.9 Example of an AD term using SKOS extensions. 101
4.10 Network of controlled vocabularies built in KoMoHe [MP08a]. 103
4.11 Example mapping between TheSoz and STW in SKOS. 106
4.12 Matching process workflow. 108
4.13 TheSoz implemented in Pubby. 113
4.14 Links from TheSoz to STW, AGROVOC and DBpedia. 114
4.15 Percentage of correct correspondences found by # matching systems [KREZ14]. 118

5.1 Matching of object properties linking to individuals of imported ontologies. 160
5.2 Example of statistical data represented as Linked Data. 161
5.3 Example individual of the seed ontology. 168

List of Tables

2.1 Statistical Linked Data used in this thesis. 27

4.1 New class swrc:Dataset. 72
4.2 Modifications on existing classes and properties. 73
4.3 New properties of the SWRC ontology. 74
4.4 Overview of SKOS classes and properties. 93
4.5 Classes and properties of SKOS-XL. 94
4.6 Mapping of TheSoz elements to classes and properties of SKOS. 97
4.7 SKOS extensions for TheSoz. 98
4.8 Mapping of TheSoz classification hierarchy to SKOS. 102
4.9 Overview on mapping types between TheSoz and STW. 105
4.10 Matching results generated by our initial approach and Silk. 108
4.11 Results of the OAEI Library Track 2012 [AEE$^+$12]. 111
4.12 Results of the manual evaluation [KREZ14]. 112
4.13 Number of correspondences: total; de-duplicated and correct [KREZ14]. . 117
4.14 Results of the majority vote [KREZ14]. 118
4.15 Comparison of different evaluation strategies [KREZ14]. 120

5.1 Overview of assessment tests. 130
5.2 Data sets used in the evaluation. 137
5.3 Results of Stage A. 138
5.4 Results of Stage A (continued). 139
5.5 Results of Stage B. 140
5.6 Summarized results for Stages A and B. 141
5.7 Overview of typical instance values and patterns. 144
5.8 Overview of instance values and their corresponding regular expressions for temporal information. 145
5.9 Overview of instance values and their corresponding regular expressions for geographical code lists. 145
5.10 Overview on instance values and their regular expressions for age groups. 146
5.11 Data sets used in the evaluation. 150
5.12 Matching tasks. 151
5.13 Summarized results of the matching systems for all tasks. 153
5.14 Results for our approach. 154
5.15 Results for COMA++. 154
5.16 Results for FALCON-AO. 155

5.17 Results for PARIS. 155
5.18 Comparison of Statistical Linked Data and OAEI data (as of December 2013). 162
5.19 Analysis of the structure of Statistical Linked Data (as of December 2013).163
5.20 Variations within the benchmark. 168
5.21 Results for both evaluation scenarios. 169
5.22 Similarity measures for detected correspondences (AgreementMaker). . . . 170
5.23 Similarity measures for detected correspondences (FALCON-AO). 171

1 Introduction

In recent years, the development of the Semantic Web, which aims to add more meaningful semantics to data and information on the web [BLHL01], has progressed and matured. The intention underlying the idea of a Semantic Web is not only to provide more information about things from the web to users, but also to develop better services and applications based on the possibilities of processing semantically rich data. In particular, the concept of Linked Open Data (LOD)[1] has gained popularity and acceptance in various communities and domains recently [BHBL09].

The paradigm of LOD describes methods to publish, share and link data of different kinds of structure and domains freely and openly on the web by applying Semantic Web-based technologies and standards. From a technical point of view, LOD is based on the unique and persistent identification of fine-grained data items – e.g. metadata elements, entities and values – via URIs (Uniform Resource Identifier), which can then be dereferenced via HTTP (Hypertext Transport Protocol) and provide meaningful and machine-interpretable information in RDF (Resource Description Framework), a graph-based data format. Finally, these data items can be linked to related and associated items of other data sets. Sir Tim Berners-Lee subsumes the idea of LOD in four principles [BL06], the so-called Linked Data principles:

1. 'Use URIs to identify things.
2. Use HTTP URIs so that these things can be referred to and looked up ("dereferenced") by people and user agents.
3. Provide useful information about the thing when its URI is dereferenced, using standard formats such as RDF.
4. Include links to other, related URIs in the exposed data to improve discovery of other related information on the web.' [BL06]

The concept of LOD has quickly encouraged organizations and institutions to publish their data according to these principles. Numerous research organizations, archives, libraries and governmental agencies are now participating in this movement worldwide. The domains of the published data range from cultural heritage in the widest sense (e.g. literature, music, media) over domain-specific scientific data (e.g. life sciences) and their schemata (e.g. classifications, ontologies, thesauri) to governmental data as well as data from the social web or user-generated content (e.g. blog posts) [BHBL09]. The links between those different data sets are visualized in the so-called LOD cloud diagram [SBJC14] (see Figure 1.1), which has become a popular symbol for the LOD idea.

[1] http://linkeddata.org

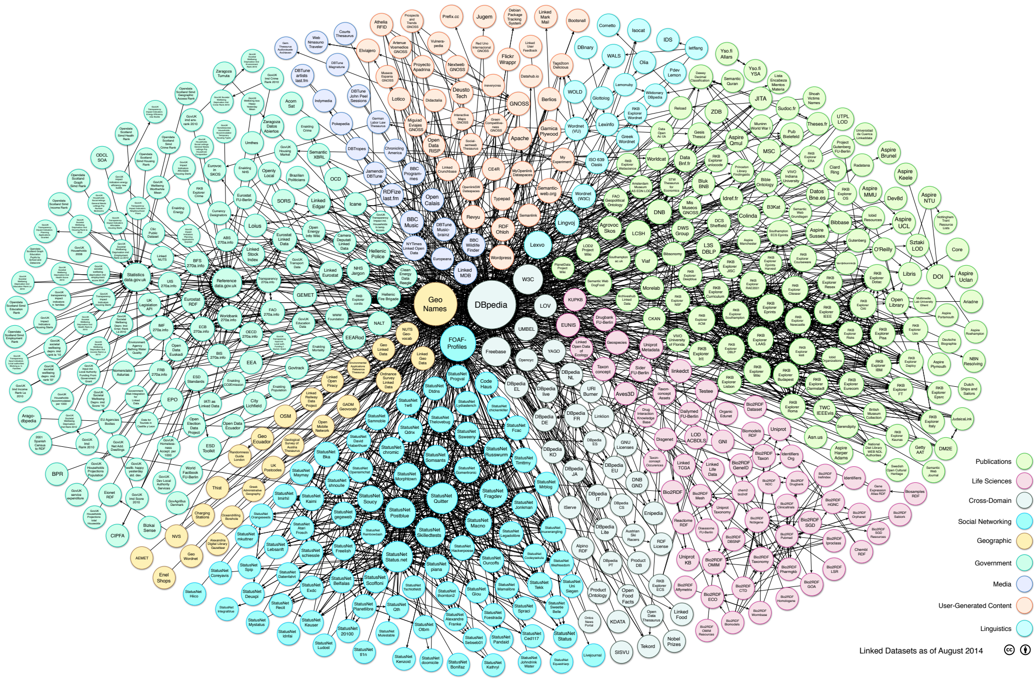


Figure 1.1: The LOD cloud diagram 2014 [SBJC14].

Due to the success of the cloud, the number of potentially interesting and relevant data sets for a scientific purpose has increased in recent years. Particularly for the domain of the social sciences, data from providers like, e.g. OECD[2], the World Bank[3] or Eurostat[4], are relevant sources for research. But the simple publication of interesting data as LOD is not sufficient for research purposes as it does not enable "reusable, shared research and the reproducibility" [BAB^{+}10] of data and results. Kauppinen et al. [KBK12] have introduced the concept of Linked Science in order to connect the idea of Linked Data and the Semantic Web with the need for their application in scientific processes, such as sharing, validating and evaluating the research results of scientific publications. Linked Science covers four aspects: Linked Data, open source and web-based environments, cloud computing and Creative Commons [KdE11]. These approaches claim that Linked Data can be useful in a scientific research process at least as a basis for the modelling and publication of data. However, in order to achieve a valuable use for a domain, its special characteristics and requirements have to be investigated.

The potential of semantic technologies and the need for innovation to support and enhance research processes and infrastructures have already been observed by science politics and organizations. In April 2011, the German Kommission Zukunft der Informationsin-

[2] http://www.oecd.org/home
[3] http://data.worldbank.org/
[4] http://epp.eurostat.ec.europa.eu/

frastruktur[5], which met under the auspices of the Leibniz Organization, approved a master plan [KII11] for the information infrastructure in Germany. In this report, the enrichment of research data by additional information (e.g. metadata) is discussed, as is interoperability and connectivity between research data sets. Furthermore, necessary integration and standardization processes are advised. These results are encouraged by the recommendations for research infrastructures in the humanities and the social sciences [Wis11] by the German Wissenschaftsrat[6]. Especially for the social sciences, a large potential regarding the search and analysis of research data, which is available as Linked Data, is expected by [GV10]. Also, [GCAR06] sees a large impact resulting from the adoption of Semantic Web technologies for eScience and its accompanied cyberinfrastructure. These can be technically and semantically enhanced by the Semantic Web. The use of Semantic Web concepts and technologies for digital libraries (DL), e.g. the use of ontologies, semantic annotations or inference engines [SS05], is discussed in a greater community, especially to improve issues concerning social and knowledge networking as well as interoperability [GCAR06]. Krause [Kra08] states that Semantic Web technologies can relieve negotiations about standardizations as well as support the identification and representation of mappings between different terminology systems. Krause [Kra08] also recommends 'that all individual DL developments should be checked even now for their adaptability to the W3C standards of the Semantic Web'. Potentials for Semantic Web technologies applied in digital libraries have also been raised by [Vat10, Sve07]. A hybrid model for an integrated retrieval of literature and research data in digital libraries, which tries to build the bridge between traditional content-indexing methods and ontology-based approaches, has been proposed by [SZ09b] and supplemented in [SZ09c, SZ09d, SZ09a].

This thesis aims to investigate whether and how far semantic technologies and the concept of LOD can be applied to the domain of the social sciences. This will be achieved using a concrete example of investigating and developing Semantic Web-based methods for matching research-relevant LOD sets. Matching heterogeneous data sets is a typical job during the analysis of research data [SHE05]. In the following Section 1.1, we briefly outline the motivation of this thesis. In Section 1.2, the objectives of this work are formulated; these are built on a typical use case from social science research, which is revisited in the course of the thesis in order to apply different stages of our work. The research questions are formulated in Section 1.3. In Section 1.4, we describe the structural composition of the thesis. Finally, Section 1.5 presents the publications and expositions that have been produced through the work on this thesis.

1.1 Motivation

The motivation of the thesis is based on two points of view. First, from a technical point of view, semantic technologies and especially the paradigm of LOD seem to be a suitable technical solution for publishing domain-specific data in a semantically enriched

[5] http://www.leibniz-gemeinschaft.de/?nid=infrastr
[6] http://www.wissenschaftsrat.de/1/home/

and standardized way on the web and to share, retrieve, combine and process it easily. With more data available for reuse, interesting and innovative web applications can be built that make it easier for a broader audience to consume and interact with complex domain-specific data. Second, from a domain-specific and scientific point of view, data on the web may be a relevant information source, but its processing in established tools, e.g. statistical tools, is not seamless because such tools usually support their own standards which are often not compliant with web standards. Despite originating from different directions, both points of view intersect by applying new technologies and standards to a scientific domain and its data. An overall question that needs to be investigated is whether claims of the Semantic Web community and expectations of a scientific domain can be fulfilled or whether both communities can complement each other. The research questions of this thesis will focus on a particular application scenario in this regard: an investigation of methods and approaches for matching LOD sets in the domain of social sciences.

Initially, we bring both together by defining the naming conventions. With the term Linked Open Social Science Data, we adapt the concept of LOD to the domain of the social sciences, as the application scenario of this thesis is located in this domain. The definition is composed of two concepts: LOD and Social Science Data.

Linked Open Data LOD describes methods and techniques for publishing, sharing and linking heterogeneous data on the web according to the standards of the Semantic Web. Details on this concept envisioned by [BL06] are given in the beginning of this chapter and in Section 2.1.

Social Science Data The term Social Science Data comprises data that is either generated or used during the research process in the social sciences (e.g. research data like survey or statistical data), but also data that documents this process, its activities, actors and results (e.g. bibliographic information, information on research projects, groups, organizations, etc.) as well as data that structures this documentation (e.g. classifications or thesauri to describe data and documents in a consistent way). The composition comprises processes, data and structures of the social sciences.

Linked Open Social Science Data By combining Linked Open Data and Social Science Data, we define Linked Open Social Science Data as Social Science Data that is published according to the Linked Data principles formulated by [BL06]. A sub-part of Linked Open Social Science Data is Statistical Linked Data, which describes statistical research data published as Linked Data. It will be introduced in detail in Section 2.2, since it will play an essential role in this thesis as a data basis for the methods for data matching.

In this thesis, we will use the term, Linked Open Data (or the shortened form, Linked Data), when referring to the underlying concepts and methods introduced by [BL06], and the term Linked Open Social Science Data when referring to Social Science Data published as LOD.

Linked Data vs. Record Linkage When adapting the technical concepts of LOD to the Social Sciences, a demarcation into existing and similar techniques in the social sciences is necessary. Since the definitions of the terms 'linked' and 'data' are kept very general in LOD as 'links between arbitrary things described by RDF' and 'data on the web' [BL06], it has to be distinguished from the linking of data performed in record linkage. In record linkage, a record from one data source (e.g. a database) is joined with one from another data source (e.g. another database) where both records describe the same thing [Dun46, Win06]. In social research, record linkage is a common technique that is often applied [SHE05]. While record linkage focuses on finding similar entities in different data sets, LOD aims to publish and connect data in a structured way in order to detect and represent related and additional information about a particular data entity. Indeed, this is often done by discovering similar entities (referred to as link discovery or entity resolution), but a link between similar entities in LOD is used more often to form a connection to the additional data attached to the detected entity.

1.2 Objectives of This Work

In this thesis, we aim to investigate methods for the data matching of statistical data as an example for a possible applicability of LOD in the domain of the social sciences, envisioned as Linked Open Social Science Data. Hence, we examine how Linked Open Social Science Data can be applied meaningfully in social science research in the concrete use case of 'Analysing Research Data', which is a typical task of researchers in the social sciences [SHE05]. This use case will be revisited throughout the entire thesis. We will adopt existing methods and technologies of the Semantic Web for matching Linked Data as well as develop new approaches that consider the particular requirements given by the domain of the social sciences.

For gaining insight into how Linked Data can be applied in a scientific domain, particularly the domain itself as well as typical activities and processes within it have to be investigated. Since the application scenario of the thesis is located in the social sciences, we have to examine how researchers of this domain conduct their research work. Similar to other scientific domains, searching, interacting, analysing and interpreting data is of high relevance in social science research. Researchers in the social sciences usually pass different stages during their research process, where they have to deal with different tasks and have different information needs. In empirical research, research data is commonly the centre of all activities [SHE05]. For secondary research purposes, i.e. where no surveying or collecting of data takes place, searching for data and information as well as the analysis of research data represent the most important phases. The tasks during these phases deal with existing data; hence, there are specific requirements regarding the data. These requirements are clarified within a typical use case scenario that is used throughout the thesis as an application example. The following subsection describes in detail the steps that researchers have to take during the secondary analysis of research data.

Use Case 'Analysing Research Data'

At the beginning of a secondary analysis, the seeking of relevant research data and information covering a particular topic of interest is one of the first steps. A qualified content indexing of such data leads to more precise search results [vR79]. Data and documents are described according to negotiated or accepted standards and their content is indexed by terms of specific classification systems or thesauri [KNS03]. Their granularity depends on the context and the discipline in which they are used. However, extensive data description and documentation is also required by researchers in order to judge its relevance and quality as well as its subsequent suitability regarding the research interest. This is especially important when seeking relevant research data, where e.g. a statistical assessment of data quality regarding bias and variance is necessary. Such a data description is usually more extensive than is needed for information retrieval purposes.

Research interests are often spread over more than one data set. It is quite common that multiple research data sets are needed in order to be compared or combined with each other. Thus, knowledge about comparable variables or indicators in other data sets is necessary. One of the reasons why this information is missing is that research data is most commonly held in a decentralized manner in governmental agencies, research data centres, archives or other research organizations. Even agreements on standards do not provide such link information or provide it only by covering very small amounts of data. The reason for this obstacle is that even when using standards, links between two or more pieces of data have to be identified and made available.

Social researchers may compare or combine data that does not seem semantically related at first sight [SHE05], e.g. unemployment rates and the fear of losing a job. In such cases, the schemata of the data sets have to be aligned in order to enable a combined analysis or at least one key variable has to be detected according to which the data sets can be merged. This task is referred to as data matching. Depending on the data source, metadata is often formatted in different ways, which complicates the analysis of such data. Traditionally, this work is carried out manually by researchers. They have to be aware of not only different metadata schemata, but also different code lists and classifications (e.g. for countries, age groups or occupations). Entries in code lists can differ in both data sets through different classification systems, naming conventions, abbreviated terms, different granularity levels, etc. For a combined analysis of different data sources, their schemata, the single schema elements (e.g. spatial coverage) as well as the contained instance values (e.g. entries of code lists) have to be mapped to each other.

There are various different ways to analyse research data, e.g. comparative analysis, time series analysis or estimation procedures (for more details, see e.g. [SHE05, KKV94]). The different methods that can be applied to research data depend massively on the questions and research intentions of the scientist. Research data is typically analysed in established statistical tools, which can compute various statistical methods (e.g. t-tests, calculations of variance and regression) and can generate output graphics like diagrams and graphs. Statistical tools provide an overall framework and support researchers in their work.

According to this use case, researchers still need to do a lot of manual work during these steps. Research data often has to be converted to specific formats of statistical tools or made semantically comparable to other research data, before it can be used for further research. This can be a very time-consuming and tedious task. Although there are established standards, research data differs widely in its format, structure and semantics. The manual work increases, because most of the research data used by scientists is traditionally obtained from different sources. Only in a few cases can researchers be supported in their work by software tools, e.g. simple data conversion or the simple matching of code lists.

In social research, scientists are often confronted with sensitive data, e.g. survey data that underlies special licence and privacy restrictions. However, this thesis focuses only on proven and available research data. Aspects of trust and privacy will not be discussed. Related work concerning these topics is recommended in Section 2.1.3.

It seems reasonable that the methods and technologies behind the LOD paradigm can be applied to the social science domain. Especially the standardization and structural format that accompany Linked Data may enable better retrieval, interpretation, processing and combination of data and information, but also an inclusion of such data into research. Whether this impression proves correct or otherwise, will be investigated in this thesis at the application scenario of data matching.

1.3 Research Questions

This PhD thesis aims to investigate methods for matching statistical data in the context of Linked Open Social Science Data. It is examined whether concepts and technologies of the Semantic Web can be applied within typical working tasks of the data matching process and whether Linked Open Social Science Data can be used as a data basis for this process. The goal is to enable the use of data published as Linked Data for social science research as well as to support and (semi-)automate data matching, which often have to be conducted manually by researchers. In this thesis, we will investigate the following research questions.

1. General Applicability of LOD for Data Matching

1a. Can Linked Data be applied for data matching, as it is conducted in social research?

1b. Which characteristics of the data have to be considered?

1c. Which requirements have to be met by matching systems?

The first research questions of this thesis focus on the possibility of whether Linked Data and Semantic Web technologies can be applied in data matching. This examination includes the technologies and methods themselves as well as particularities of the data that is to be matched.

2. Suitability of LOD for Data Matching

2a. Can the suitability of Linked Data for data matching be determined?

2b. Does such a suitability have an impact on the matching process?

2c. Is it possible to determine whether the matching of particular Linked Data sets will be successful before the matching process?

Since data matching can be a time-consuming task, especially when the data sources have to be inspected beforehand, it may be worthwhile to ascertain whether the effort involved in matching is reasonable. Furthermore, Linked Data sets may not be easily inspected by users with limited technical experience who are not familiar with their RDF representation.

3. Influence on the Matching Result

3a. Is it possible to influence the matching result by a targeted application of matching systems that consider the particular characteristics and requirements of Linked Data?

3b. What kind of impact on the results is measurable by such matching methods in comparison to state-of-the-art approaches?

It will be investigated whether particular characteristics and structures of data typically used in social research can be addressed particularly by matching systems and whether the matching results can be influenced.

4. Limitations of the Approach and Applicability on Other Data Sources

4a. Which limitations of these targeted matching methods can be observed in comparison to state-of-the-art matching systems?

4b. Can these methods be applied to other Linked Data sources?

We will examine the general applicability of the developed methods for matching Linked Open Social Science Data and will report where we have identified limitations.

5. Beneficial Use of LOD in the Social Sciences

5a. Can a value addition through the application of Semantic Web technologies and Linked Data in the social sciences be observed?

5b. Which preconditions have to be met?

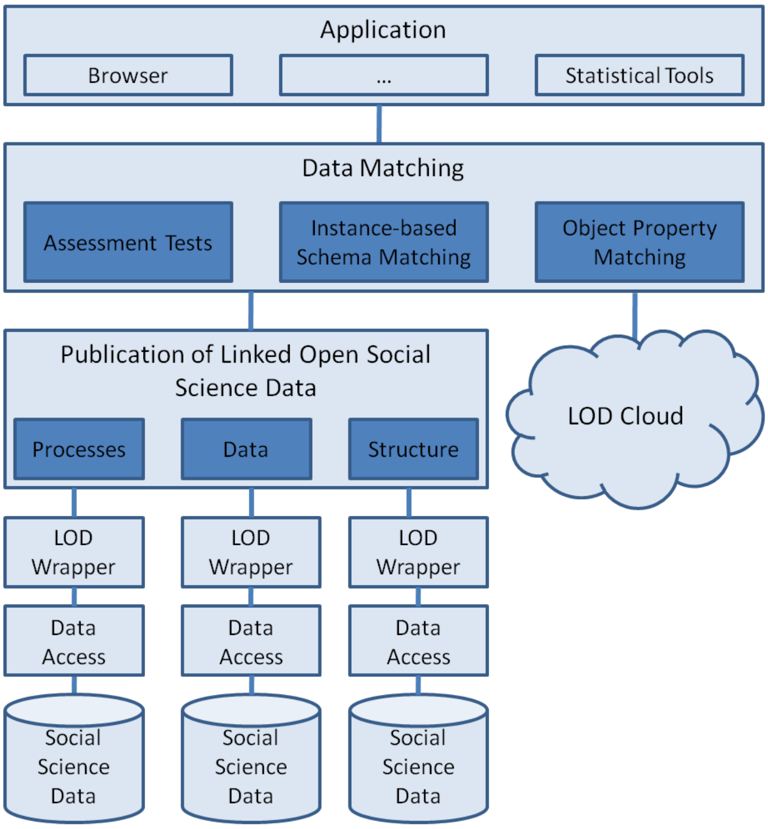


Figure 1.2: Structure of the thesis.

Finally, we will investigate whether a beneficial use of Linked Data in the social sciences can be observed using the developed methods.

1.4 Structure of the Thesis

This thesis comprises six chapters. Followed by this introductory chapter, we present an overview on background material and related work relevant to the main topics of the thesis in Chapter 2. In Section 2.1, an overview of LOD, its conceptual ideas and technical standards is presented. This serves as a foundation for Chapters 3 and 4. In Section 2.2, we introduce the term Statistical Linked Data, since it will be a major component of the approaches presented in Chapter 5. We provide the fundamentals of schema and ontology matching in Section 2.3, which also serves as a basis for the later work in Chapter 5.

Figure 1.2 illustrates the structure of the thesis by means of a software architecture for consuming Linked Open Social Science Data. In Chapter 3, we proceed from the top of the architecture (i.e. the application layer) in order to identify which issues have to be addressed for meaningful applicability of Linked Open Social Science Data in terms of data matching. Thus, we analyse the current consumption of LOD in the social sciences. First, in Section 3.1, we conduct expert interviews with researchers from the social sciences. In Section 3.2, a prototype application for a Semantic Data Library and its requirements are presented and analysed. The findings of these sections form five topics of interest: data modelling, data access, data retrieval, data matching, and data interaction. Since our objective is to investigate methods for data matching, we continue to examine data modelling (since the publication of data is the foundation for its use) and data matching in the following chapters in detail.

Chapter 4 focuses on the modelling and publication of Linked Open Social Science Data, since this builds the foundation for our work on data matching. This is examined in three examples that consider the processes, data and structure of Social Science Data (see Figure 1.2). The first example (processes) introduces an extension of an existing ontology (SWRC – Semantic Web for Research Communities)[7] [SBH$^+$05] by classes and properties that support the representation of social science research processes, activities and entities. In the second example (data), a widely applied metadata format for documenting social science research data, called DDI (Data Document Initiative)[All], is conducted into a RDF representation in order to represent person-level research data (e.g. survey data) as Linked Data. Finally in the third example (structure), a thesaurus, which is a crucial instrument for information retrieval in document and data collections, is converted into the SKOS (Simple Knowledge Information System) format [MB09b], a popular Semantic Web vocabulary for the representation of terminologies, classifications and thesauri. Using these examples, we enable a complete publication of Linked Open Social Science Data.

The problem of data matching is discussed in Chapter 5. Again, three methods for supporting and improving data matching are presented (see Figure 1.2). In this chapter, we focus on Statistical Linked Data (introduced in Section 2.2) as part of Linked Open Social Science Data. Necessary assessment tests for a scientific analysis of Statistical Linked Data sets are presented in Section 5.2. This approach allows for assisting non-expert researchers in deciding whether particular Statistical Linked Data sets are technically and semantically suitable for use in a scientific analysis and whether they can be matched with particular other data sets. An instance-based schema matching approach, which considers typical patterns of instance values of statistical data for finding similar schema elements, is presented in Section 5.3. Finally, we present an approach to match the object properties of Statistical Linked Data by utilizing the overlap between the imported ontologies in Section 5.4. By these assessment tests and the treatment of datatype and object properties, we provide a full consideration of matching Statistical Linked Data.

In the concluding Chapter 6, our research questions are answered and our contributions are discussed. Finally, we present an outlook on ongoing and future work.

[7] http://ontoware.org/swrc/

1.5 Publications

This work is partly based on former publications and presentations that have been presented at national and international conferences and workshops and that discuss different aspects of this work.

LOD Prototype Application for the Social Sciences In [ZHM11], the general use and potentials of LOD for enriching and analysing statistics is proposed. This publication has influenced the follow-up publication [GHHZ11], which adopted and supplemented the use case and implemented a prototype covering the main aspects of the approach. The work presented in [GHHZ11] builds the major part of Section 3.2. I worked jointly with the co-authors on concept and technical implementation of the prototype and developed the use case for the social sciences.

Publication of Linked Open Social Science Data Section 4.3, which describes the DDI-RDF Discovery Vocabulary, is based on [BCWZ12, BZWG13] and the results of two workshops conducted in 2011 (see the introduction of Section 4.3 for the acknowledgement of the participants). During these workshops, the model and use cases described in [BCWZ12] were defined and discussed by the participants at the events. Together with two participants of the workshops, I implemented the conversion scripts in XSLT. In Section 4.3.3, I also add an explicit subsumption of DDI into the context of Linked Data as well as a more extensive discussion on different formats for describing research data (Section 4.3.1).

The transformation of the Thesaurus for the Social Sciences, described in Section 4.4, has been planned, conducted and introduced by myself in [ZS09a]. I have revised this first implemented version with self-defined extensions in [MZS10a], added links to other thesauri [MZS10a, MZS10b, MZJ$^+$11] and established a multi-thesauri setting. In [ZSMM13], the resultant and mature SKOS version of the thesaurus is presented. All publications are revised and put into context in Section 4.4. The SKOS version of the thesaurus was the first dataset of GESIS, which has been included within the LOD cloud diagram.

Work in the context of the Ontology Alignment Evaluation Initiative (OAEI) is described in Section 4.4.3.4 and work based upon these results is presented in Section 4.4.4.3. Both have been conducted together with the co-authors of [KZ13, KREZ14]. I contributed by co-organizing the Library Track since OAEI 2012 [AEE$^+$12] and by supporting the domain expert during the manual evaluation of the detected mappings. Together with the domain expert, I also performed the evaluation in [KREZ14].

Data Matching The basic idea for the assessment tests for analysing statistical data as published in [ZM11a] has been developed by me together with the co-author. The implementation and evaluation of the tests has been conducted by me. This publication has been the basis for Section 5.2 and for further research in the field of schema matching with statistical data. The approach has been revised and extended for this thesis.

In [ZZS12], I developed the approach of instance-based schema matching considering the patterns of instance values. I defined the setting and environment for the experiment, conducted the experiment together with the co-authors and supported the technical implementation of the jointly developed algorithm. This publication has built the basis for Section 5.3. However, the approach has been revised and extended for this thesis.

The work presented in Section 5.4 is based on [ZM14]. The concepts and the identification of the problem statement, the underlying analysis of Statistical Linked Data as well as the implementation, the development of the benchmark and the evaluation of the approach have been conducted by me.

Further Publications and Preliminary Work Additional published work has influenced this thesis. A model of Text-Fact-Integration, which organizes an integrated retrieval on literature and research data by combining traditional content-indexing methods with ontological approaches, has been presented in [SZ09b] and revised and supplemented in [SZ09c, SZ09a, SZ09d]. To these publications, I contributed parts of the overall concept of the model and the majority of the development of the model itself. Additionally, in [ZS09b, ZS09c], I investigated how semantic technologies can be applied for Digital Libraries. These publications have also supported the motivation of the thesis.

The following publications show a meaningful applicability of Linked Data and semantic technologies in the context of the social sciences. The concept for performing statistical calculations on Linked Data [ZM11b] has been developed jointly by both authors, while the technical implementation has been carried out by me. This publication is included as Section 6.2. I have contributed to [WBZMZ13] by envisioning a semantic technology-empowered infrastructure for a social science research organization. In [BZT13], which describes initial works regarding a data restore model for enabling the reproducibility of research experiments, I supported the author in the joint development of the concept.

Publications with minor influence on this thesis are [HZSM11a, HZSM11b, HZSM12, WWH$^+$10, WAZM10, BZTM11]. In [HZSM11a, HZSM11b, HZSM12], the use of Open Data and LOD in visualization tools is presented. In [WWH$^+$10], we have jointly developed concepts and methods for connecting and linking data holdings at GESIS. In [WAZM10], an approach has been presented for extracting and collecting snippets, which describe relations between Wikipedia articles. I have supported the author by designing the use case for that approach. In [BZTM11], I have supported the development of a concept for linking social network sites with scholarly information portals by using SKOS thesauri.

2 Background & Related Work

In this chapter necessary background information and related work according to the objectives of this thesis is presented. In regard to the use case addressed in Section 1.2 and the research questions raised in Section 1.3, the main research fields covered by this thesis are LOD and schema matching. These topics are fundamental for Chapters 3, 4, and 5. Additionally, we define the term Statistical Linked Data in order to specify research data published as LOD. Further related work regarding particular standards, models, methods, and approaches presented in the Chapters 3, 4, and 5 are discussed in the particular chapters.

In Section 2.1, an overview of LOD is given. Since the publication of data as Linked Open Data is the basis for its consumption, i.e. for data matching, we cover all aspects from the modelling and publication of data as Linked Data to the consumption of such resources by users or applications. This chapter also serves as a technical foundation for the work presented in Chapters 3, 4 and 5.

Since research data plays a significant role in scientific research and the number of research data published as Linked Data has increased, we define the term Statistical Linked Data in Section 2.2. This term comprises Linked Data sets, which represent data commonly serving as source for research, e.g. statistical or survey data. Statistical Linked Data is a sub-part of Linked Open Social Science Data as introduced in Section 1.1 and is used for the methods developed in Chapter 5.

Finally, Section 2.3 provides an overview of schema matching. The problem of matching is described as are the different approaches for treating this challenge. The matching of heterogeneous data is necessary when researchers intend to compare multiple datasets or examine possible correlations. Thus, this section builds the technical foundation for the work presented in Chapter 5.

2.1 Linked Open Data[1]

The paradigm of LOD [BL06] is derived from the Semantic Web, which addresses the need to enhance the web and resources published on the web with more expressive

[1] Since the openness of data is often an issue of privacy, licensing and political laws, we are using the terms ‘Linked Open Data’ and ‘Linked Data’ interchangeably throughout the thesis. Both terms refer to the same vision created by [BL06]. From a technical perspective, Linked Data cannot only be published and consumed on the web, but also in similar internal or restricted infrastructures, that have recently been called Linked Closed Data [CBG^{+}11].

semantics [BLHL01]. Instead of only enriching data separately with additional semantics, one of the main elements of the web, hyperlinks to other resources came into play. By establishing and describing hyperlinks between web resources that are enhanced by additional information, a network of semantically interlinked web resources arises. Furthermore, the understanding of resources has been expanded from web documents to real-world objects and concepts, even of an abstract nature. The idea of Linked Data was introduced by Tim Berners-Lee in [BL06]. Heath et al. [HB11] summarize the benefits of Linked Data as follows:

> ‘Linked Data provides [...] a unifying data model [...] a standardized data access mechanism [...] hyperlink-based data discovery [...] self-descriptive data’.

In the following subsections, we provide an overview of the modelling and design consideration of Linked Data and the approaches for publishing Linked Data technically. We also describe how Linked Data can be searched and consumed on the web in general.

2.1.1 Design Considerations and Modelling Patterns

The design considerations and modelling patterns for LOD are extensively described in detail in [BHBL09, HB11, BCH07, DD12]. The following sections comprise the most relevant design and modelling issues with regard to the research questions and use cases of the thesis alongside the four Linked Data principles introduced in [BL06]. They are picked up again in Chapter 4, where they are considered for the modelling and standardization of social scientific data.

URIs for Naming Things

The identification of things is the basis for publishing and linking them. Thus, URIs are used to identify resources on the web. This may include not only web documents, but also a description of any object or concept of the real world. Heath et al. [HB11] argue that URIs allow ‘a simple way to create globally unique names in a decentralized fashion’. They also state that URIs provide not only names for things, but also the possibility to access information about the resource directly.

By creating URIs for identifying things it is important that they are sufficiently expressive and self-descriptive so that third parties can imagine the resource to which they refer. Besides the technical requirement that URIs should be stable and persistent, [HB11] recommend using Cool URIs [SC08] in the context of Linked Data.

Bizer et al. [BCH07] state three different types of URIs for a single resource, by which a generic identification of a particular resource, a HTML representation of the resource, and a RDF representation of the resource can be distinguished. Examples for these different types of URIs are:

‘1. http://dbpedia.org/resource/Berlin
2. http://dbpedia.org/page/Berlin
3. http://dbpedia.org/data/Berlin
or
1. http://id.dbpedia.org/Berlin
2. http://pages.dbpedia.org/Berlin
3. http://data.dbpedia.org/Berlin
or
1. http://dbpedia.org/Berlin
2. http://dbpedia.org/Berlin.html
3. http://dbpedia.org/Berlin.rdf’ [BCH07]

Bizer et al. [BCH07] also recommend defining URIs only in namespaces and domains that are controlled by the provider of the exposed resources and to leave the details of technical implementations out of URIs, e.g. no server ports, etc. Additionally, several patterns and best practices for constructing URIs can be found in [DD12].

Dereferenceable HTTP URIs for Looking up Things

Since Linked Data is built upon the web architecture, the given URIs should be dereferenceable. This means that each URI can be looked up by HTTP clients, e.g. web browsers, and can provide a description about the identified resource. Content negotiation [Fie99] allows distinguishing between machine-processable and human-readable information by sending HTTP headers including the preferred type of description. Thus, servers can choose for any incoming requests the appropriate response types, e.g. HTML for humans and RDF for machine processing. Heath et al. [HB11] outline two strategies for making URIs dereferenceable: 303 URIs and hash URIs, both of which are described in detail in [SC08].

- 303 URIs are using the HTTP response code ‘303 See other’. This strategy is often used when identifying real-world objects with an URI, which is obviously not inside the web, e.g. the city of Berlin. By retrieving a 303 response, this particular URI is referred to another URI, which provides a description about the real-world object, e.g. a description about the city of Berlin.
- The hash URI strategy applies hash symbols at the end of URIs in order to address special parts of a URI. Such separated parts can be called directly by HTTP clients, but the particular URI without any hash symbol cannot be retrieved directly. In this way, the URI can be used to identify a real-world object, while the special parts of it addressed by hashes can be treated as special representations of the object.

Advantages and disadvantages of both strategies are discussed in detail in [SC08]. Heath et al. [HB11] states that ‘303 URIs are often used to serve resource descriptions that are part of very large data sets’, while ‘Hash URIs are often used to identify terms within RDF vocabularies’.

Providing Useful Information in RDF

RDF (Resource Description Framework) [KC04] is a simple graph-based data model that has been designed for publishing structured data on the web. In RDF, data is modelled in the shape of triples. Each triple is seen as a statement consisting of a subject, a predicate and an object.

```
http://geonames.org/cities/berlin rdfs:label Berlin
```

Heath et al. [HB11] distinguish between two types of RDF triples: Literal Triples and RDF Links. Both types differ in the object of their statements. Literal Triples contain a plain literal as object, which is sometimes accompanied with a language tag, e.g. `'Berlin'@de`. In contrast, RDF Links contain a URI as object, which refers to another resource, identified by this particular URI.

```
http://geonames.org/cities/berlin foo:isCapitalOf
    http://geonames.org/countries/germany
```

Heath et al. [HB11] list several benefits of using the RDF data model for LOD. HTTP URIs and RDF fit together because both are designed for use on a global scale. Setting up links between different sources is enabled through the use of RDF. Information of different sources can easily be merged into to one graph. This implies an independence of schemata because the use of different schemata inside one graph is possible. Furthermore, the complexity of the structure in which information is described can be leveraged by using schema languages like RDF Schema [BG04] or OWL [MvH04].

However, three features of RDF are identified by [HB11], which should be avoided, if possible, in context of Linked Data: RDF reification ('reified statements are rather cumbersome to query' [HB11]), collections and containers ('also problematic if the data needs to be queried with SPARQL' [HB11]), and blank nodes ('it is not possible to create RDF links to them from external documents' [HB11]).

There exist several vocabularies for describing resources in RDF, each of them covering specific domains, contexts or types of resources to be described. According to [BCH07, BHBL09] it has been considered as good practice 'to reuse terms from well-known RDF vocabularies such as FOAF [Bri10], SIOC [BB10], SKOS [MB09b], DOAP [Dum], vCard [HISW10], Dublin Core [Ini], OAI-ORE [LdS08] or GoodRelations [Hep08] wherever possible in order to make it easier for client applications to process Linked Data' [BHBL09]. New terms or complete vocabularies should be defined only if the required terms or relations are not available in existing vocabularies [BCH07]. An overview of the vocabularies used for data sets of the LOD cloud diagram [SBJC14] is generated in the project Linked Open Vocabularies [LOV12], which not only describes the used vocabularies using metadata, but also exposes links between different vocabularies.

Linked Open Vocabularies (LOV)[2] provide an overview of the RDF(S) and OWL vocabularies that are used in the Linked Data cloud. According to LOV, popular vocabularies for describing RDF resources are::

[2] http://lov.okfn.org/dataset/lov/

- **Dublin Core - DCMI Metadata Term [Ini].** All terms of the Dublin Core Metadata Initiative are also included into the same entitled RDF vocabulary. Dublin Core provides a basic core of terms for describing resources (e.g. a publication) and agents (e.g. a creator or an editor) and activities involving them.
- **FOAF - Friend of a Friend vocabulary [Bri10].** This vocabulary is used to represent people, social groups and their relationships to each other. Also, some terms for describing online activities and accounts are available.
- **SKOS - Simple Knowledge Organization System [MB09b].** The SKOS vocabulary focuses on the description of knowledge organization systems (KOS) like thesauri, taxonomies or nomenclatures. The terms and entries of such systems are modelled as concepts that are connected to each other via hierarchical or associative relationships. We describe SKOS in detail in Section 4.4.1 and utilize it for the representation of a domain-specific thesaurus in Section 4.4.2.
- **SIOC - Semantically-Interlinked Online Communities [BB10].** SIOC allows for describing information and relationships of online communities like message boards, wikis or blogs.
- **BIBO - The Bibliographic Ontology [DG08].** This ontology allows the representation of bibliographic records, but also of relationships between agents, activities and entities involved in the publication process.
- **EVENT - The Event Ontology [RA07].** With properties and classes of this ontology, it is possible to model events in relation to when and where they took place and which agents were involved.

Another important issue when describing data in RDF as Linked Data is to decide what further information should be added. Bizer et al. [BCH07] state that besides the description of a particular resource, associated backlinks (although redundant), descriptions of related resources and metadata about the resource itself (e.g. provenance information, licensing terms, creation data) should be included into a response to a HTTP client. Especially the provision of metadata information is claimed by [BCH07, BHBL09, HB11, BL06] as highly relevant for the utilization of Linked Data by data consumers. The Vocabulary of Interlinked Datasets (voiD) [CZAH11] provides the possibility to describe published Linked Data sets according to particular metadata terms, e.g. data provider, topics of the data set, number of triples, number of links to other data sets, etc. Provenance information about the data provides descriptions regarding the creator of the particular resource as well as regarding the creation methods and dates of the data [HZ09]. The PROV Ontology [LSM13] can be used to encode such information.

Including Links to Other Resources

Since hyperlinks are a fundamental element of the web, they are also fundamental for Linked Data. Hyperlinks are used for connecting different identified resources with each other. In the context of Linked Data they can be described as an RDF triple.

Heath et al. [HB11] categorize three different types of RDF links: Relationship Links, Identity Links and Vocabulary Links. Relationship Links refer to somehow related resources of a particular resource, e.g. associated persons, books. Identity Links address a different description of the same resource from a different data source, e.g. another description of the city of Berlin. These links enable the retrieval of additional information about a resource even from a different perspective or context. Vocabulary Links provide definitions of specific terms in the same or other vocabularies. These links support the self-descriptiveness of data and resources.

There are two ways to establish links to other resources: manually and (semi-)automatically. Setting up links manually can lead to major efforts according to the size of the source data set and the size and amount of targeted data sets. It is recommended that either targeted data sets should be chosen or tools providing a keyword-like search on URIs should be used. Sindice[3] [TDO07] and Falcons [4] [CQ09] create an index of URIs that can be searched for adequate candidate URIs. For large or multiple target data sets or if there is no target data set, links should be generated with the help of (semi-)automatic approaches. Recently, various Link Discovery Tools have been developed that mostly follow similarity-based or machine-learning techniques. Examples of similarity-based approaches are Silk [VBGK09], LIMES [NA11], SERIMI [AHSdV11] and Amalgame [vOHdB11], while RiMOM [LTLL09] builds on machine-learning techniques.

The linking of instances is also a well-known problem in the research area of databases (often known as entity resolution, record linkage or duplicate detection) [Win06, EIV07]. As Linked Data sets are often available as A-Box ontologies (containing the instances) and come along with an underlying schema, the T-Box ontology, the problem of instance linking is also addressed by schema and ontology matching approaches (see Section 2.3 for more details). Because of the substantial meaning of instance linking methods for Linked Data and for Ontology Matching, the Ontology Alignment Evaluation Initiative (OAEI) [OAE] has picked up an instance matching track since 2009.

The Five Stars of Linked Data

In 2010, Tim Berners-Lee added a five-star rating system to his Linked Data design issues [BL06] for awarding data sets regarding if and to what extent they are published as LOD.

- ‘1 Star: Available on the web (whatever format) but with an open licence, to be Open Data
- 2 Stars: Available as machine-readable structured data (e.g. excel instead of an image scan of a table)
- 3 Stars: as (2) plus non-proprietary format (e.g. CSV instead of excel)

[3] http://sindice.com/

[4] http://iws.seu.edu.cn/services/falcons/objectsearch

- 4 Stars: All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
- 5 Stars: All the above, plus: Link your data to other people's data to provide context' [BL06]

2.1.2 Publishing Linked Data

Bizer et al. [BCH07] introduced several methods for publishing Linked Data technically, which consider different infrastructure situations, e.g. how is the data stored and accessible originally and how it should be published on the web. These methods are revised in [HB11] and defined as six publishing recipes that are depicted in Figure 2.1 (taken from [HB11]) and are summarized briefly in the following list.

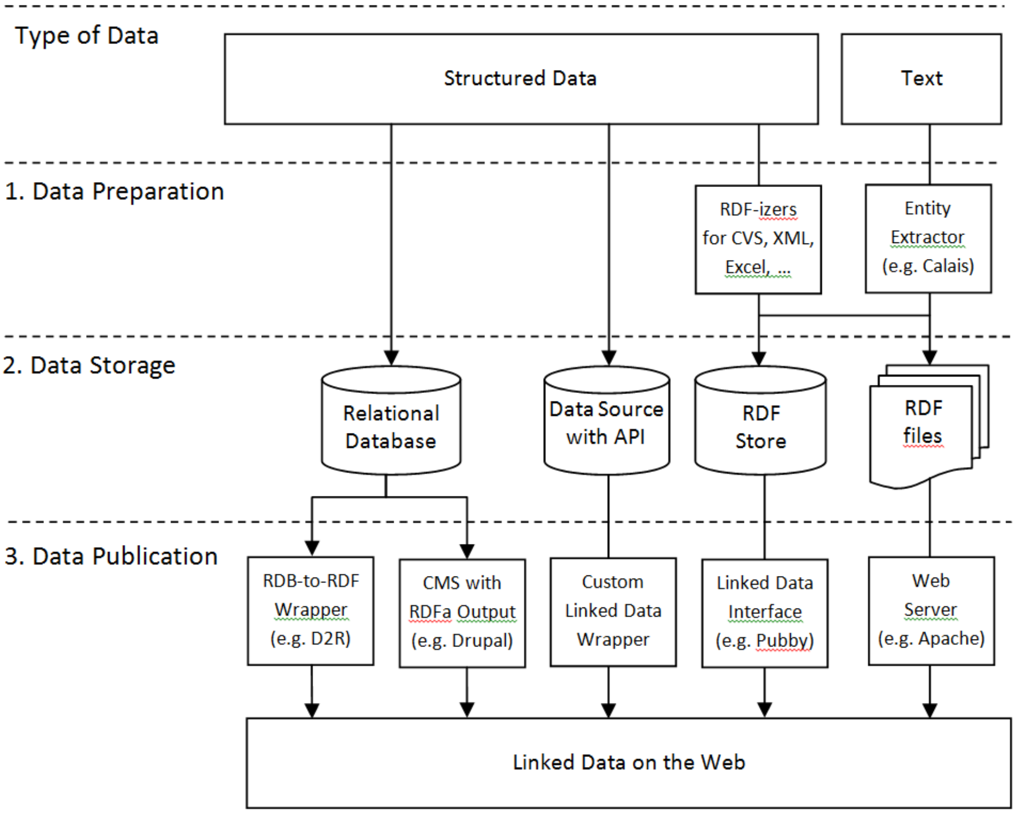

Type of Data
Structured Data
Text
1. Data Preparation
RDF-izers for CVS, XML, Excel, ...
Entity Extractor (e.g. Calais)
2. Data Storage
Relational Database
Data Source with API
RDF Store
RDF files
3. Data Publication
RDB-to-RDF Wrapper (e.g. D2R)
CMS with RDFa Output (e.g. Drupal)
Custom Linked Data Wrapper
Linked Data Interface (e.g. Pubby)
Web Server (e.g. Apache)
Linked Data on the Web

Figure 2.1: Linked Data publishing recipes and workflows [HB11].

- **Publishing Linked Data as Static RDF Files.** This recipe describes a simple way to publish static RDF files on a web server. This method is suitable for data that changes infrequently or not at all, e.g. small files or metadata vocabularies, where only from time to time is a new version published.

- **Publishing Linked Data as RDF Embedded in HTML Files.** This method includes RDF data into HTML pages by using RDFa [AHSB13], which allows an easy enhancement of generic HTML output with Linked Data content. The use of RDFa can be useful, e.g. in content management systems, when templates are used to generate an output [HB11]. Drupal[5] supports the publication of RDFa content since version 7.
- **Publishing Linked Data by Custom Server-side Scripts.** Many web applications are using custom scripts, e.g. PHP or methods for generating HTML output. In such a scenario, custom server-side scripts addressing particularly the existing infrastructure have to be developed.
- **Publishing Linked Data from Relational Databases.** In many organizational infrastructures, data is stored in relational databases and it is desired to retain existing database management infrastructures. For such cases, publishing 'a Linked Data view of the relational databases' [HB11] should be considered. There are several approaches for mapping relational databases to RDF, some of which define a mapping between RDF and the database schemata (e.g. D2R Server [BC06], OpenLink Virtuoso[6]) or between RDF and SQL queries (e.g. RDQuery [PZC06]) resp. their results, e.g. Triplify [ADL+09]. A detailed overview of existing approaches can be found in [SHH+09]. Currently, the W3C RDB2RDF Working Group[7] develops a generic language for mapping relational databases to RDF. The aimed result, the R2RML (RDB 2 RDF Mapping Language) [DSC12], has the status 'W3C Recommendation' since September 2012[8].
- **Publishing Linked Data from RDF Triple Stores.** Typically, RDF triple stores (e.g. AllegroGraph[9], OpenLink Virtuoso, OWLIM[10], 4store[11], Mulgara[12]) provide a Linked Data frontend for publishing data on the web. However, since not all triple stores allow such a frontend, approaches like Pubby [CB07] have been developed, which query the SPARQL endpoint of a triple store and generate a web-based output.
- **Publishing Linked Data by Wrapping Applications or Web APIs.** Many websites expose their content or their data via web services or APIs. In order to retrieve this data as Linked Data, it is common to implement a wrapper around the existing interfaces, which retrieves the data via the API and converts it into Linked Data.

The publishing patterns are examined further in later sections of the thesis. In Section 4.4.3.5, a domain-specific thesaurus is published as a static RDF file via the Pubby Linked

[5] http://drupal.org/
[6] http://virtuoso.openlinksw.com/dataspace/dav/wiki/Main/
[7] http://www.w3.org/2001/sw/rdb2rdf/
[8] http://www.w3.org/TR/r2rml/
[9] http://www.franz.com/agraph/allegrograph/
[10] http://www.ontotext.com/owlim
[11] http://4store.org/
[12] http://www.mulgara.org/

Data Frontend [CB07]. In Section 3.2.4, data from a statistical agency is published as Linked Data by a wrapper that is built upon a web API.

2.1.3 Consuming Linked Data

Heath et al. [HB11] state that consumption of Linked Data is typically enabled by creating a Linked Data mashup. A mashup is built by starting with data of a specific regard and is then accumulated with additional data that is somehow associated and, in the context of Linked Data, connected to it. There are different ways of creating such a Linked Data mashup. Heath et al. [HB11] highlight three so-called architectural patterns for Linked Data applications, which are briefly described in the following paragraphs.

- **The Crawling Pattern.** Available Linked Data on the web is crawled by applications beforehand and is integrated in order to provide a consistent view of the data. The performance of such an application can be adjusted by caching the crawled data. A tool that crawls Linked Data is the LDSpider [IUBH10]. Typically, Linked Data Search Engines are used to crawl the Web (see below).
- **The On-The-Fly Dereferencing Pattern.** This pattern describes a method that dereferences only those URIs that are currently in use by the application. Heath et al. [HB11] state that this pattern holds the advantage that no unused data is processed, but that the live dereferencing of many URIs in the background decreases the runtimes of such applications. Linked Data applications using this pattern are typically Linked Data Browsers (see below). Approaches for on-the-fly dereferencing are presented in [HBF09, HT12].
- **The Query Federation Pattern.** Applications using this pattern send complex SPARQL queries to SPARQL endpoints determined in advance. General challenges and solutions in distributed query processing are discussed in [Kos00]. An approach applying SPARQL query federation is presented in [LWB08].

These architectural patterns for retrieving heterogeneous Linked Data are also discussed in detail in [HL10]. Freitas et al. [FCOO12] discuss different approaches in constructing queries for an intuitive search in Linked Data resources.

In general, [BHBL09] distinguish between three categories of Linked Data applications: browsers, search engines and indexes, and domain-specific applications.

- Linked Data Browsers allow users to navigate through Linked Data resources by following RDF links between them. Examples for Linked Data Browsers are the Disco hyperdata browser[13], the Tabulator browser[14] [BLCC$^+$06] and LinkSailor[15].
- Linked Data Search Engines and Indexes can be distinguished in human-oriented and application-oriented search engines. Both usually crawl Linked Data from the

[13] http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/

[14] http://www.w3.org/2005/ajar/tab

[15] http://linksailor.com/

web. While human-oriented engines present the retrieved data sources as results in a more or less user-friendly way, application-oriented approaches focus on providing results via an API or on creating an index over Linked Data sets, both for reuse in third-party applications. Examples of human-oriented Linked Data Search Engines are Sig.ma[16] [TCC+10], VisiNav[17] [Har12], SWSE [18] [HHD+07] and Falcons[19] [CQ09]. Well-known application-oriented search engines are Sindice[20] [TDO07], Swoogle[21], and Watson[22].

- Domain-specific Applications focus on available Linked Data resources for a specific regard in order to offer users mashed up or visualized information based on Linked Data. Examples of domain-specific Linked Data applications are U.S. Global Foreign Aid mashup[23], DBpedia Mobile[24], NCBO Resource Index[25], Diseasome Map[26], and paggr[27].

Detailed overviews of Linked Data applications can be found in [BHBL09, HB11]. There are also lists comprising Linked Data applications based on governmental data of the US[28] and the UK[29]. Hausenblas [Hau09] discussed ways to exploit Linked Data for creating applications. As seen in the above classification and in the examples, most of the existing Linked Data applications focus on the searching, browsing and visualising of data. Only a few approaches like [vHEM12, Pau12, NKIV11, KOH12] go beyond this state and allow users to interact with Linked Data. Moreover, most Linked Data applications share the commonality that they are built on the technical standards and on available Linked Data sets, which is reasonable, because data is one of the most important fundamentals for building applications. Approaches that are inspired by real-life use cases and the information needs of humans instead of available data still lack their volume, quality, and the persuasiveness of a benefit to users. There is no ‘Killer App’ for Linked Data, a fact that is widely discussed in the community[30]. Some argue that Linked Data should not be seen as an application, but rather as enabling infrastructure[31]. The concern over why Semantic Web technologies have yet not arrived among web developers has also been the subject of controversy and debate at the 10th International Semantic Web Conference[32]

[16] http://sig.ma/
[17] http://sw.deri.org/2009/01/visinav/
[18] http://www.swse.org/
[19] http://iws.seu.edu.cn/services/falcons/documentsearch/
[20] http://sindice.com/
[21] http://swoogle.umbc.edu/
[22] http://kmi-web05.open.ac.uk/Overview.html
[23] http://data-gov.tw.rpi.edu/demo/USForeignAid/demo-1554.html
[24] http://wiki.dbpedia.org/DBpediaMobile
[25] http://bioportal.bioontology.org/resources
[26] http://diseasome.eu/map.html
[27] http://paggr.com/
[28] http://www.data.gov/communities/node/116/apps
[29] http://data.gov.uk/apps
[30] e.g. see http://dataliberate.com/2012/03/the-pointless-search-for-the-killer-app/ or [KPGC11]
[31] http://www.niso.org/news/events/2011/nisowebinars/authoritydata/questions/
[32] http://iswc2011.semanticweb.org/

during the Semantic Web Deathmatch panel discussion[33]. In Section 3.2, we present a prototype application for the consumption of Linked Data, which has been inspired by a real-world use case.

The publication and consumption of Linked Data generates new and increases existing challenges and open issues, which accompany a decentralized, open and self-organizing web. Issues of privacy and licensing, trust, relevance and quality of Linked Data as well of maintaining the links have gained in importance in recent years. As these topics are beyond the scope of this thesis, we will not discuss them further at this point. Relevant work regarding privacy can be found in, e.g. [HP09, WSRH10, SP11]. Issues of trust, relevance and the quality of Linked Data are discussed in, e.g. [BFF08, GGSL12, HZ09, MMB12]. The problem of maintaining links is addressed in [VBGK09, VHC10, PH10].

2.2 Statistical Linked Data

In this section, we define the term Statistical Linked Data. An overview of the typical structure, semantics and modelling issues is given. This section builds upon the foundations of LOD as given in Section 2.1. The focus of this section is on the currently available Statistical Linked Data sets and their typical characteristics, attributes and patterns. Since Statistical Linked Data may be used for scientific research in the social sciences, it is seen as a part of Linked Open Social Science Data as introduced in Section 1.1.

Statistical Data

Statistical data has a long tradition, starting from the late 1800s as a means for kings to keep track of the economic development of their country. In the early 1900s, the very first automatic data storage systems, e.g. Hollerith cards from IBM[34], were developed to enable large scale statistical operations, like the US Census in 1890. Since then, statistical data may have lost its position as a front runner in the technological race towards better data management, but it has never lagged too far behind either.

Statistical data is periodically collected by administrative sources [fECoD02] and attempts to describe the state of a nation in numbers, typically by collecting demographic and economic data. Commonly known examples include population number and unemployment ratios, but also soft measurements like general well-being. When the data is collected, it is usually stored in table-like data structures, like Excel, or the diverse formats of current statistical programs, like SPSS[35], STATA[36] and R[37]. For larger-scale processing, these are

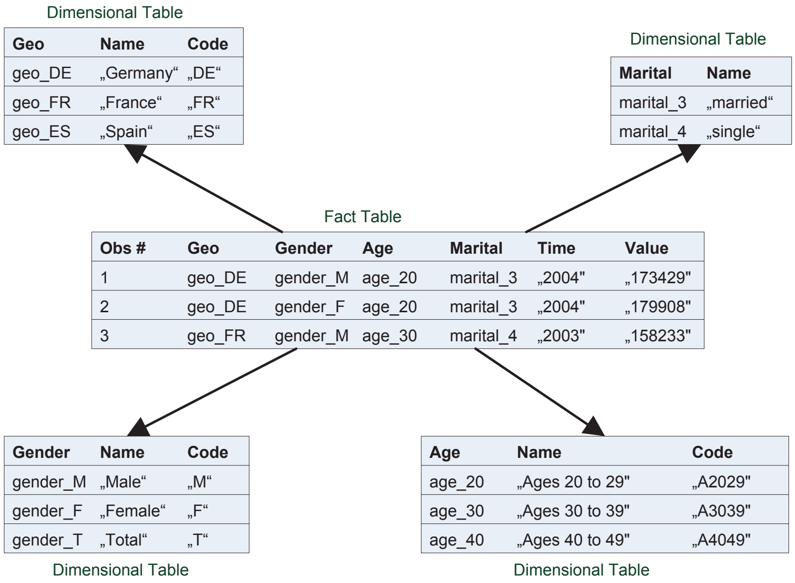
[33] see http://videolectures.net/iswc2011_panel/ and http://semanticweb.com/semantic-web-death-match-at-iswc_b24249 for a summarisation

[34] http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/tabulator/

[35] http://www-01.ibm.com/software/uk/analytics/spss/

[36] http://www.stata.com/

[37] http://www.r-project.org/



Geo	Name	Code
geo_DE	„Germany“	„DE“
geo_FR	„France“	„FR“
geo_ES	„Spain“	„ES“

Marital	Name
marital_3	„married“
marital_4	„single“

Obs #	Geo	Gender	Age	Marital	Time	Value
1	geo_DE	gender_M	age_20	marital_3	„2004"	„173429"
2	geo_DE	gender_F	age_20	marital_3	„2004"	„179908"
3	geo_FR	gender_M	age_30	marital_4	„2003"	„158233"

Gender	Name	Code
gender_M	„Male“	„M“
gender_F	„Female“	„F“
gender_T	„Total“	„T“

Age	Name	Code
age_20	„Ages 20 to 29"	„A2029"
age_30	„Ages 30 to 39"	„A3039"
age_40	„Ages 40 to 49"	„A4049"

Figure 2.2: Statistical Data organized in a Star Schema.

then transferred to relational databases or data warehouses. The structure of statistical data can be compared with multidimensional models in data warehouses [Inm05], where a fact table determines the centre of the model in a star join data structure.

> ‘A fact table is a structure that contains many occurrences of data. Surrounding the fact table are dimensions, which describe one important aspect of the fact table’ [Inm05]

This structure is also reflected in the SDMX information model [SDM09], a multidimensional standard model for describing statistical data. In terms of statistical data, an occurrence of data refers to a statistical observation. In general, statistical data consists of multiple observations [SDM09], which have been made at some points. Observations can be organized among specific dimensions (e.g. temporal or geographical dimensions), which describe the logical space on which the observations have been applied (e.g. time and geographical area). This information is often coded [SDM09], i.e. its values are taken from classifications or code lists. For example, the dimension "reporting country" may refer to values of a classification consisting of country names. The values of the fact table can be organized along with their surrounding dimensions as data cubes [CRT14]. Figure 2.2 depicts an example of statistical data organized in a star schema.

Code lists are used for encoding information on geographical concepts, gender, age groups and others [SDM09]. Especially for geographical codes, there exist several code lists, which are traditionally used for statistical data. These are widely reused for Statistical Linked Data. The ISO norms[38] 3166-1 for codes of countries and dependent territories

[38] http://www.iso.org/iso/home/standards/country_codes.htm

and 3166-2 for codes of subdivisions (e.g. states or provinces) of countries are well known and used internationally. The Nomenclature of Territorial Units for Statistics (NUTS)[39] denotes a common standard for referencing regional areas in the member states of the EU, where the three levels stand for different levels of subdivisions of the countries. For Germany, for example, NUTS level 1 denotes the federal states, level 2 government regions and level 3 the smallest subdivision, the districts. But various agencies maintain their own code lists, which increases heterogeneity.

Statistical Linked Data

Due to governmental pressure and other effects, the number of statistical data sets available as LOD has recently seen a considerable increase. This is a welcome step towards governmental transparency, as professionals from many domains rely on the analysis of such raw data as opposed to the often graphic-based representations that are preferred by laypersons.

In this thesis, we define Statistical Linked Data as statistical data that is technically published according to the Linked Data principles. Statistical Linked Data includes both, aggregated and micro data, and can be generalized to survey data. However, these are so far rarely found as Linked Data. In this thesis, the term does not include statistical data from other domains, such as experimental data from clinical trials or laboratory experiments of any kind. Represented as Linked Data, a statistical data set consists of several instances of data entries, each of which determines a particular data value, e.g. `548215`. The data values are supplemented by additional objects which provide further information, e.g. in which country or at which time the data value has been collected. This sets the data values in a context. Such objects are referenced in the data value instances by object properties. However, the objects themselves are classes or individuals of other external or separate data sets (e.g. classifications or code lists). Figure 2.3 depicts an example of Statistical Linked Data.

A vocabulary designed for representing Statistical Linked Data is the RDF Data Cube vocabulary [CRT14]. It is based on the SDMX information model [SDM02] and is capable of modelling observations, dimensions and measures for multi-dimensional data sets. However, there are other vocabularies for other types of research data, e.g. the DDI-RDF Discovery vocabulary (see Section 4.3 and [BCWZ12, BZWG13]), which aims to represent micro data (i.e. person-level data) as Linked Data. An overview on other vocabularies for representing statistical data can be found in Section 4.3.1.

Statistical Linked Data Sets Used in this Thesis

Table 2.1 depicts the Statistical Linked Data that is used in this thesis. Additional overviews regarding available Statistical Linked Data sets can be found at the Data

[39] http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction

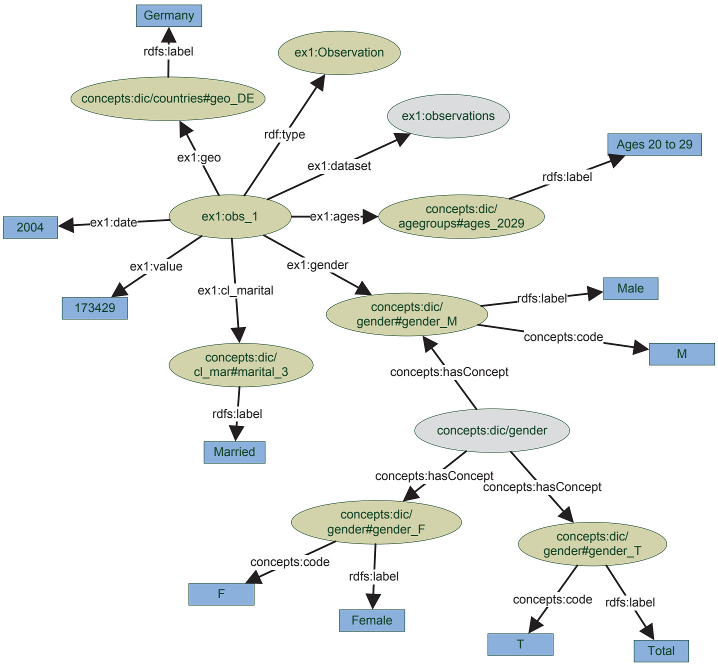


Figure 2.3: Example of Statistical Linked Data.

Hub[40], a data repository that currently contains 9,864 data sets including the 570 data sets of the LOD cloud diagram [SBJC14], and the wiki of Planet Data[41], which collects data sets published in the RDF Data Cube vocabulary [CRT14]. It lists 21 data sets.

In this thesis, all implementations and evaluations that use Statistical Linked Data have been carried out with these data sets. In Section 3.2, two additional data sets have been used. Their RDF generation is presented in Section 3.2.4.

German General Social Survey – ALLBUS The German General Social Survey ALLBUS[42], which collects up-to-date data on attitudes, behaviour and social structure in Germany, is archived at GESIS – Leibniz Institute for the Social Sciences[43]. Due to data privacy restrictions, we use a special edited version of a subset of ALLBUS/GGSS 1980-2008 (Cumulated German General Social Survey 1980 - 2008) [ALL10], which includes only few variables that are relevant for our use case ('Current Economic Situation in Germany' and 'Resp. own Current Financial Situation'). Additionally, only the data of participants from North Rhine-Westphalia has been included into the subset in order to make it comparable to the election statistics from North Rhine-Westphalia on a

[40] http://thedatahub.io/
[41] http://wiki.planet-data.eu/web/Datasets
[42] http://www.gesis.org/en/allbus/allbus-home/
[43] http://www.gesis.org/

geographical level. Because of omitting a lot of relevant information and variables for the subset, it has been explicitly created for technical feasibility experiments only.

Data Provider	Description	URL
data.gov	US governmental data from diverse agencies like energy, environment, veterans affairs, housing and urban development, commerce, healthcare, etc.	http://data-gov.tw.rpi.edu/wiki/Data.gov_Catalog
data.gov.uk	Data about education, transport, environment, etc. from the British government.	http://data.gov.uk/linked-data
Eurostat	Data publicly available from the statistical office of the EU. Covers various topics ranging from population, economics and industry to education.	http://estatwrap.ontologycentral.com/
ISTAT	Immigration statistics from Italy.	http://www.linkedopendata.it/datasets/istat-immigration
OECD	Organisation for Economic Co-operation and Development published as Linked Data	http://oecd.270a.info/
World Bank Climate Change	Data about climate change, including the response of the global climate system to increasing greenhouse gas concentrations.	http://worldbank.270a.info/dataset/world-bank-climates.html
Global Hunger Index	The Global Hunger Index (GHI) offers a multidimensional overview of global hunger recording the state of global, regional and national hunger.	http://datahub.io/dataset/global-hunger-index-2011
EnAKTing Energy	Data extracted from the statistics for road transport consumption compiled by the UK Department for Business, Enterprise and Regulatory Reform.	http://energy.psi.enakting.org/

Table 2.1: Statistical Linked Data used in this thesis.

Election Statistics of North Rhine-Westphalia The election statistics from the German federal state of North Rhine-Westphalia are provided by IT.NRW[44], the statistical office and IT service provider of the federal state. They are published as tables on HTML pages and are accessible as CSV via a web service. The statistics contain election votes and results for the elections of the German parliament, the parliament of North Rhine-Westphalia as well as for elections of the European parliament. Both, votes and results, can be retrieved on different administrative levels, e.g. from the federal state itself, administrative districts down to single electoral districts.

2.3 Schema Matching

The representation of data with Semantic Web technologies and standards with regard to Linked Data alone does not allow its seamless consumption for research. Considering the use case of 'Analysing Research Data' introduced in Section 1.2, more data processing has to be done, i.e. merging and integrating multiple data sets. Schema matching has a long tradition in different application domains of databases, such as database integration, semantic query processing or data warehousing [RB01]. Despite this wide range of application areas, [MBR01] argue in favor of schema matching independently as a generic problem because of the similar approaches for solving the problems of data matching and integration. Applied on Linked Data, schema matching can be used for merging and integrating multiple heterogeneous data sets.

In the following sections, we describe the general problem of matching and provide an overview of existing techniques and matching systems. Finally, we describe how matching systems can be evaluated.

2.3.1 The Matching Problem

Rahm [Rah11] summarizes the process of schema matching as follows:

> 'Schema matching aims at identifying semantic correspondences between metadata structures or models, such as database schemas, XML message formats, and ontologies' [Rah11]

According to [MBR01], the problem of schema matching starts with two given schemata $S1$ and $S2$, each consisting of related elements, e.g. tables, columns, classes or attributes. A *mapping* between two schemata subsumes several *mapping elements* describing a correspondence between a particular element of schema $S1$ to a particular element of schema $S2$. The process of detecting a mapping element is independent of data models and can be supported by auxiliary information such as input mappings. The matching process can include validation of mappings by users.

[44] http://www.it.nrw.de/

Euzenat et al. [ES07] define the matching problem in the context of ontology matching very similarly to [MBR01], but aim to formalise the approach for keeping it as general as possible. The matching process of this definition is depicted in Figure 2.4. According to [ES07], an alignment A' for two ontologies o and o' is detected by a matching process. Three optional input parameters are allowed:

- an input alignment A, which has to be completed or can serve as reference,
- matching parameters p, e.g. weight or thresholds and
- external resources r for supporting matching functions, e.g. knowledge bases like thesauri.

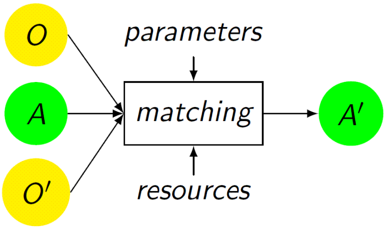


Figure 2.4: The matching process [ES07].

The matching process is defined as the following function [ES07]:

$$A' = f(o, o', A, p, r)$$

An alignment A consists of multiple correspondences, which determine a relation between elements of both ontologies. A correspondence is defined in [ES07] as the 5-tuple

$$\langle id, e, e', r, n \rangle$$

where id is the unique identifier of the correspondence for the particular relation r between the ontology entities e and e'. n describes a particular confidence value of the relation r.

Schema Matching vs. Ontology Matching

This thesis focuses on schema matching, although the principles and methods of ontology matching are also of relevance. Shvaiko et al. [SE05] distinguish between both concepts as follows:

> ‘Database schemas often do not provide explicit semantics for their data. Semantics is usually specified explicitly at design-time, and frequently is not becoming a part of a database specification, therefore it is not available. Ontologies are logical systems that themselves obey some formal semantics, e.g., we can interpret ontology definitions as a set of logical axioms.’ [SE05]

Furthermore, they state that the focus of schema matching is usually to guess the meaning of schema elements, while ontology matching focuses on exposing the encoded information. But the similarities of both approaches (i.e. ‘both provide a vocabulary of terms and somewhat constrain the meaning of terms used in the vocabulary’ [SE11]) and the general applicability of schema matching onto various types of data models lead to mutual benefits of solutions from both problems [SE05, SE11].

Both concepts are stated as being relevant for Linked Data [BMR11, SE11] because their results can be used as links between resources. In particular, instance matching with respect to entity resolution has been addressed only recently. This thesis focuses on schema matching. Considering the use case of ‘Analysing Research Data’, it becomes clear that research data such as survey data or statistical data has its origin in databases and data warehouses instead of ontologies. Usually, only small semantics accompany the data; even if they are represented in RDF or OWL, they are mostly created and exposed out of databases.

2.3.2 Classification of Matching Techniques

Madhavan et al. [MBR01] defined an initial taxonomy of schema matching techniques, which comprised current approaches. This taxonomy has been revised and supplemented in [RB01]. While this categorization has been very database-centric, [SE05] proposed a classification considering techniques from both, schema and ontology matching approaches. The different approaches have been organized according to three criteria:

> ‘(i) general properties of matching techniques, (ii) interpretation of input information, and (iii) kind of input information’ [SE05]

This classification has been updated in [ES07] (e.g. by the sporadically consideration of instances as input) and is depicted in Figure 2.5.

The difference from the classification of [RB01] is that [ES07] characterize the matching approaches on a different level of granularity and separate the general matching technique from the input data. In general, [ES07] distinguish between element-level and structure-level matching techniques, both of which are subdivided into syntactic, external and semantic techniques. While syntactic techniques examine the input according to its syntactical structure, external techniques use additional (external) resources as auxiliary information in order to analyse the input data. Such information can include, e.g. domain knowledge provided by thesauri or human interaction. Semantic techniques apply model-based semantics like reasoners in order to interpret the input.

Element-level Techniques

Techniques of this category compute correspondences between single entities (e.g. schema elements, ontology classes or their instances) without considering their relation to other

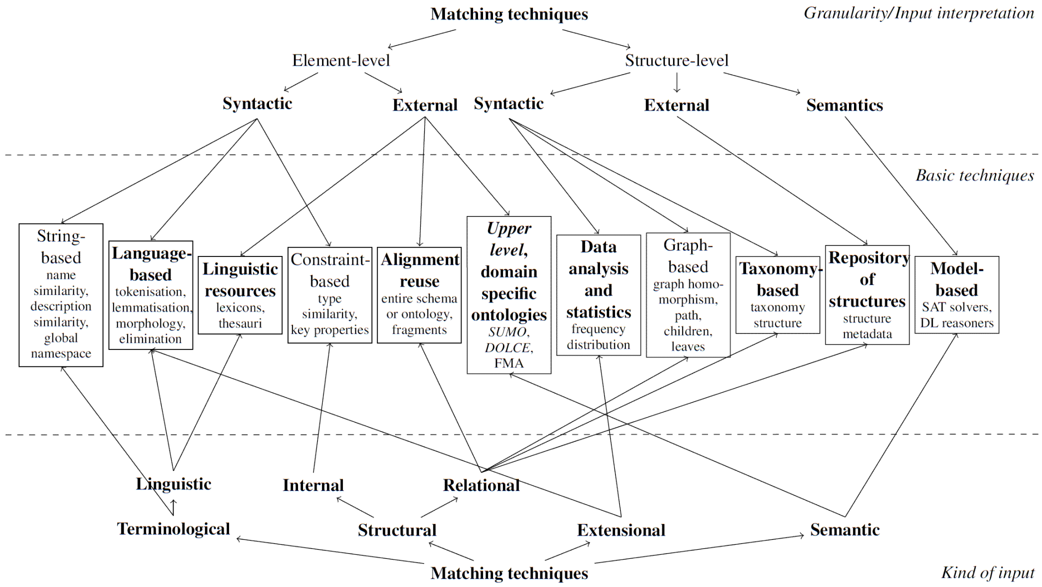


Figure 2.5: Classification of matching approaches [ES07].

entities within the input data. In comparison to the classification of [RB01], element-level techniques comprise the instance-based and parts of the schema-based matching techniques. Typical approaches and techniques are described in brief.

String-based techniques examine the strings of entities as a sequence of characters. Correspondences are computed via distance functions. Language-based techniques consider entities as words of a particular language. Intrinsic techniques like tokenization, lemmatisation or term extraction are applied for identifying correspondences. Using constraint-based techniques, attributes of the entities are analysed regarding particular constraints, e.g. cardinality, multiplicity or specific data types. Using linguistic resources like thesauri or lexicons allows the detection of correspondences by identifying linguistic relations, e.g. synonyms or hyponyms. A different approach for using external information is the reuse of existing alignments. This is a reasonable approach when matching input data of a domain, for which alignments of other, but similar input data are already available. A variant of this approach is to exploit the semantics of external upper level and domain-specific formal ontologies for detecting correspondences.

Structure-level Techniques

These techniques compute correspondences between single entities of the input data, but in contrast to element-level techniques, they analyse the relation between entities and the structure of entities in which they appear. The structure-level techniques comprise parts of the schema-based matching techniques of [RB01]. Common techniques are described briefly.

By using graph-based techniques, the input data is analysed as the named graph structures. If two nodes from two ontologies (or schemata) are similar, it is assumed that their neighbours are also somehow similar. Thereby, taxonomy-based techniques consider only specialization relations of a graph, i.e. hierarchical is-a relations. A repository of structures can be used for identifying correspondences between structures of the input data by comparing them with existing alignments between other data structures. This approach is similar to the reuse of existing alignments with the difference that not entities, but complete structures or fragments of structures of entities are considered. Model-based techniques analyse the input semantically by, e.g. describing logical reasoning techniques. Data analysis and statistical techniques compute representative samples of the input or of fragments of it in order to detect regularities or discrepancies.

2.3.3 Overview on Matching Systems

The following section provides an overview of current matching systems. As we introduce two novel instance-based schema matching approaches in Sections 5.3 and 5.4, the focus in this section is on instance-based and mixed approaches, i.e. approaches that consider the schema and instance levels. This section does not aim to provide a complete overview of developed systems. Further overviews of matching systems can be found in [RB01, ES07, EFvH$^+$11, BMR11, Rah11]

Similar schema (or ontology) elements can be derived out of the similarity of their instance values [BMR11] and can thereby deliver valuable results to the matching process. Instance-based approaches are located as element-level techniques in the classification of [ES07]. A survey with different matching functions considering instances conducted by [IMSW07] showed that instance-based approaches deliver excellent results. Our approach presented in Section 5.3 is an instance-based approach, which considers instances with respect to their patterns and features in order to match schema elements. It can be relocated as an instance-based and constraint-based technique according to [BMR11], and as an element-level technique in [ES07] applying constraint-based functionalities.

Different instance-based approaches for schema matching can be identified in current systems. COMA [DR02] was originally a schema-based approach that has been extended by several functionalities (and thereby entitled as COMA++) like importing schemata, adding new matchers and performing complex matching algorithms, e.g. fragment-based matching [ADMR05]. Engmann et al. [EM07] have enhanced the matching process by considering instances of the input data. This has been done by adding constraint-based and content-based matching. A similar approach for considering constraints in instances is described in [ZSC09], where the use of regular expressions and catchwords is considered for instance-based schema matching. NOM [ES04b] and its variation QOM [ES04a] detect alignments by considering instances, schema and structure information. NOM applies rules into the matching process, which exploit knowledge from the input ontologies like information about super classes. QOM aims to improve the efficiency of

NOM by restricting some time-consuming rules. Falcon-AO [HQ08] combines a linguistic and a structure matching. Doan et al. [DMD$^+$03] describes the GLUE system, which uses similarity measures, exploitation of domain constraints and heuristic knowledge for instance-based schema matching. Some schema matching systems use machine learning and rule based approaches for detecting features inside schema elements and instance values, e.g. Automatch [BM02] or the system presented in [JFNP12]. Features of instance values are also computed in [ZC09]. Lexical similarities of instance values are computed in CODI [HSNM11] in order to create object properties, which can be seen as a similar approach to the use of features. In PARIS [SAS11], degrees of matchings are measured between instances and schema elements based on probability estimates.

Additional matching systems interpreting instance-level information are LSD [DDH01], DUMAS [BN05], FCA-merge [SM01], SEMINT [LC00], Clio [HPV$^+$02], and RiMOM [LTLL09]. Approaches that consider primarily schema-level information are H-Match [CFM06], Cupid [MBR01], Similarity Flooding [MGMR02], OntoMerge [DMQ05], and S-Match [GSY04].

Although instances have already been considered extensively by schema and ontology matching approaches, they have gained in importance due to the popularity of Linked Data [BMR11, SE11] recently. In the context of Linked Data such approaches of entity resolution regarding instance matching are mostly applied with a focus on detecting similar instance values (see Subsection 'Including Links to Other Resources' of Section 2.1.1 for more details). This is grounded in the assumption that identical or similar instance values belong to the same or a similar schema element.

2.3.4 Evaluation Methods of Matching Approaches

Several works focus on the evaluation of schema and ontology matching approaches [BBDV11, DMR03, ES07]. In order to assess the performance and quality of alignments computed by a matching system and compare these results with other systems, evaluations and evaluation results give developers insights into the strengths and weaknesses of their systems. Moreover, users can choose particular matchers based on which ones are most appropriate for their requirements. This section presents which aspects should be considered for conducting a methodological correct and comparable evaluation. We summarize the principles and different types of evaluations and provide an overview of the data sets to be used for evaluation. Additionally, we describe common evaluation measures and benchmarks.

Evaluation Principles

[ES07] state five rules, which should underlie each evaluation process.

- **Systematic Procedure.** The evaluation process has to be reproducible at different moments and on different systems.

- **Continuity.** An evaluation process is encouraged to be an ongoing effort in order to detect and observe progresses.
- **Quality and Equity.** The setting of the evaluation (e.g. the task, the data sets, the measures) should be defined precisely before the conduction. The evaluation process must not be distorted or changed by the approaches that are to be examined.
- **Dissemination.** The evaluation and especially all of its data sets and results, etc. have to be published and made available.
- **Intelligibility.** To ensure that the evaluation and its results can be understood by third parties, the intermediate (e.g. the alignments themselves) and final results should be published and explained.

An evaluation is typically conducted via three methodological steps [ES07]: planning, processing and analysing. The planning stage includes the definition of the task and the choice of the used data sets, systems, measures and the desired output. While the evaluation is conducted during the processing stage, its results are examined in the analysing stage according to determined measures.

Types of Evaluation

Euzenat et al. [ES07] consider three types of evaluation that can be distinguished by their purpose, i.e. what has to be evaluated. The three types are competence benchmarks, comparative evaluations and application-specific evaluations.

- **Competence Benchmark.** Benchmarks are defined in order to evaluate the performance of systems according to a pre-defined task. Such a task often isolates particular characteristics, e.g. inside the used data or the applied matching technique. Benchmarks help improve individual systems. Typically, they are conducted multiple times, e.g. annually in order to observe improvements over time. A few popular benchmarks are described in detail below.
- **Comparative Evaluation.** This type of evaluation aims to compare several matching systems according to a defined task. A comparative evaluation aims to find the best system for this particular task. It is important that the task is defined precisely and that the data set for the task is published shortly before the evaluation. This prevents the matching systems from being tuned and adjusted specifically to the characteristics of the data set.
- **Application-specific Evaluation.** This kind of evaluation focuses on the input data, which is delivered by a particular application or domain. The evaluation aims to test the performance of the systems on these specific input data and its characteristics.

Data Sets

The data sets on which an evaluation is conducted play an important role because they can influence the matching process and, therefore, the achieved results. Euzenat et al. [ES07] state six factors, according to which a data set can be chosen or designed.

- **Input Ontologies.** As ontologies differ in their structure, extent and according to the described knowledge, they can influence the matching process and the performance of the matching approaches. Typically, there are two ontologies on which a matching process is conducted. But there are scenarios in which multiple ontologies are considered as inputs.
- **Input Alignment.** Pre-defined alignments (by a user or a previous matching process) can support the matching quality as they can serve as a reference for the newly detected alignments.
- **Parameters.** Algorithms and functions of a matching approach can be influenced by setting parameters, e.g. weights or thresholds.
- **Resources.** Additional resources like thesauri or lexicons can support matching approaches by delivering background knowledge, which can be used for specific linguistic or terminological functions. Moreover, training data, which is especially necessary for machine learning approaches, is seen as an additional resource. This factor also includes user interaction in terms of reviewing proposed alignments.
- **Output Alignment.** Several characteristics of the output alignment influence how the matching process has to be conducted. The multiplicity of the output alignments defines the cardinality of the detected correspondences, e.g. 1:1, 1:n or n:m correspondences. Furthermore, the relations within the correspondences, i.e. how the two entities of a correspondence are associated with each other, have to be determined. Relations can be of an equivalent kind, a subsumption or describe an incompatibility. Finally, a value for each correspondence can be computed, which indicates the confidence of the correctness of the detected correspondence.
- **Matching Process.** This factor influences the process as a whole by constraints, e.g. time constraints or specific types of entities, e.g. only instances or schema elements which should be considered for a correspondence.

Evaluation Measures

The resulting alignments produced by a matching system during an evaluation can be assessed and analysed using specific evaluation measures. We focus on the most common measures of precision, recall and F-measure, because they are the most frequently used measures. Moreover, [ES07, BBDV11] also describe measures of fallout, overall and strength-based similarity as well as performance and user-related measures.

The measures precision, recall and F-measure have been originally defined for information retrieval purposes [vR79]. They have been applied to ontology and schema matching by [DMR03].

> 'Precision measures the ratio of correctly found correspondences over the total number of returned correspondences' [ES07].

$$P(A,R) = \frac{|R \cap A|}{|A|}$$

where A determines the resulting alignments, R describes the reference alignments and $R \cap A$ the correct and found alignments.

> 'Recall measures the ratio of correctly found correspondences over the total number of expected correspondences' [ES07].

$$R(A,R) = \frac{|R \cap A|}{|R|}$$

A combination of precision and recall is the F-measure. It is commonly used in order to achieve a better comparability of results because the sole investigation of precision and recall is often insufficient [ES07]. By computing the F-measure, the importance of precision and recall can be determined by the variable α. If it is intended to give precision and recall the same relevance, then the value of α is 0.5. The F-measure M is then computed as follows.

$$M_{0.5}(A,R) = \frac{2 \cdot P(A,R) \cdot R(A,R)}{P(A,R) + R(A,R)}$$

Benchmarks

The Ontology Alignment Evaluation Initiative[45] [OAE] is the best-known platform for the evaluation of ontology matching. It has been held annually since 2004 and comprises several tracks covering different evaluation types like the OAEI benchmark as well as various comparative and application-specific evaluations. The benchmark track of the OAEI is based on ontologies describing bibliographic resources. There is one reference ontology, which is then transformed into 50 variant ontologies. Different kinds of transformations are computed, like modifying element names, suppressing or restricting comments, instances or properties, or expanding or flattening hierarchies or classes. During the benchmark test, the reference ontology is to be matched with each of the variant ontologies.

The STBenchmark [ATV08] has been created for evaluating mappings between schemata. It provides various pre-defined mapping scenarios, which can be applied on the input

[45] http://oaei.ontologymatching.org/

schema in order to generate a modified target schema. But it is also possible to generate new scenarios and corresponding variations. STBenchmark provides only a simple usability model as a measure. No reference alignments are created or can be used. These lacks prevent a complex and comparative evaluation of systems.

The Islab Instance Matching Benchmark (IIMB)[46] [FLMV08] has been created for measuring and evaluating instance matching systems and their results. The creation of the benchmark and its evaluation are similar to those of the OAEI benchmark. There is a reference ontology, of which variations are generated by modifying only the instance values.

[46] http://islab.dico.unimi.it/content/iimb2009/

3 Linked Data Consumption in the Social Sciences

LOD can fulfil the vision of the Semantic Web by bringing more meaning to the World Wide Web and its data [BL06]. Due to the impact of using standard technologies like those of the Semantic Web and the popularity of Linked Data [BHBL09], an application of these technologies is often recommended for various disciplines, among them scientific research (for a detailed discussion, see Chapter 1). Meanwhile, Semantic Web technologies and tools are mature enough to allow broad applicability across various disciplines. However, the consumption of Linked Data may be problematic because it depends on the acceptance of these technologies by application developers and end users. This could evolve to a core problem for the Semantic Web community.

In order to investigate methods for the data matching of LOD in the social sciences, we have studied the actual consumption of Linked Data in this domain. This study serves as a foundation to enhance and develop methods for matching Linked Open Social Science Data. By conducting qualitative interviews with experts from the social sciences, we examine how Linked Data is currently consumed in social scientific research. Although the interviews show that there is currently no consumption of Linked Data, experts see a large potential for beneficial application. Based on these observations, we have built a prototype that demonstrates a possible consumption of Linked Data in the social sciences by integrating two Statistical Linked Data sets. During the development, we have elaborated the technical requirements for the use of Linked Data in such an application. Finally, we draw conclusions from the interviews and the prototype development that will be used in the following chapters. In a list of points of interests, we can address the problems named by the experts explicitly using Linked Data technologies that may resolve or improve some of the problems.

In Section 3.1, the conduction and results of the qualitative interviews are presented. The development of a prototype for Linked Data consumption in the social sciences is described in Section 3.2. Finally, in Section 3.3, we present the summarized list of points of interest for a beneficial consumption of Linked Data in the social sciences.

3.1 Qualitative Interviews

We have conducted qualitative interviews with domain experts from the social sciences in order to investigate whether and how Linked Data is consumed and whether Linked

Data technologies can be applied beneficially for social scientific research. The aim of the interviews is not only to find out whether Linked Data is consumed, but also to examine how and for which particular tasks Linked Data technologies can be applied. Moreover, it is investigated where enhancements by using these technologies can be expected, e.g. concrete problems where such technologies may ease working tasks. According to our initially introduced use case of 'Analysing Research Data' (see Section 1.2), the interviews focus on typical working tasks occurring during this use case.

In this section, we present the design, preparation and analysis of the interviews. In Section 3.1.1, we describe the applied methodology on which the interviews have been designed, conducted and analysed. The preparation and conduction of the interviews are presented in Section 3.1.2. Finally, in Section 3.1.3, we analyse and discuss the results of the interviews.

3.1.1 Methodology

This section presents the methodology applied for the conduction of the qualitative interviews in this thesis. We provide a short overview of qualitative research in order to describe and argue our methodological choice and design. Detailed information about the foundations, concepts, and methodologies of qualitative research can be found in [SHE05, KKV94, May02, FvKK$^+$95].

Qualitative research aims to access the social reality through considerably different theoretical and methodological approaches [FvKK$^+$95]. V. Kardorff [vK90] defines the starting point of qualitative research as an experiment that determines a meaningful access to a social reality, which is interactively created and represented by language and non-language symbols. It is aimed to reconstruct a detailed and complete representation of the reality to be analysed. Thus, experiences and perceptions of the examined objects are considered for research. In contrast to quantitative research, where complex social facts are explained objectively by analysing a predefined hypothesis on large samples, qualitative research focuses on the understanding of complex social facts through the reconstruction and investigation of subjective perceptions and opinions [Kru11]. In the social sciences, qualitative research is typically carried out when the examined topic or object seems to be too complex, differentiated, unclear or contradictory to be investigated and analysed by quantitative, i.e. metrical, methods [FvKK$^+$95].

There are different methods for collecting data in qualitative research. Among participating observations and group discussions, conducting interviews is a typical method [May02]. Mayring [May02] categorizes qualitative interviews based on their degree of openness, their degree of structure and standardization, and the kind of analysis. The openness of an interview describes the freedom of the interviewed person, i.e. is he allowed to answer freely or asked to choose from among predefined answers. The structure and standardization of an interview describes the freedom of the interviewer, i.e. whether he has a guideline or questionnaire for the interview or not and how strictly the guideline is followed. The kind of analysis indicates how the results of the interviews are investigated, e.g. by qualitative interpretable methods or by quantitative metrical methods. These criteria can be relocated in other classifications of qualitative interviews [Kru11, FvKK$^+$95, Hel05].

To examine how Linked Data can be used in social science research, we have conducted open and unstructured expert interviews. We chose researchers from the social sciences as experts for the interviews. This research design offers the best insight into the work, problems and challenges of a specific field, where the researchers act as experts [Lit08, Kru11]. In order to gather as many subjective opinions of the participants as possible, we have chosen to conduct open interviews, where the interviewed person is allowed to give free answers. The interview is planned as an unstructured interview in order to focus on the interviewed person and his or her opinions on the topic. Although our interviews have been planned to be open and unstructured, we have chosen to define a guideline for the interviews as [Kru11] classifies expert interviews as guideline-supported interviews, where the strictness of following the guideline can vary massively. Moreover, [Lit08] recommends using a guideline in order to act as a competent interview partner to the expert. The creation of the guideline according to [Kru11] is discussed in Section 3.1.2.

In all fields of qualitative research, the sample design, i.e. the choice of persons that are observed, is of high relevance. While in quantitative research, a sample that structurally represents a whole group is taken randomly, the sample in qualitative research is taken manually and represents the heterogeneity of a group in a case-dependent manner [Kru11]. Typically, a qualitative sample is rather small and deliberately designed with a maximum of structural heterogeneity according to the whole group.

Before the interviews can be analysed, the collected material has to be prepared for scientific analysis. The conducted interviews in this thesis have been transcribed literally in order to preserve all spoken words during the interviews. Because we have conducted open and unstructured interviews with experts, some relevant statements might be mentioned in side comments. Besides, a literal transcription allows the extraction of strong and relevant statements for an analysis.

The qualitative interviews are analysed and interpreted by using reconstructive text analysis [Kru11], which aims to expose central intentions and perspectives inside the text. [May02, Kru11] list various techniques for analysing such data depending on the available material, the conducted methodologies and the intended research gain. For analysing the interviews in this thesis, we are conducting a qualitative content analysis that is performed closely to the transcribed text. The transcribed interviews are analysed in a sequence analysis, i.e. the text is divided into segments, which can vary from paragraphs to single sentences. Each segment is then analysed separately from the others. Text passages are classified and subsumed. Afterwards, they are compared with predefined analysis categories and used to modify these categories. These modified analysis categories allow an abstraction to central intentions and perspectives of the interviewed person. The beginning is analysed with special attention as it holds a special role for the whole interview. The results of the analysis are presented in Section 3.1.3.

Example Each transcription is investigated according to statements regarding working tasks in social science research. The corresponding text passages are labelled. Afterwards, all labelled passages are collected and summarized according to their main message. The messages are compared to each other in order to identify similar perspectives and opinions between the participants of the interviews. Similar or equal perspectives indicate a common sense of specific topics.

For judging the goodness of the conducted interviews and their analysis [May02, Kru11] describe the generic quality criteria of qualitative research:

- **Procedural Documentation.** The process of data collection has to be documented in detail, including planning, methodology, conduction and analysis.
- **Argumentative Interpretation.** Interpretations have to be argumentative and logical.
- **Rule-guided Safety.** The research has to be conducted according to common methodologies.
- **Closeness to the Object.** The researcher has to get as close as possible to the object in order to observe its everyday life.
- **Communicative Validation.** The results of an analysis should be presented to the observed object again in order to get feedback about the validation.
- **Triangulation.** The quality of research results can be judged, if they accomplish several cycles of analysis, e.g. with different data or different methods.

The design, conduction and analysis of the interviews in this thesis have been followed these criteria.

3.1.2 Preparation and Conduction of Interviews

The qualitative expert interviews have been prepared and conducted in accordance with [May02, Kru11]. Thus, we have defined a guideline as per [Kru11], which supports the conduction of the interviews in order to keep the focus on the main topic of the interview. The guideline is not followed too strictly during the interviews, but is used as preparation and support to enable the interviewer to be a competent interview partner to the expert [Lit08]. The interview is roughly separated into two parts by the guideline. The first part deals with general problems of the interviewed person regarding the analysis of research data from their working perspective. The second part discusses the identified problems related to a specific example and introduces an alternative scenario, where Linked Data technologies are applied. Finally, the implications of this alternative scenario are discussed. We have chosen to avoid using the technical terms surrounding Linked Data and Semantic Web (and the terms themselves) because our intention has been to investigate the potential benefits of these technologies for an end-user, who should not be bothered with technical details, but with the implications enabled by them. Introducing technical terms to participants who have no or little technical background can increase the duration of the interviews massively and holds the possibility of leading the interview into a direction that is not intended by us. The complete guideline for the interviews can be found in Appendix A.

Since we conduct qualitative interviews, we have chosen to use a qualitative sample as per [Kru11]. It is based on a consciously chosen small sample, which consists in our case of six experts. Hence, we have to preserve the maximum of structural variability. Thus, we have decided to vary the sample according to the grade of scholarship of the participating persons as well as the data with which they typically work or conduct research. This has led to a sample ranging from postgraduates to senior researchers in the fields of sociological and political science, who are working with different types of data (e.g. survey data, statistical data, or others) and with data of different subjects (e.g. election data, behavioural data, crime data, all on national or international levels). The selection of the interviewees has followed the snowball system described in [Kru11], where participants are asked for additional suitable or interested participants.

Information on the six participants:

- **Age:** 32 to 59 years
- **Profession:** 4 research associates (2 postgraduates and 2 postdocs), 2 professors
- **Experience in social sciences:** 4 to 32 years
- **Experience in LOD:** 0 years

The intended duration for each interview was estimated at one hour. The effective length of the interviews is pending at between 33 and 57 minutes. The interviews have been conducted in German because all participants are native speakers. The interviews have been recorded with the software No23 Recorder[1]. The recordings have been transcribed literally using audiotranskription.de[2] to ensure a detailed analysis. The transcriptions have been created according to the simple rules described in [DP11], which means a literally transcription including temporal annotated pauses and special accentuations of words. This has been sufficient for our purpose.

3.1.3 Results

The qualitative interviews have been conducted with six experts, who are researching in the social and political sciences. They have been recorded and, afterwards, transcribed literally. We have conducted a qualitative content analysis of our interviews.

Analysis Categories Following the guidelines for qualitative content analysis, we have defined four analysis categories. We have analysed the interviews considering these categories in the context of the integration and connection of research data and information.

[1] http://no23.de/no23web/MP3_OGG_Aufnahme_Software.aspx?smi=1

[2] http://www.audiotranskription.de/

All of the presented categories are required in order to investigate a beneficial use of Linked Data for social scientific research. The categories have been abstracted from our guideline and are presented in the following list:

- **Working Tasks.** Statements about concrete working tasks deriving from the information needs of the interviewed experts are considered in this category. This supports the understanding of the context, in which Linked Data standards and technologies can be applied beneficially.
- **Problems and Challenges.** In this category, statements regarding problems and challenges during the integration and connection of data and information are examined. An analysis of these statements aims to identify connection points for technical solutions and implementations of Linked Data standards and technologies.
- **Benefits of Linked Data[3].** Statements regarding the potential benefits of a linked and connected representation and availability of research data and information are analysed in this category. This also includes requests of the participants regarding what would be helpful or supportive for their work. In this category, we aim to identify true benefits of Linked Data for social science research, which are generated by the use of Semantic Web standards and technologies.

A major result of the interviews has been that two experts have heard of Linked Data, but nobody has ever consumed Linked Data sets for research. Further results of the interviews are ordered among the three identified analysis categories.

Working Tasks

All six experts referred to various precise information needs that emerge directly from their research or service work. By summarizing these different research intentions to working tasks, four central tasks can be identified. The two most frequent tasks have been the design of an own data set comprising and integrating variables and data from distributed data sets, and the merging and accumulation of data according to a specific variable mostly over a range of time. Five of the six experts have mentioned that they typically integrate or merge data by enriching a first data set with additional data from a second data set over a specific key variable. This key variable usually contains entries of taxonomies or code lists like geographical codes, political parties, etc. Additionally, respectively one person mentioned the extension and documentation of data sets with context information like literature and the searching for data.

Problems and Challenges

Although the statements of the experts were similar in the first analysis category, their personal problems and challenges during these tasks expose a broad variety. This could

[3] In order to avoid misunderstandings of the technical idea and concepts of LOD, we have used an abstracted imagination of interlinked and connected data during the interviews.

have been expected because of the variety of research questions and interests in social science research. Most of the problems can be grouped into four topics: data retrieval, data access, data documentation and modelling, and data matching and integration

- **Data Retrieval.** All six experts mentioned that gathering research-relevant data is one of the most time-consuming tasks during research. This problem occurs not only on the web, where it is not always known which data is where available and to what extent. Four of the experts addressed the problem to the non-digital world as well, where agencies and organizations have to be asked whether specific data is available. Complicating the problem is the fact that data is often published incomplete, e.g. some values have been proven as wrong, have been lost over the time or have never been collected for a specific country.
- **Data Access.** Four experts complained that relevant data is not always available on the specific required level of aggregation, i.e. data to a particular variable or indicator is not always available for, e.g. countries, districts, and cities simultaneously. Another problem (claimed by two experts) is that some information is only available following payment. This includes especially data mappings, whose creation has been expensive and extensive, e.g. geographical coordinates with specific geographical context information. Three experts mentioned that when working with data on the individual level, data privacy restrictions hinder researchers from accessing and, especially, reusing particular collected data. This is especially the case when aiming to connect sensitive data with further context information, which might allow for an identification of the individual persons. Five experts conclude that when specific data is unavailable or when information is missing, e.g. inside a time series, the researcher can either leave it out with respect to the original research intention or, alternatively, investigate for alternatives or try to reconstruct or calculate the specific missing data value or variable by himself or herself. Again, five experts have mentioned that such a reconstruction or calculation is common practice.
- **Data Documentation and Modelling.** Additional problems concerning the search and an intended use of data lie in their lack of documentation. Often, not all the specific attributes of data items are included in the documentation and are, therefore, unavailable to the user. But changes in variable definitions or questions, e.g. over time, are of high relevance for researchers. Three experts have complained of this problem. Also, differences in the definitions of variables, e.g. unemployment, between different data providers are referred to as relevant decision criteria by two participants of the interviews. It is not unusual that such specific information about data is not available in its documentation, because the data has still not been preprocessed for scientific use, i.e. necessary information about the process of data collection, how variables are constructed and defined or information on question filtering is missing (stated by one expert).
- **Data Matching and Integration.** Major challenges regarding the mapping of variables and, specifically, the mapping of entries of code lists for these variables can be identified. Since data is typically matched or enriched with context information

according to a key variable, e.g. countries or political parties, not only does the key variable itself have to be identified within the involved data sets, but also their possible entries have to be mapped to each other. Three experts have referred to this problem. Moreover, this challenge varies in complexity. In some cases, it can be carried out very easily and clearly, e.g. mapping of country names. But problems can occur during such mappings if there are ambiguous or incomplete mappings (claimed by two experts), which is, for instance, the case with administrative districts and electoral wards. In some cases, entries of such code lists describe different granularities or summarize attributes, which also increases the complexity.

- **Data Preprocessing and Data Reuse.** In general, two further aspects have been mentioned by the experts as being problematic and time-consuming. Before analysing data, major effort has to be put into the preprocessing of data (three experts), which typically means the conversion of data from formats like PDF or printed documents into formats required by statistical tools. However, all experts also emphasize that this effort need not be made in many cases because a lot of data is available in processable formats. The second aspect is that a lot of work especially regarding the mapping of variables or structural information has been carried out repeatedly by researchers, although one can be sure that specific work has already been done by others. However, this information is mostly not available as two experts have mentioned.

Benefits of Linked Data

In order to focus on the main ideas behind LOD and since none of the participants has ever used Linked Data sets, technical details have not been discussed with the participants. Moreover, an alternative scenario that accords with their research tasks has been presented and discussed. In these scenarios, specific data sets, variables or sources of context information are already connected and interlinked. Also, a detailed and fine-grained description of data and information is available.

Three major benefits have been identified by the experts.

1. **Data documentation.** First, as contribution is seen as a more detailed and fine-grained description of data with respect to the specific information necessary for scientific use. This has been stated by three experts.
2. **Data linking.** Three experts have claimed that the enrichment of data, e.g. statistics, with context information would be a value addition for researchers. But they also have doubts with respect to choosing context information to link to, which is relevant to a preferably large group of researchers. This once again pertains to the variety of research interests.
3. **Data matching.** The third expected benefit, which has also been named by three experts, is the creation and availability of mappings between entries of code lists, which is used when enriching or integrating data according to specific key variables.

> This especially refers to the mappings of structural information like geographical regions, electoral wards, etc. that have to be done repeatedly and which could be omitted, if such mapping information were available for reuse.

An additional benefit, as identified by the participants, lies in the possibility of an easy search for available and relevant data based on detailed data documentation. This has been named by two experts. The following benefits have each been claimed by one expert. They could imagine a technical possibility for an easy connection of data, better cross-search over data sets for, e.g. similar variables, and accessing everything that they need together at one time, i.e. data that simultaneously offers precise and detailed documentation, relevant literature and context information.

3.1.4 Discussion and Limitations

The results of the interviews show that there are working tasks in scientific research, where Linked Data technologies may be applied beneficially. Although only two experts have heard about the ideas of LOD and none of them has used it yet, all participants were open-minded regarding new technologies and methods that may support their research.

The major problem of finding data can be addressed by a detailed, fine-grained and inter-connected description of data and information as well as by effective information retrieval methods. The mapping between variables or entries of code lists is currently usually done manually. Methods of schema and ontology matching can support such tasks. The semantic richness of Linked Data sets can decrease the barriers for reconstructing and calculating data values or variables by the researchers themselves. In general, the reuse of data is increased when it is available as Linked Data. This affects links between different data sets, variables, mapping information or self-created data sets. In times where the reproducibility of science gains in importance, Linked Data can make a valuable contribution.

Finally, all the interviewees have expressed one major concern: the identification and decision of what exactly to link to and what kind of links might be relevant for researchers. This issue is difficult to address since research interests and questions in the social sciences are of such a variety that only a small common denominator can be identified by the experts so far. The general benefits of linking data can be derived by linking and mapping general or structural data and information like geographical regions, coordinates, etc. A solution in this regard can be the possibility of finding and establishing links independently according to their particular research interest.

3.2 Prototypical Semantic Data Library for the Social Sciences

Following the results of the qualitative interviews, we now investigate how Linked Data can be applied technically into the scientific research process and the challenges that have to be addressed. In this section, we present a framework for a Semantic Data Library for the social sciences. In contrast to a traditional notion of libraries, the focus of this section is on research data and not on literature, since it is based on the use case 'Analysing Research Data' introduced in Section 1.2. By the use of semantic technologies, this framework aims to identify and overcome the typical challenges that digital libraries and archives are currently facing, when dealing with heterogeneous data and trying to provide useful services to ambitious users. The framework initially sought to exploit the potential of enriching and analysing statistics with LOD, as presented in [ZHM11], and has been moved forward by the development of a LOD pilot application for the social sciences, which serves as prototype for the Semantic Data Library and is presented in Section 3.2.4. The work presented in this section is the result of a cooperation between technology experts from the Karlsruhe Institute of Technology (KIT)[4], the Institute for Web Science and Technologies (WeST)[5] and GESIS on the one hand and data experts from the statistical office and IT service provider of the federal state North Rhine-Westphalia IT.NRW[6] and the research data centres ALLBUS[7] and Elections[8] of GESIS on the other hand. The results have been published in [GHHZ11].

In Section 3.2.1, we provide the theoretical background of our work. We present a detailed description of the use case 'Analysing Research Data' applied in this framework in Section 3.2.2. In Section 3.2.3, the framework of a Semantic Data Library and its key modules are presented as a requirements list. A first prototype implementation is presented in Section 3.2.4. In Section 3.2.5, the remaining open issues are discussed.

3.2.1 Theoretical Background

Libraries and archives follow a long tradition of surveying, collecting and classifying available knowledge and of providing access to these high-quality information resources. With the distributed publishing paradigm of the web, providing such services has grown in complexity, driven by a multiplicity of exchange formats, different terminologies for metadata annotations and missing connections between distributed data sets. However, researchers cannot use distributed data on the web in the same way as they are used to in libraries and archives. One reason is that digital libraries and digital archives are often still disconnected from each other – not only because of historical and disciplinary reasons, but also because they use different standards and formats. The results from several studies prove that a change in the organization, management and offering of library data is necessary in order to support scientists in their research [MBD$^+$07, TP11, KPGC11]. Semantic Digital Libraries and Archives [KM09b] aim to overcome these challenges. They address key challenges like information integration and interoperability as well as user-friendly interfaces, all supported by semantic technologies and community interactions [KM09a]. Semantic Digital Libraries can be seen as the next step and further evolution

[4] http://www.kit.edu/english/index.php
[5] http://west.uni-koblenz.de/
[6] http://www.it.nrw.de/
[7] http://www.gesis.org/en/institute/competence-centers/rdc-allbus/
[8] http://www.gesis.org/en/institute/competence-centers/rdc-elections/

of traditional digital approaches, which often lack the implementation of Semantic Web and social networking technologies. Considering a digital library of distributed data, semantic technologies can facilitate the integration of data from disparate sources.

Quantitative research in the social sciences heavily relies on the analysis of survey and statistical data. The emerging field of ‘Computational Social Sciences’ leverages the possibility of collecting and analysing large-scale data sets to potentially reveal patterns of behaviour of individuals and groups [LPA^{+}09]. The necessary data for such an approach is often difficult to find, integrate and process, which is due to a mostly decentralised and historically grown distributed publication and archiving of data in, e.g. government agencies, research data centres or universities. Scattered information due to organic growth also occurs on the web at large. To be able to judge the relevance and quality of the data for any upcoming analysis in research, it is important to gain deep insights into both, data and especially its documentation. Besides descriptive standard information, the metadata of data used in analysis shall provide extensive information about the methodology, sample design, necessary weights or notes on the safe and correct handling of the data concerning privacy and provenance. A lack of metadata annotation complicates the process of data search on the web as well as the comparison of different data sets, e.g. regarding concrete indicators or samples.

While sizeable amounts of data useful for research are attainable through the web, the data is published in a large variety of data formats. To process and analyse data, one has to convert data into the particular formats of statistical tools and integrate data from multiple sources. In general, data conversion and integration is not a technical barrier, but the effort spent in conversion is a nuisance, especially for necessary but tedious routine tasks, such as gaining a first insight into the data, or in cases where the expected research gain is minor. All these problems hinder a reuse of available and valuable data resources; when comparing or integrating multiple data sets these efforts increase.

To overcome the challenges that digital libraries and archives are facing with regard to distributed data on the web, we propose a framework for a Semantic Data Library of Linked Data, which is relevant for research in the social sciences. While the framework provides central services for the accessing, processing and integration of distributed data sources, their physical storage location remains distributed and will not be collected or hosted by the data library. The difficulties in searching, modelling and annotating distributed data are addressed not only on the metadata level, but also on the directly connected underlying numerical data, which provides researchers with on-the-fly usage of the data in visualizations or for statistical analysis. We present a prototype implementation that demonstrates the automatic aggregation and integration of data using wrappers and a common exchange format.

3.2.2 Application Scenario

The Semantic Data Library revisits our use case of ‘Analysing Research Data’ introduced in Section 1.2. As an organization providing infrastructure for the social sciences, GESIS

– Leibniz Institute for the Social Sciences offers a wide range of different study series as well as empirical primary data from survey research and historical social research. At the beginning of any research, scientists usually have a first idea regarding what kind of data they will need and which analysis method they would like to perform on the data. For example, a researcher would like to investigate possible correlations in a correspondence analysis of unemployment rate, immigration quota and the subjectively perceived risk of unemployment in Germany. However, the desired data is only available from different authorities. While the researcher can retrieve statistics from German statistical offices, data on attitudes, behaviour and social structure in Germany is part of the German General Social Survey ALLBUS, which is archived at GESIS. On the web portals of GESIS, the ALLBUS metadata can be searched, so the researcher can gain insight into the documentation of the data and is able to decide whether ALLBUS is (completely or partly) relevant to the research interests. To make a decision regarding whether the data is suitable for the intended analysis method, a comprehensive and detailed documentation of the data is essential. Information on, e.g. sample design, populations or possible bias and variance has to be provided. In case researchers would like to analyse more than one data set, the individual data sets have to be aligned, i.e., not only technically, but also considering differences in populations or aggregation levels.

Using statistical tools such as STATA[9], SPSS[10] or the R Project[11] might require the data to be converted into application-specific formats. When dealing with different data sets, it has to be clear what dimensions and samples the data is comparable to and thus how data can be matched up. For example, data from ALLBUS has to be aggregated for it to be comparable to any statistics, because ALLBUS is micro data and, therefore, determined at an individual level due to its origin as survey data. What sounds feasible from a technical point of view is not trivial from a researcher's perspective. Weightings and potential transformations vary according to the intended research query, which means that there exists a wide range of possible calculations and a certain degree of openness in the way of using survey data. The matching is mostly done manually before importing the integrated data into statistical tools, although some tools can automatically detect comparable dimensions like time or geographic regions. Finally, the researcher analyses data as well as defines and executes statistical functions depending on the desired analysis method, such as multidimensional analysis, time series analysis, correspondence analysis or estimation procedures in complex designs [KKV94, SHE05].

After completing research and data analysis, researchers ought to cite the used data sets in the resulting publications. Referencing the analysed data helps fellow researchers to comprehend the analysis done with the data. Data can be cited and afterwards identified by using a URI (Uniform Resource Identifier) or a DOI (Digital Object Identifier). Newly created data during research obtains an identifier only if it is published afterwards.

In general, retrieving and analysing data on the web is nothing new to researchers in the social sciences. The providers of research data are interested in offering the possibility

[9] http://stata.com/

[10] http://www-01.ibm.com/software/uk/analytics/spss/

[11] http://www.r-project.org/

for browsing, analysing and downloading their data, even if it is only metadata due to privacy restrictions. Examples are ZACAT[12] by GESIS and SOEPinfo[13] by the Research Data Centre of the SOEP[14] (Socio-Economic Panel Study). Both portals offer a wide range of tools for processing, analysing, visualizing and exporting data to different data formats. However, both are restricted to the data holdings of their particular organisation. There is no connection point to other external data sources except that the user exports the data and loads it in an extra application for further processing and calculations. A web-based application without data restrictions is GraphPad QuickCalcs[15], a collection of free online calculators for different analysis purposes. It is offered by GraphPad Software and enables statistical calculations based on data, i.e. numbers entered by the user manually. However, calculations are only possible on single numbers and not on complete data, which separates the data from its context and meaning. A combination of different data sources seems to be possible, but a lot of manual work is left to the user. To summarize, current approaches are either restricted to data sources that can be used for an analysis, or restricted in their functionality.

3.2.3 Framework for a Semantic Data Library

In this section, we introduce a generic framework of a Semantic Data Library in order to address the key challenges for semantic library services providing survey and statistical data in the social sciences. Thus, the framework is oriented on the goals of Semantic Digital Libraries [KM09a], but sets the focus on research data. As we focus on the retrieval, integration and analysis processes, our approach can be distinguished from the concept of Semantic Digital Archives, because we do not address typical archiving services like long-term preservation or data curation. Our framework is composed of modules for identifying and exchanging, searching and integrating, evaluating and publishing data. Thereby, we address the main obstacles for reusing statistical or survey data in the social sciences. Figure 3.1 depicts an overview of the architecture of the framework.

The framework consists of seven key modules, each of which covers specific aspects from the path of publishing data as Linked Data to the graphical preview, statistical analysis and export of data. These modules are: (1) Common Identifier Format, (2) Common Data Exchange Format, (3) Retrieval of Data, (4) Linking and Integration, (5) Graphical Data Preview, (6) Data Analysis and (7) Export and Referencing. The first two modules are comprised in the depicted LOD wrapper, which exposes data as Linked Data. The data can then be queried via SPARQL. Based on the exposed dimensions and measures (by the SPARQL queries), the user can link and integrate data according to his research intentions. This means that multiple data sets are not linked automatically during the retrieval with SPARQL queries. Moreover, it is left to the user to decide which data sets and which of their dimensions should be linked and integrated. This decision has been

[12] http://zacat.gesis.org/

[13] http://panel.gsoep.de/soepinfo/

[14] http://www.diw.de/en/soep

[15] http://www.graphpad.com/quickcalcs/index.cfm

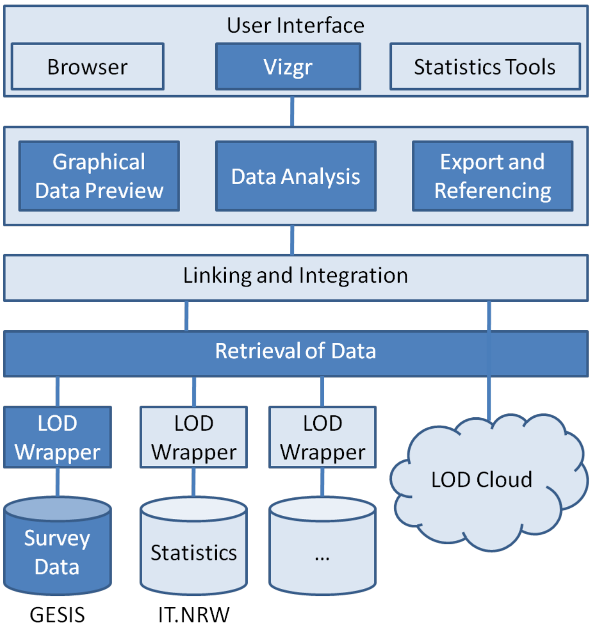


Figure 3.1: Framework for a Semantic Data Library.

made together with the domain experts from the research data centres working on this prototype based on the fact that research questions can vary massively. The integrated data can be previewed in a graphical way, e.g. as a line diagram analysed with simple statistical calculations, and exported in common formats in order to be used in additional tools. Users can access the Semantic Data Library from a web browser, but a connection to other tools capable of data analysis, like statistical tools or the Vizgr [HZSM12] tool, is possible. The different modules are describes in detail in the following paragraphs.

Common Identifier Format The identification of data sets, measurements or dimensions is of importance for a variety of reasons. On the data level, a unique identifier enables referencing the data set itself. Referencing is crucial in the context of making data sets citable in scientific publications, thereby providing valuable metadata about the scientific work. For this reason, identifiers like DOI, URN or Handles are typically used. But, according to the Linked Data principles, the use of URI, a core ingredient to Semantic Web technologies, is preferred because it is resolvable by HTTP with the need of an extra resolver. After discussions regarding the use of DOI as an identifier for Linked Data[16], CrossRef[17], the official DOI link registration agency for scholarly and professional publications has announced that DOIs can be used as HTTP URIs with

[16] Exemplary discussion can be found at
http://www.crossref.org/CrossTech/2010/03/dois_and_linked_data_some_conc.html and
http://bitwacker.com/2010/01/19/the-doi-datacite-and-linked-data-made-for-each-other/

[17] http://crossref.org/

content negotiation[18].

Within the data, the Linked Data identifiers provide a way to identify the semantics of dimensions, measures and observations as well as detailed metadata information. URIs fulfil this requirement. With respect to the integration and aggregation of data sets, particularly the semantics of the dimensions is of interest.

Common Data Exchange Format There are a few well-established and proven formats for statistical calculations. Amongst others, Excel spreadsheets, SPSS, SAS, Stata and R native formats are used to exchange data including the respective formulas. Unfortunately, these formats are proprietary (locked) and/or in binary format, which makes it difficult to transform data seamlessly from one format to another. Moreover, these well-known formats do not describe their data in an expressive way, i.e. in a way that is expressive enough to deliver self-explanatory data via metadata. For the purpose of a data library for the social sciences, it is necessary to integrate various heterogeneous data sources and perform calculations directly on data or on aggregated items coming from these sources. To achieve direct calculations, we are interested in self-explanatory or self-descriptive data sources that deliver generic structures which can be semantically processed further. Thus, we aim for annotated or metadata-enriched data formats that promote easy exchange, integration and annotation using data from many heterogeneous sources. These requirements are well met by the Data Cube format [CRT14] since it (a) is an open, non-proprietary metadata model in the RDF format, (b) is widely based on the established SDMX information model [SDM02] and also includes other vocabularies, and (c) provides a semantic and self-descriptive annotation of the data. Given these advantages, it is likely that this metadata model will be supported by established statistics packages or that converter programs will be developed. The advantages of Data Cube foster a thorough adoption by practitioners as well as facilitate an easy deployment and publication of statistical and survey data. Another advantage of Data Cube is that thanks to its flexibility and simplicity, it is easy to convert existing data. In our prototype implementation presented below, we actually use efficient wrapper modules to convert proprietary or other non-semantic formats on the fly to the Data Cube vocabulary. Modelling examples using the Data Cube vocabulary are presented in Section 3.2.4 through the implementation of a prototype.

Retrieval of Data The ability to find relevant data sets is a key factor for enabling social scientists to make use of existing data sets. Therefore, an efficient retrieval module is necessary to ensure that the search of data is suitable for the respective research topic. Metadata, which is semantically annotated highly expressive, delivers more processable input for traditional information retrieval algorithms. Thus, more details about the requested data become evident during the retrieval process, e.g. the granularity of

[18] see the announcement at http://www.doi.org/news/DOINewsApr11.html#2 and a detailed description of the implementation at http://www.crossref.org/CrossTech/2011/04/content_negotiation_for_crossr.html

specific dimensions or the frequency of observations. To provide researchers with useful information about a data set, extensive metadata must be available. Metadata not only supports the retrieval process, but nust also be considered afterwards to be able for evaluating relevance, quality and suitability for the following analysis process. For comparative research, the description and attributes of for example different indicators, sample designs and populations have to allow for comparisons to those of other data sets. Eventually, the retrieval module should provide the underlying data itself.

The semantic description of the data also enables more complex search tasks. For instance, if a researcher is interested in the GDPs of European countries, the available data provides these figures in the currency of the corresponding countries and not all of the data might be provided using Euro as a currency. If a second source can deliver the conversion rate, it is possible to combine the data sets and produce the requested information. Beyond the actual retrieval of the data sets, the module will need to provide a simple interaction component to define possible common dimensions by which data sets should flexibly be merged and integrated, i.e. temporal or geographical areas. Therefore, the task of the retrieval module is twofold: retrieve (a) metadata about the data sets (e.g. using taxonomies as is common in libraries like SKOS [MB09b]) and (b) the data sets themselves.

Semantic technologies can aid in the combined querying of data. Both, descriptions of a data set (such as author, publication date) and the data itself (individual observations), can be encoded and interpreted by machines. Thus, an integration is made possible. Both are required: descriptions of the data (e.g. author, responsible organisation) and the data itself (the individual observations). Once data has been published in a uniform base format (e.g. RDF), machine-supported integration is possible. There are several services possible on integrated data, e.g. keyword search [LT10] or faceted browsing [WLT11]. VisiNav, in particular, offers navigation functionality over data integrated from the web [Har12]. Combined querying of distributed and heterogeneous data is also discussed in detail in [FCOO12, HL10] and in [HBF09], which traverses RDF links in order to discover potentially relevant data during the query execution.

Linking and Integration The semantic representation and annotation of data allows for services far beyond the simple retrieval and provisioning of data sets. As the semantics of dimensions, values and metrics is explicitly modelled in the data, automatic linking and integration of data is at a researcher´s fingertips.

To correctly join and merge two data sets, it is necessary to identify common dimensions, align and map the according values, and possibly aggregate some of the data entries. Based on the dimension concept in Data Cube and the possibility for semantic annotation, the identification step can be carried out easily. Alignment of the values requires some more insights and may be achieved by a more detailed model and description of the data. In data with a temporal dimension, for instance, it is necessary to define its format and distinguish between frequencies or between values in percentages and values in absolute numbers. Aggregation becomes necessary when there is no comparable representation and the data values need to be summed up, or averaged. Again, the semantic description of the dimension has to provide the exact information that is necessary to know which aggregation function to apply. This is not only relevant for further calculations, but also for combined visualizations, e.g. the combined representation of different data sets in a line graph where the axes have to be aligned to each other.

Graphical Data Preview For any existing or newly created (by the means of linking and integration) data set, the first approach for a researcher is typically to examine some key characteristics of the data. Therefore, together with the provision of the data itself, the library presents some results of a simple statistical analysis. For existing data sets, key characteristics can be pre-computed; for freshly integrated data, an overview will be generated on the fly. We benefit from a semantic representation of the data that allows for a better notion of which characteristics are of interest and which dimensions need to be looked at.

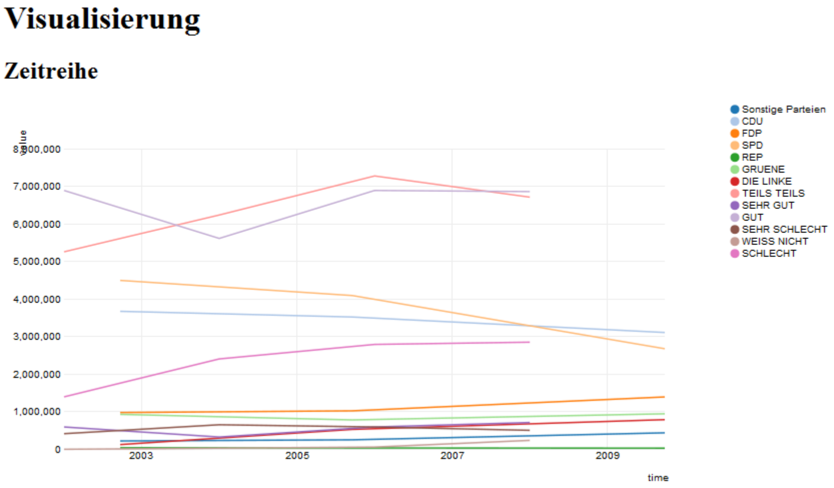


Figure 3.2: Visualization of a combined query on different data sets [ZHM11].

To make a first glance analysis easier, data sets can be presented in a graphical form, plotting key indicators over the main or common dimensions of integrated data sets. Figure 3.2 depicts a combined visualization of two heterogeneous data sets [ZHM11]. Such visualizations depict the need for an alignment of values according to the axes of the diagram. While the depicted election votes from the first data set range from values of zero to approximately 10,000,000, the values of the other data set, the survey results of a study with a sample of 1,000 persons always lie beneath 1,000. This complicates the legibility of the graph. Furthermore, the use of flexible queries on the data allows easy adjustment of the graphs. In Figure 3.3, the first visualisation has been filtered to show only one party and one specific type of answer of the chosen data sets.

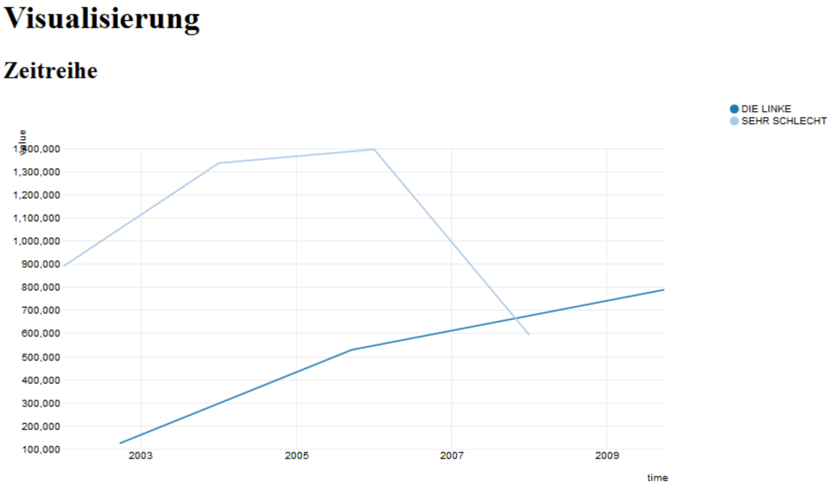


Figure 3.3: Filtered Visualization with one party and answer type.

Data Analysis The information infrastructure for working with research data are often too strictly adjusted to data sources that are traditionally used or to specific domains or purposes. Considering the LOD movement, a method for performing mostly secondary research tasks is needed to allow for statistical queries and calculations on standardised data sets. The calculations for weighting and transforming the data as well as the statistical methods applied afterwards (e.g. regression analysis) is performed on the level of the integrated data corpus. Since the use case consists of precise tasks and we do not claim to replace statistical tools, our prototype provides the possibility of performing basic secondary analyses based on a small set of implemented functions. For implementing more complex statistical calculations, existing sources from the R Project for Statistical Computing can be reused. An alternative is the extension of the SPARQL query language by query rewriting in order to transform SPARQL queries to particular statistical functions. To allow more comprehensive analyses, the system provides export capabilities to standard tools such as SPSS and STATA. Our scenario focuses on the processing of data of the same hierarchical (aggregation) levels. The data integration layer is virtual, i.e. the integration layer provides access to data that remains at its original source.

Recently, there have been various approaches for analysing Statistical Linked Data sets. Most similar to our approach regarding the retrieval and manipulation of data are the SPARQL plugin[19] for the R Project and the OLAP-based approach in [KOH12]. Both use SPARQL queries for loading the data into the particular systems and allow, in a subsequent step, various operations on the data. The LiDDM system [NKIV11] is similar to our approach and offers a complete framework for retrieving, integrating, filtering and mining Linked Data. Statistical analysis of data is enabled by mining the integrated

[19] http://cran.r-project.org/web/packages/SPARQL/index.html

data for, e.g. patterns using the Weka tool [HFH$^+$09]. The user is involved in the full process and can decide what data is retrieved, how it is integrated and according to which constraints it is filtered and mined. User interaction during these processes is very important in order to support the wide range of various different information needs. A different approach, which uses Linked Data as background knowledge, is presented in [Pau12]. The approach Explain-a-LOD aims to explain statistical information by enriching it with Linked Data. Hypotheses are then generated for explaining a particular statistical fact based on a combination of statistics and Linked Data.

Export and Referencing While the preview and basic analysis can provide first insights into the data it neither can nor is supposed to replace the analysis based on a full statistics application. Therefore, the system needs to allow for exporting the data to enable downstream processing. An export service providing data sets in a selection of common formats (like CSV, RDF, or Excel) is crucial for feeding into the individual scientific processing pipelines of research groups. Exporters are needed particularly as long as the RDF Data Cube format itself is not supported by all major statistics tools.

As each data set is compiled based on user-defined parameters and needs, the data set can be reproduced at any time. Parameters can also be used in a unique identifier to a data set. Thereby, data sets can be referenced and cited.

3.2.4 Prototype Implementation

The motivation behind the prototype has been to implement a first use case relevant to social science research and to investigate further areas of research utilising state-of-the-art technologies. However, we focus on the integration and analysis of data since search/retrieval of data on the Semantic Web is an already established field of research [FCOO12, HL10, Tra10]. The prototype aims to integrate, aggregate and visualize data from two sources, ALLBUS and IT.NRW.

The research question underlying the use case of the prototype has been whether there are correlations between the number of votes per party and the people's ratings of the economic situation (both personal and national prospect) in the federal state of North Rhine-Westphalia. Therefore, ALLBUS[20] provides survey data of individuals rating the personal and national economic situation. IT.NRW[21] provides the number of votes per party of elections to the "Bundestag" for the state of North Rhine-Westphalia.

ALLBUS The German General Social Survey ALLBUS, which collects up-to-date data on attitudes, behaviour, and social structure in Germany, is archived at the GESIS – Leibniz Institute for the Social Sciences. Due to data privacy restrictions, we use a special processed version of a subset of ALLBUS/GGSS 1980 - 2008 (Cumulated German

[20] http://www.gesis.org/en/allbus/allbus-home/

[21] http://www.it.nrw.de/

General Social Survey 1980-2008) [ALL10], which includes only a few variables that are relevant for our use case ('Current Economic Situation in Germany' and 'Resp. own Current Financial Situation'). Additionally, only data from participants from North Rhine-Westphalia has been included into the subset in order to make it comparable to the election statistics from North Rhine-Westphalia on a geographical level. Due to the omitting of a lot of relevant information and variables for the subset, it has been explicitly created for technical feasibility experiments only.

Election Statistics The election statistics from the German federal state of North Rhine-Westphalia are provided by IT.NRW, the statistical office and IT service provider of the federal state. The data are offered as HTML tables as well as CSV via a web service. The statistics contain the election votes and results for the elections of the German parliament, the parliament of North Rhine-Westphalia and the European parliament. Both, votes and results, can be retrieved on different administrative levels, e.g. from the federal state and from administrative districts down to single electoral districts.

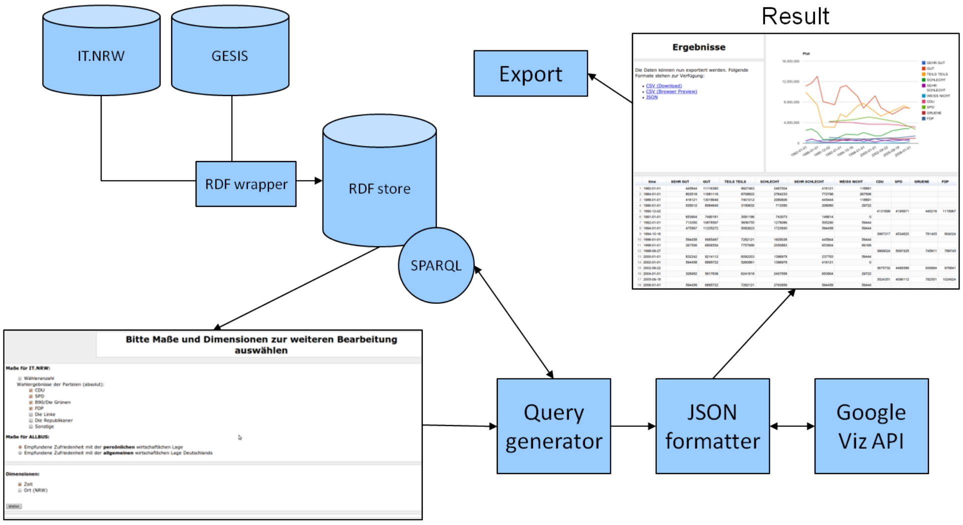
Architecture

Figure 3.4 provides an overview of the architecture for the prototype which can be accessed online[22]. We do not include search capabilities in the prototype for retrieval of data sets since we only process a few data sets. Therefore, we enable the user to select manually the data sets to be used from a fixed list. For the integration step, all data in Data Cube format is then collected in an RDF memory store and accessed via a SPARQL endpoint on top of the RDF store. In our case, we use OpenRDF's Sesame library including a SPARQL interface since the prototype is implemented as a Java-based web application on an Apache Tomcat infrastructure using servlets. Via these servlets, the user can choose which parties and which economic rating should be visualized. Based on the chosen data, a SPARQL query is formulated by the system and the desired data is then retrieved from the Sesame store. The data is formatted into JSON format, which is necessary for the inclusion into a diagram and table representation generated by the Google Visualization API. The visualized result can be converted in JSON or CSV format, which contains the integrated data set of ALLBUS and IT.NRW.

Generating and Exposing Data

To identify data items and corresponding dimensions, measures and attributes, we use RDF URIs. The participating data sets are represented in the Data Cube vocabulary. Data Cube-compliant data is generated by on-the-fly wrappers from our IT.NRW data source and by a conversion of data exported from the ALLBUS database. The wrapper receives data in CSV format from a web interface of IT.NRW and converts it to RDF.

[22] http://lod.gesis.org/gesis-lod-pilot/



Ergebnisse

Die Daten können nun exportiert werden. Folgende Formate stehen zur Verfügung:

- CSV (Download)
- CSV (Browser Preview)
- JSON

Bitte Maße und Dimensionen zur weiteren Bearbeitung auswählen

Maße für IT.NRW:

□ Wähleranzahl
Wahlergebnisse der Parteien (absolut):
☑ CDU
☑ SPD
☑ B90/Die Grünen
☑ FDP
□ Die Linke
□ Die Republikaner
□ Sonstige

Maße für ALLBUS:

● Empfundene Zufriedenheit mit der **persönlichen** wirtschaftlichen Lage
○ Empfundene Zufriedenheit mit der **allgemeinen** wirtschaftlichen Lage Deutschlands

Dimensionen:

☑ Zeit
□ Ort (NRW)

Weiter

Figure 3.4: Implemented prototype.

One observation of a measure (the voting results for a German political party) from the IT.NRW data appears as illustrated in Listing 3.1.

Listing 3.1: Example observation of an election statistics (Source: IT.NRW).

```
[   a qb:Observation ;
    qb:dataset <./data?code=14111#ds> ;
    dcterms:date "2009-09-27" ;
    geo <geo.rdf#051> ;
    partei <./parteien.rdf#CDU> ;
    sdmx-measure:obsValue "845318" . ]
```

The data sets of our use case consist of a high complexity according to the number of observed dimensions and measures. Especially the ALLBUS data set holds hundreds of variables, which determine questions that have been asked during surveys. We represent each variable that has formed a query (e.g. the estimation of the personal economic situation) as a separate observation. This is justified by the used Data Cube Vocabulary, which advises splitting up multiple measures of one observation to separate observations according to better querying possibilities. Listing 3.2 depicts an observation of the ALLBUS data set exposed as RDF using the Data Cube vocabulary.

Listing 3.2: Example observation of ALLBUS (Source: GESIS).

```
[   a qb:Observation ;
    qb:dataset <./ZA4570agg_unscaled.rdf#ds"> ;
    gesis:time <./time#1990> ;
    gesis:geo <./geo#D-NRW> ;
    gesis:variable <#var9> ;
    gesis:valuelabel <#scale4> ;
    sdmx-measure:obsValue "13" . ]
```

The excerpt above describes an observation carried out in `1990` in the geographical area with the code `D-NRW`, which can be resolved as `North Rhine-Westphalia`. The observation belongs to `#var9`, which is resolvable as the variable `v9` with the label `Current Economic Situation in Germany`. Since survey data often organizes possible answers to a surveyed question in a scale, the observation value refers to the scale value `#scale4`, which is resolved as `Bad`. Hence, the example above depicts that 13 participants of the survey have estimated a bad economic situation for Germany in this observation.

According to the RDF Data Cube vocabulary, the RDF representation of both, ALLBUS and IT.NRW, have been arranged using a simple Data Structure Definition (DSD), which is illustrated for the election statistics of IT.NRW in Listing 3.3.

Listing 3.3: Data Structure Definition for the election statistics of IT.NRW.

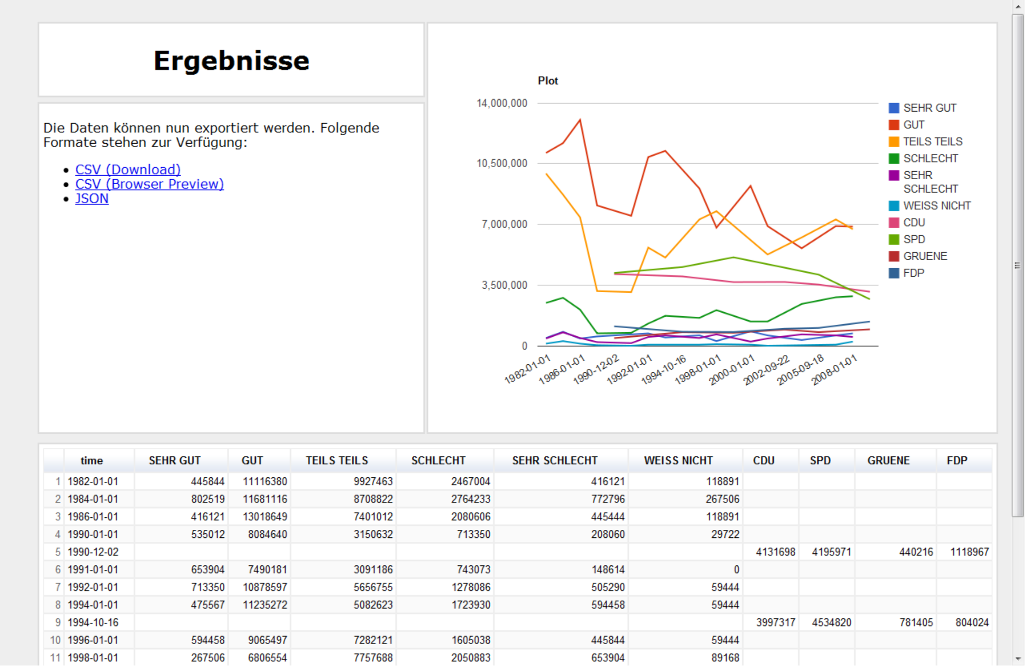
```
<http://lod.gesis.org/gesis-lod-pilot/dsd/14111.rdf#dsd>
  a qb:DataStructureDefinition ;
  dcterms:publisher "IT.NRW" ;
  qb:component _:ag0 ;
  qb:component _:ag1 ;
  qb:component _:ag2 .

_:ag0 qb:dimension <http://lod.gesis.org/lodpilot/ITNRW/vocab.rdf#geo> .

_:ag1 qb:dimension <http://lod.gesis.org/lodpilot/ITNRW/vocab.rdf#partei> .

_:ag2 qb:measure sdmx-measure:obsValue .

<http://lod.gesis.org/lodpilot/ITNRW/vocab.rdf#geo> qb:codeList
    <http://lod.gesis.org/lodpilot/ITNRW/geo.rdf#list> .

<http://lod.gesis.org/lodpilot/ITNRW/vocab.rdf#partei> qb:codeList
    <http://lod.gesis.org/lodpilot/ITNRW/parteien.rdf#list> .
```

Data Integration

During the implementation, we have identified challenges regarding the aggregation of data using current technologies. This is necessary because the ALLBUS data holds data from individuals in contrast to the election statistics from IT.NRW. In order to make both data sets comparable, they have to be available on the same aggregation level, i.e. the ALLBUS data has to be scaled up. Aggregation can be done on application level or data modelling level. Since we use SPARQL 1.0 for querying, aggregation on the query level is still not possible due to the lack of such functionalities in the SPARQL language. Thus, the ALLBUS data is aggregated on the data level. This aggregation has been calculated by considering the number of people giving a particular answer to a question according to the whole population of North-Rhine Westphalia. Such processes of data manipulation have to be included into the metadata in order to reproduce changes on the data.

Result Output

A graphical preview is generated on the integrated data (see Figure 3.5). For this visualization, we use the 2D line chart and table component from Google Visualization API[23], which processes data in the JSON format. Thus, our SPARQL results are transformed into the JSON format. Our visualization allows a time-series analysis of election results in comparison to future prospects of participants of a study by analysing line charts or table data. Additionally, data can be seamlessly exported to CSV and JSON for further analysis in, e.g. external statistical tools. An approach for calculating and analysing statistical data independent of statistical tools is presented in [ZM11b].



Ergebnisse

Die Daten können nun exportiert werden. Folgende Formate stehen zur Verfügung:

- CSV (Download)
- CSV (Browser Preview)
- JSON

	time	SEHR GUT	GUT	TEILS TEILS	SCHLECHT	SEHR SCHLECHT	WEISS NICHT	CDU	SPD	GRUENE	FDP
1	1982-01-01	445844	11116380	9927463	2467004	416121	118891				
2	1984-01-01	802519	11681116	8708822	2764233	772796	267506				
3	1986-01-01	416121	13018649	7401012	2080606	445444	118891				
4	1990-01-01	535012	8084640	3150632	713350	208060	29722				
5	1990-12-02							4131698	4195971	440216	1118967
6	1991-01-01	653904	7490181	3091186	743073	148614	0				
7	1992-01-01	713350	10878597	5656755	1278086	505290	59444				
8	1994-01-01	475567	11235272	5082623	1723930	594458	59444				
9	1994-10-16							3997317	4534820	781405	804024
10	1996-01-01	594458	9065497	7282121	1605038	445844	59444				
11	1998-01-01	267506	6806554	7757688	2050883	653904	89168				

Figure 3.5: Result output of the prototype.

3.2.5 Discussion and Limitations

There are several open issues in the realisation of a large-scale Semantic Data Library for the Social Sciences. Some of these are of a technical nature on a higher level (relative to the technical details identified in the prototype implementation), while others are more related to the research culture of the potential user community.

Data Privacy One rather technical issue is how to deal with privacy. Survey data is anonymized to ensure the privacy of the participants. When merging and integrating

[23] https://developers.google.com/chart/interactive/docs/reference

data sets, these anonymization efforts can be annulled, as the combination of information allows for the identification of individuals. To avoid such problems, it is necessary to formalize, model and describe implications on the kind and type of data sets another data set with which another data set may be combined and integrated. The modelling of such information is still an open issue [BHBL09]. Approaches regarding fine-grained privacy preferences for Linked Data have been made in [SP11] with the Privacy Preference Ontology (PPO)[24]. Another approach allows users to add access control to RDF documents [HP09].

Data Modelling A similar meta-information that is crucial to a valid scientific analysis is the description of any bias present in the data. Statistical data is based on a sample of a larger population. The initial producers of such a data set are typically aware of any sampling bias they might have in the data (over- and underrepresentation of age groups, geographic location, cultural background, etc.). When publishing a data set in a library, the knowledge of any bias needs to be preserved, which is of particular importance in a scenario where data sets are integrated and joined, as a bias may lead to the wrong conclusions (e.g. joining data on perceived job security and preferences for political parties sampled from different income groups). There is an approach for representing data quality [FH11], but it focuses on managing the quality of data from a database perspective, e.g. monitoring and assessing of data quality and data cleansing. The modelling of biases is not considered in that approach.

Data Integration and Interaction To address the issue of biased data in an adequate manner as well as to enable the (semi-)automatic merging, aggregation and integration of different data sources, it is possibly necessary to further extend existing metadata models like Data Cube and/or complement them with other vocabularies specifically dealing with data transformation. Bias in statistical data or other limitations of the data in use should have standardized support in terms of the vocabulary in metadata models (e.g. descriptive comments are currently supported but lack the advantage of standardized vocabulary for automatic processing). However, increasingly automatic data merging or aggregation needs standardized ways of applying transformation rules to deal with heterogeneous data structure. Here, specific vocabularies or ontologies for data transformation and mathematical functions come into play. Solutions and methods in this regard have been proposed by [Lan12], who suggests using mathematical knowledge of (yet non-Semantic Web conform) ontologies by translating their XML mark-up to RDF.

Community Behaviour A less technical issue is rooted in the scientific culture of the social sciences. The preparation and curation of data sets is a labour-intensive and time-consuming task. The work invested pays off in the production of high-quality papers and the resulting reward in the sense of scientific reputation in the form of citations.

[24] http://vocab.deri.ie/ppo

Publishing a data set itself does not create citations (as there is no established process), and thus no scientific reputation. Therefore, single data sets are rarely published, as data publication might actually bear the risk that other research groups will come up with important findings quicker and thereby exploit the development of the data set without acknowledging the original work. While this behaviour is a cultural issue in the community of the social sciences, a Semantic Data Library that supports the citation of data sets might have an impact on the behaviour. If a data set can be cited and thereby provide the authors with scientific credits, they might be less reluctant to publish their data. Another issue related to citing data sets is the question of granularity. URIs actually allow for the 'deep linking' of individual observations. How to enable fine-grained linkage and referencing with DOIs is still an open question. Overall, the ideas of Linked Science [KdE11] aim at the same direction of making research transparent, accessible and reusable.

3.3 Summary

Considering the fifth block of research questions in Section 1.3, all participants in the expert interviews shared the opinion that there may be a beneficial application for Linked Open Social Science Data **(5a)**. However, based on the results of the interviews conducted in Section 3.1 as well as the requirements and open issues identified in Section 3.2, these are topics of interest that we identified and that we will further investigate.

- **Data Modelling.** The interviews show that a meaningful and fine-grained documentation of data is fundamental for its retrieval and usage. Only when particular detailed information about the data set are available through extensive data modelling can one decide whether a particular data set is useful, relevant and of high quality. The use of Semantic Web standards like RDF enables a semantically rich modelling of complex data. We will address problems in the context of data modelling in Chapter 4.
- **Data Access.** All of the experts raised the point that data is not always available in the fullest extent and, sometimes, not at all. Using sensitive data depends on licence and privacy restrictions. While information on privacy restrictions can be represented in RDF, technologies for handling data access on Linked Data are in early development.
- **Data Retrieval.** Even if data is well documented, it has to be found by users on the web. Our work on the framework of a semantic data library revealed that while information retrieval methods focus traditionally on bibliographical data, research data has barely been addressed yet. However, promising approaches on searching and browsing Linked Data exist that may support the retrieval of such data (for a detailed discussion, see Section 2.1.3).
- **Data Matching.** According to the experts in the interviews, multiple heterogeneous data sets are often compared and analysed together. Hence, they have to be

integrated, merged or matched. Data matching enables the matching of particular data elements between multiple data sets. Thus, it is a method for supporting data integration. For data published as Linked Data, schema and ontology matching approaches can support the data matching task. We will investigate problems in the context of data matching in Chapter 5.

- **Data Interaction.** The interview participants stated that data is sought and integrated for a certain reason, e.g. for performing statistical methods or to visualize it for better understanding and analysis. While a lot of related work can be found for visualizing Linked Data, only a few approaches addressing Linked Data consumption go further (for a detailed discussion, see Section 2.1.3).

Since the modelling and publication of Linked Open Social Science Data is the basis for its use, e.g. for data matching, we will investigate the problems of data modelling (in terms of publishing Linked Open Social Science Data) in Chapter 4 in detail. Afterwards, in Chapter 5, we will focus on developing methods for matching such data. In this thesis, we will not address problems of the other points of interests in detail, since it would go beyond the scope of this work.

4 Publication of Linked Open Social Science Data

Chapter 1 discusses that Semantic Web standards and techniques may have a major influence on a standardized and interlinked publication of data and information on the web. There are several reasons for publishing domain-specific data on the web (for a detailed discussion, see Chapter 1). This chapter focuses on the standardized modelling and publication of Linked Open Social Science Data on the web. In Section 3.3, we identified the topic of 'data modelling' as a fundamental problem that needs further investigation because data published in a structured and semantically expressive way is the basis for its further usage, e.g. in data matching, which is our main objective.

Bechhofer et al. [BAB^{+}10] states, 'Linked Data provides some of the infrastructure that will support the exposure and publication of data and results, but will not alone enable reusable, shared research and the reproducibility required of scientific publication'. Hence, based on the fundamentals of publishing data as LOD on the web given in Section 2.1, we adopt these standardization and modelling techniques on Social Science Data (see Section 1.1 for details). The results of this chapter build the foundation to publish, share and (re-)use Linked Open Social Science Data on the web to the fullest extent. This is the first step regarding the use of such data for scientific research addressed in our use case of 'Analysing Research Data', which was introduced in Section 1.2.

The publication of Linked Open Social Science Data is applied in three examples by which we aim to cover all the relevant aspects. Thus, we are considering processes, data and structures of such data. The examples and their interplay are introduced in Section 4.1. Each of the following three sections focuses on one example in particular. Section 4.2 focuses on publishing the overall research process in the social sciences. Section 4.3 presents a model for publishing person-level research data, which is an essential part of the research process. Section 4.4 focuses on the transformation and linking of a thesaurus in SKOS (Simple Knowledge Information System) format [MB09b]. Using these examples, we enable a complete publication of Linked Open Social Science Data. We summarize the results of this chapter in Section 4.5.

4.1 Levels of Publishing Linked Open Social Science Data

As introduced in Section 1.1, Linked Open Social Science Data includes data and information on different abstraction levels. In its entirety, this composition includes processes, data and structures of the social science domain. Thus, our investigation on publishing Linked Open Social Science Data is applied in three examples that capture these aspects. The examples lead from the top-level perspective on research processes and activities (processes), over a detailed representation of a specific type of information (data) to a thesaurus as an instrument for content indexing of particular information types with expressive terms (structure). By choosing these three examples, we cover all levels of abstraction of Linked Open Social Science Data. Moreover, by (1) extending the processes, (2) developing a new vocabulary for representing data, and (3) providing expressive structure, we are able to complete the composition of Linked Open Social Science Data, since it was not possible to represent all of its components previously (e.g. person-level research data).

Process Section 4.2 focuses on the overall research process in the social sciences. An ontology for representing research entities, communities, activities and results, which is called SWRC (Semantic Web for Research Communities) [SBH^{+}05], is updated in accordance with current Semantic Web standards and links to other vocabularies. It is also extended with regard to missing entities, activities and relationships that are relevant in social science research. The connection and semantic annotation of relationships between different entities in a research process is necessary because research interests and information needs are spread over different entities, activities and data, e.g. a search for relevant literature and survey data or the research outcomes of a particular research topic.

Data Section 4.3 presents an ontology for publishing person-level research data, which typically refers to survey data in the social sciences. This information type is an essential part of the research process. A metadata format for describing and documenting the full life cycle of research data in the social, economic and behavioural sciences, called DDI (Data Documentation Initiative) [All], is transferred to a vocabulary based on Semantic Web standards. Extensive data documentation is necessary for ensuring better and more precise results during information retrieval and for providing users with detailed insights into the content of the research data, which is mostly marked by high complexity and consists of multiple descriptive levels (e.g. study level and variable level). Currently, there is no suitable standard in the Semantic Web for describing research data in such a complex manner. Parts of this section have been published in [BCWZ12, BZWG13].

Structure Section 4.4 focuses on the publication of a thesaurus in the SKOS (Simple Knowledge Information System) format [MB09b]. A domain-specific thesaurus is an instrument that is used for the content indexing of data and documents in a consistent systematic and semantic structure, e.g. for indexing bibliographical data. Thesauri are necessary instruments for information retrieval in large document collections [Kra03b]. Therefore they play an important role for describing and indexing Semantic Web data.

This section contains the work published in [ZS09c, MZS10a, MZS10b, MZJ+11, AEE+12, ZSMM13, KZ13, KREZ14].

4.2 Publishing the Research Processes of the Social Sciences

The exploitation of research activities, entities and outcomes as well as the relationships between them is very important in order to get an overall view on specific topics, research groups or domains. This can be important for administrative purposes, but also for the research interests of scientists, which are not compulsorily restricted to a single information type like literature.

In [SZ09d] and [GHHZ11], we identified the change that digital libraries and archives are undergoing due to the technical possibilities of modern web technologies and their potential support in answering the complex research questions of scientists. Results from several surveys [Pol04, MBD+07, WSB+09, CCGC09, AR10] indicate that harvesting or linking metadata from different sources and making them available for retrieval by applying only a minimum of standardization techniques on data and retrieval features is no longer sufficient for the information needs of users. Scientific users are expecting a tight integration of different types of information (full text, bibliographic references, surveys and other primary data, time-series data, project information, researchers' profiles etc.). This reflects their use of these types of information at different stages and in different combinations throughout the research cycle. At an early stage, for example, a scientist might search for publications and project information, whereas at later stages of his research, he might be looking for research data used in a specific project or study to do secondary analysis or for conferences to present his results. According to [MBD+07], researchers seek to connect pieces of data with other pieces of separately published work. It is important to identify and establish the missing links between such different and heterogeneous types of research information. The connections between research activities, actors and results provide valuable information and support users during their search for answers to their research questions. It enables the desired big picture context [HE09], which is of high relevance especially at the beginning of the research process.

Particularly in the social sciences, on the one hand, data archives documenting empirical data in a very detailed manner are organized at an international level and create dedicated entry points to their holdings; on the other hand, these information and infrastructures are still only minimally connected to the holdings of libraries and information centres [SZ09d, Moc11]. This not only challenges information providers in establishing and organizing collaboration with each other to bring together all resources, but also raises research questions on how to integrate research information at the technical, structural and semantic levels. The complexity involved in supporting the full life cycle of data, including the accompanying documentation – i.e. different versions of questionnaires, the final data set of a survey, the accompanying codebook, sample frequency distributions and summary statistics for variables – creates domain-specific semantics that are at present insufficiently matched to the semantic representations produced for, e.g. research literature. According to [Moc11], current services are far away from satisfying the needs of end-users unless few exceptions.

The emerging paradigm of eScience [Gol07], understood as ‘enhanced’ science, places the focus on creating a holistic infrastructure of hardware, software and (collaboration) networks to support advanced scientific activities that begin with data acquisition and laboratory notes, lead to a new level of scientific publishing (e.g. electronic publishing, open access repositories), and simultaneously make all research results available for retrieval by fellow researchers. Scientific models and methods are therefore needed to uniformly express the structure and semantics of all types of research information. This leads to the problem of data integration, which can basically be observed on two levels. On account of historical and organizational reasons, heterogeneous types of research information are often stored in a physically distributed manner. This situation is complicated by the use of different metadata standards to describe and document the data or information, which is justified in the context of different disciplines. [GCAR06] states that applying Semantic Web technologies to eScience concepts and infrastructures promises support in various tasks like, e.g. ‘the development of controlled vocabularies, flexible metadata modelling, intelligent searching and document discovery,[...] advanced content syndication and publishing, data integration, aggregation and cross linking [...]’. The use of ontologies as a modelling or representation layer can overcome the problem of data integration and can establish the desired missing links between these types of information. But, at the same time, existing infrastructures are retained and metadata standards are still recognized. SWRC (Semantic Web for Research Communities)[1] [SBH$^+$05] is an ontology for modelling entities of research communities, such as persons, organizations, publications (bibliographic metadata) and their relationships.

This section focuses on the extension of the SWRC ontology while paying attention to the demands for modelling the research process in the social sciences. We provide an overview of the SWRC ontology, including its structure and original design issues in Section 4.2.1. Furthermore, we distinguish the SWRC ontology from other approaches, which aim to describe research processes and activities. We present the conceptual requirements needed for an extension of the ontology in order to accomplish the suitability for the domain of the social sciences in Section 4.2.2. In Section 4.2.3 we describe the extension technically in detail including added classes and properties as well as added links to other ontologies.

4.2.1 Analysis of Existing Approaches

There are several approaches for representing scientific research processes. Since we focus on publishing Linked Data, we will introduce concurrent ontologies.

[1] http://ontoware.org/swrc/

SWRC Semantic Web for Research Communities

The SWRC ontology, which is also known as the Semantic Web Research Community Ontology, is introduced in [SBH$^+$05] as an ontology for ‘representing knowledge about researchers, research communities, their publications and activities as well as about their mutual interrelations’. Therefore, key entities and relations of a typical research community can be represented. This comprises a total of 53 concepts and 42 object properties, 20 of which are participating in ten pairs of inverse object properties.

SWRC consists of six top-level concepts that mark the key entities in a research scenario: *Person*, *Publication*, *Event*, *Organization*, *Topic* and *Project*. The different types of research publications, which are subsumed in the publication concept, correspond closely to BibTex publication types. Although the SWRC ontology has been developed before the standardization of SKOS [MB09b], domain-specific topic classifications had been already treated as lightweight ontologies. Therefore, they can be linked with the SWRC ontology through the specialization of the *ResearchTopic* concept with a *Topic* concept of a specific topic hierarchy.

The SWRC ontology has been designed and developed in a modularized way, i.e. additional ontologies, can be imported via `owl:imports` statements, whose definitions become applicable and valid for the importing ontology. According to [SBH$^+$05], it became clear that the SWRC ontology should be able to cover very different use cases. This was why a modularized ontology design became necessary in order to facilitate the reuse of individual ontology modules and to decrease maintenance efforts.

The SWRC ontology is widely used in portals that contain, integrate and manage different information resources, such as the portals of the institute AIFB[2] or the SEKT project[3]. It is also used to publish Linked Data of, e.g. DBLP[4] and Bibsonomy[5]. The Linked Open Vocabularies (LOV) project[6], a project to expose the interlinking between ontologies and the vocabularies used to describe data in the LOD cloud, reveals by which other vocabularies the SWRC ontology is referenced[7]. These are also the SwetoDblp Ontology of Computer Science Publications[8] for exposing DBLP data and SWC – the Semantic Web Conference Ontology[9].

Besides the SWRC ontology, there are additional ontologies and models for representing knowledge related to research activities and entities.

AKT (Advanced Knowledge Technologies) Reference Ontology The AKT (Advanced Knowledge Technologies) Reference Ontology [AKT01] and its sub-ontologies have been

[2] http://www.aifb.kit.edu/

[3] http://www.sekt-project.com/

[4] http://www.informatik.uni-trier.de/~ley/db/

[5] http://www.bibsonomy.org/

[6] http://labs.mondeca.com/dataset/lov/index.html

[7] http://labs.mondeca.com/dataset/lov/details/vocabulary_swrc.html

[8] http://knoesis.wright.edu/library/ontologies/swetodblp/

[9] http://data.semanticweb.org/ns/swc/swc_2009-05-09.html

developed for representing the process of knowledge. They evolved from the AKT project, which aimed to ‘develop and extend a range of technologies providing integrated methods and services for capture, modelling, publishing, reuse and management of knowledge’.

VIVO ontology The VIVO ontology [MCA$^+$11] focuses on researchers and networks of researchers. It enables representing researchers’ teaching activities, their expertise, their research and which service activities they provide. The ontology arose from the VIVO project, which enables the exploitation and discovery of researchers across an institutional context by providing an open source application, which can be populated with Semantic Web-compliant data.

CERIF (Common European Research Information Format) The CERIF (Common European Research Information Format)[10] standard is the most similar approach to the SWRC ontology. It enables the management and exchange of research data and information, and is a European Union recommendation that was originally developed with the support of the European Commission. CERIF provides a model for describing research domains, their entities and activities as well as how they change over time. In contrast to SWRC, CERIF is currently not available in a Semantic Web standard format, but activities regarding the application of CERIF to Linked Data have recently begun.

Other Approaches There are other approaches for modelling parts of the research process. Since they do not cover the overall research process as discussed in Section 4.2 and do not concentrate on research data, they can be seen as a bridge between both representation layers. Rijgersberg et al. [RTM08] propose a quantitative research ontology in eScience. While there is no formal representation of the ontology, it aims to model concepts required in a quantitative research process such as units of measurements, quantities and dimensions. It is motivated by the lack of such vocabularies in information systems in order to enable advanced services for researchers. Bechhofer et al. [BAB$^+$10] introduces research objects for sharing, publishing and reproducing research and its results. Research objects are defined as an aggregation layer on top of exposed Linked Data resources in order to conduct scientific research in context. An abstract vocabulary for representing and connecting research entities (e.g. researchers, methods, hypotheses, conclusions) and activities (e.g. participations, confirmations, falsifications) has been presented with the Linked Science Core Vocabulary [BKK11].

In contrast to the other presented approaches, the SWRC ontology provides a comprehensive and balanced representation of the research process. Other approaches set a particular focus (AKT, VIVO, Linked Science) or are still not available in RDF (CERIF). Furthermore, SWRC is a mature ontology that is already established among other popular ontologies. This decreases the effort for developing a new and separate ontology. At the same time, data integration is eased and interoperability with other research-associated Semantic Web data can be ensured.

[10] http://www.eurocris.org/Index.php?page=CERIFintroduction&t=1

4.2.2 Design Considerations for an Extension

While the SWRC ontology is a mature and established ontology among other Semantic Web ontologies, there are a few aspects that have to be considered for applying it to the domain of the social sciences. Since its introduction in 2005, there have been further developments regarding standard vocabularies for describing research-relevant types of information that have to be considered as well, e.g. scientific literature. All requirements presented in this section have been chosen while following the original ontology design decisions made by the authors and engineers [SBH$^+$05].

Essential for social science research is research data, which is collected, e.g. in studies or surveys or held by statistical agencies. As research data marks also a relevant entity for other research domains, it seems reasonable to introduce a new top-level concept called *Dataset* in order to remain as generic and applicable as possible.

In order to ensure interoperability to include other established ontologies, the top-level concepts of SWRC can be linked to the corresponding and suitable classes, e.g. via `owl:subClassOf` or `owl:equivalentClass` statements. The definition of equivalent classes should be considered if the linked ontology is a major and acknowledged standard among other Semantic Web ontologies. This can be considered for links to the Dublin Core Metadata Initiative (DCMI) Terms [Ini], to the Friend-of-a-Friend (FOAF) vocabulary [Bri10] or the Simple Knowledge Organization System (SKOS) format [MB09b]. However, the definition of the *Topic* concept of the SWRC as a subclass of a SKOS concept is especially reasonable because SKOS gained tremendous popularity as a format for representing thesauri and classification hierarchies in recent years. When the *Topic* concept is treated exactly like SKOS concepts, the reuse of existing vocabularies already published in the SKOS format is possible without any constraints. For modelling research data, links to recently established ontologies like the RDF Data Cube vocabulary [CRT14] for modelling statistical data should be considered as should links to the DDI-RDF ontology presented in Section 4.3, which is based on a model for describing research data in the economic, social and behavioural sciences.

When analysing the datatype properties and object properties of the SWRC ontology, it becomes clear that properties relating to the new top-level concept of *Dataset* have to be added. This includes properties that describe a connection to other concepts of the ontology, e.g. `swrc:outcomeData`. Defined domains and ranges of existing properties have to be adjusted as well.

For some existing properties, it can be considered whether a specialization into sub-properties supports a precise modelling of specific aspects. As an example, the property `swrc:identifier` describes in a very generic way all kinds of identifiers. But a specialization in, e.g. `swrc:isbn`, `swrc:doi` or `swrc:urn`, allows a concrete representation of a specific identifier, which is very important especially when thinking of the different types of research information that can be represented using the SWRC ontology. Thus, the usage of a specific identifier can be ensured and modelled precisely.

Additional properties can be added that relate to the overall research process and support the representation of relationships between single entities, e.g. the funding of projects by organizations or the production of research data during research projects. A design principle in this regard has been to add classes and properties that are covered in only smaller and lesser-known vocabularies, but that are of relevance for the scope of the SWRC ontology. This has been done in order to avoid an extensive use of single classes and properties of a wide range of different and small vocabularies because this leads to an inconsistent and mixed modelling of resources.

4.2.3 Extending the Semantic Web for Research Communities (SWRC) Ontology

In this section, we describe the extension of the SWRC ontology in detail. The focus is on extending and completing relations and activities that affect more than one entity, and in terms of SWRC, more than one top concept. This decision has been made because single entities should be modelled as per the established and adequate standards for each. SWRC models the relationships and activities between them.

In order to include research data, which is a relevant entity in social science research, the new top concept `swrc:Dataset` has been defined (see Table 4.1). It is defined as an equivalent class to `qb:DataSet` of the RDF Data Cube vocabulary [CRT14].

Class	Description
swrc:Dataset	The class Dataset represents a research data set. It is defined as owl:equivalentClass to qb:DataSet of the RDF Data Cube vocabulary.

Table 4.1: New class swrc:Dataset.

Table 4.2 provides an overview of additions to the existing classes of the SWRC ontology. Six of the classes of the SWRC have been set in relation to other classes of existing vocabularies. The classes `swrc:Document`, `swrc:Organization` and `swrc:Person` have been defined as equivalent classes to the corresponding classes of FOAF and Dublin Core in order to expose their similar definition, scope and usage. The class `swrc:Topic` has been defined as a subclass of a `skos:Concept` because their use and scope is exactly the same. This allows a full integration of data in the SKOS format into the SWRC ontology. Finally, the class `swrc:Project` has been defined as an equivalent class to `foaf:Project`.

Various adjustments have been made to properties by extending the domains and ranges to the new concept `swrc:Dataset`. Furthermore, inconsistencies regarding inverse properties have been fixed. The datatype properties `swrc:journal` and `swrc:series` have been reorganized as object properties, because this information is mostly modelled as own classes[11].

[11] This issue has been raised by members of the Pedantic Web Group at http://groups.google.com/group/pedantic-web/browse_thread/thread/385353b4fcd9452e

Original Class or Property	Modification
swrc:Document	Has been defined as owl:equivalentClass to foaf:Document.
swrc:Organization	Has been defined as owl:equivalentClass to foaf:Agent and dc:Agent. It is not referenced directly to foaf:Organization in order to keep the same conceptual level with dc:Agent.
swrc:Person	Has been defined as owl:equivalentClass to foaf:Agent and dc:Agent. It is not referenced directly to foaf:Person in order to maintain the same conceptual level as dc:Agent.
swrc:Project	Has been defined as owl:equivalentClass to foaf:Project.
swrc:Topic	Has been defined as rdfs:subClass to skos:Concept.
swrc:journal and swrc:series	Have been moved to object properties due to the major use as such.
swrc:cite	Changed to swrc:cites due to naming conventions. Domain and range have been expanded to the new class swrc:Dataset.
swrc:citiedBy	Range and domain have been expanded to swrc:Dataset.
swrc:head and swrc:headOf	The definition of the inverse properties has been repaired.
swrc:isAbout and others	Diverse adjustments in domains and ranges and inverse properties

Table 4.2: Modifications on existing classes and properties.

Table 4.3 depicts additional properties that have been added to SWRC. The properties `swrc:startHour`, `swrc:endHour` and `swrc:fee` allow a detailed description of events. Persons and organizations can be additionally related to events with the properties `swrc:organizes` and `swrc:hosts` and their inverse properties `swrc:organizedBy` and `swrc:hostedBy`. Subclasses for different kinds of persistent identifiers have been added to `swrc:identifier` in order to enable a precise representation of them and for their clear distinction. In addition, two subclasses, `swrc:acronym` and `swrc:subtitle`, have been added to `swrc:title`; these are not only relevant for publications, but also for the titles of projects or organizations. Funding information can be represented by `swrc:funds` and `swrc:fundedBy`. Alongside the new class `swrc:Dataset`, additional properties have been defined that set this resource in relation to other top concepts, e.g. `swrc:datasetInfo`, `swrc:describedDataset`, `swrc:outcomeDataset`, `swrc:hasPrimaryResearcher` and `swrc:hasDataCollector`.

Property	Description
swrc:startHour and swrc:endHour	For some events it is important not only to provide the start and end data, but also the start and end hours.
swrc:fee	This property has been added, because the fee of an event is a relevant bit of information.
swrc:doi, swrc:handle, swrc:isbn, swrc:issn and swrc:urn	Because persistent identifiers are of a high relevance for resources, subclasses for describing DOI, Handles, URN as well as ISBN and ISSN numbers have been added to swrc:identifier in order to provide precise distinctions between them.
swrc:fundedBy and swrc:funds	Have been defined to expose funding information and to distinguish them from swrc:financedBy and swrc:finances.
swrc:producer and swrc:producedBy	Have been defined for representing product and data set production information.
swrc:datasetInfo	Has been defined in orientation to swrc:projectInfo for the inclusion of swrc:Dataset.
swrc:describedDataset	Has been defined in orientation to swrc: describesProject for the inclusion of swrc:Dataset.
swrc:outcomeDataset	Has been defined in orientation to swrc: outcomeProduct for the inclusion of swrc:Dataset.
swrc:graduatedIn	Has been added for presenting the graduation of a person.
swrc:acronym	Has been added as rdfs:subProperty of swrc:title.
swrc:subtitle	Has been added as rdfs:subProperty of swrc:title.
swrc:hostedBy and swrc:hosts	Have been added for specifying relations from and to an event.
swrc:organizedBy and swrc:organizes	Have been added for specifying relations from and to an event.
swrc:hasPrimary Researcher	Describes a relationship between a data set and a person.
swrc:hasDataCollector	Describes a relationship between a data set and a person or organization.

Table 4.3: New properties of the SWRC ontology.

Due to the extension with a fundamental new entity `swrc:Dataset`, we have decided to provide the updated ontology as Version 0.8. Furthermore, we have decided to move the namespace to a PURL[12] web address because of maintenance and availability reasons. Until the PURL registration is complete, the updated and extended SWRC ontology is

[12] http://purl.org/

available at `http://lod.gesis.org/lodpilot/swrc/swrc_v0.8.owl`. The extensions of the ontology can also be found in the Appendix B.

4.3 Publishing Person-level Research Data Using Data Documentation Initiative RDF (DDI-RDF)

The modelling the research process as presented in Section 4.2 considers top-level activities and the relationships of scientific research in the first place. However, all entities of a research process typically hold various elements and relations of a fine-grained level of detail within. The Data Documentation Initiative (DDI) [All] is an international standard for the documentation and management of data from the social, behavioural, and economic sciences. The DDI metadata specification supports the entire research data life cycle. Such an extensive description of research data is not only necessary during its collection, analysis and archiving, but also for information seeking. Since research data is an important information source for social science research, it is reasonable that it is connected with other data sources for providing integrated information and, hence, should be represented as Linked Data as well. This can enrich DDI data with context information from other data sources.

DDI focuses on micro data, which describes data on a fine-grained level, e.g. survey data from single participants of a study. But aggregated data can also be described. Aggregated data is derived from micro data by statistics on groups, or aggregates such as counts, means or frequencies. So far, the DDI data model is expressed in XML Schema. This section focuses on the development of DDI-RDF, a RDFS and OWL ontology for a basic subset of DDI. DDI-RDF enables the representation of DDI data for discovery and dissemination purposes in accordance with Semantic Web standards. Thus, the DDI model is opened up to the Linked Data community. When DDI data and metadata is published as Linked Data, it can be processed by Semantic Web applications and can be linked with other data sets from the web, which can provide additional context information and metadata. Furthermore, querying multiple, distributed and merged DDI instances is possible. Currently, there is no ontology with a comparable level of detail available in the Semantic Web that represents complex entities and relations regarding the complete life cycle of research data in the way that DDI does. For social science research, the research data in such a fine-grained level can be queried together with other data sets and can easily be combined with, e.g. relevant publications and statistical data.

This section summarizes work that has been presented in [BCWZ12, BZWG13]. Work on DDI-RDF[13] has been started at the workshop ‘Semantic Statistics for Social, Behavioural,

[13] The development of DDI-RDF includes the contributions of (in alphabetical order) Archana Bidargaddi (NSD - Norwegian Social Science Data Services), Thomas Bosch (GESIS – Leibniz Institute for the Social Sciences), Sarven Capadisli (Bern University of Applied Sciences), Franck Cotton (INSEE - Institut National de la Statistique et des Études Économiques), Richard Cyganiak (DERI, Digital Enterprise Research Institute), Daniel Gilman (BLS - Bureau of Labor Statistics), Arofan Gregory (ODaF - Open Data Foundation and DDI Alliance Technical Implementation Committee (TIC)), Rob

and Economic Sciences: Leveraging the DDI Model for the Linked Data Web' at Schloss Dagstuhl - Leibniz Center for Informatics, Germany in September 2011 and has been continued at the follow-up workshop in the course of the 3rd Annual European DDI Users Group Meeting (EDDI11) in Gothenburg, Sweden. We present metadata formats and vocabularies for the representation of research or statistical data in Section 4.3.1. In Section 4.3.2, we present an overview on the DDI metadata standard. We discuss design considerations and goals for DDI as Linked Data in Section 4.3.3. We present the modelling and technical implementation in Section 4.3.4. The conversion process from DDI-XML to DDI-RDF is presented in Section 4.3.5. Relationships to other vocabularies that complement DDI-RDF are described in Section 4.3.6. Finally, in Section 4.3.7, we discuss the current limitations of the presented approach.

4.3.1 Metadata Formats and Vocabularies for Representing Research Data

Beyond the Semantic Web, there are already several metadata standards for representing complex statistical information like SDMX (Statistical Data and Metadata Exchange) [SDM02] as well as for related information like ISO 19115 [ISO03] for geographic information or PREMIS [oC] for preservation purposes. The metadata registry ISO 11179 [ISO04] marks a standard for the modelling of metadata, e.g. reference models and for their registry. Elements are often used as top-level components, while other standards and concrete implementations are derived. But besides standards in XML for describing and documenting such complex metadata models, there are still only a few adequate RDF-based vocabularies. DDI-RDF has a clearly defined focus on describing micro data, which has not been covered to this extent by other established vocabularies yet (see SDMX, SCOVO and Data Cube below). Therefore, it applies well alongside other metadata standards on the web and can clearly be distinguished. Connection points to classes or properties of other vocabularies ensure equivalent or more detailed possibilities for describing entities or relationships.

Grim (Tilburg University), Marcel Hebing (SOEP - German Socio-Economic Panel Study), Larry Hoyle (University of Kansas), Yves Jaques (FAO of the UN), Jannik Jensen (DDA - Danish Data Archive), Benedikt Kämpgen (Karlsruhe Institute of Technology), Stefan Kramer (CISER - Cornell Institute for Social and Economic Research), Amber Leahey (Scholars Portal Project - University of Toronto), Olof Olsson (SND - Swedish National Data Service), Heiko Paulheim (University of Mannheim), Abdul Rahim (Metadata Technologies Inc.), John Shepherdson (UK Data Archive), Dan Smith (Algenta Technologies Inc.), Humphrey Southall (Department of Geography, UK Portsmouth University), Wendy Thomas (MPC - Minnesota Population Center and DDI Alliance TIC), Johanna Vompras (University Bielefeld Library), Joachim Wackerow (GESIS – Leibniz Institute for the Social Sciences and DDI Alliance TIC), Benjamin Zapilko (GESIS - Leibniz Institute for the Social Sciences), and Matthäus Zloch (GESIS - Leibniz Institute for the Social Sciences).

SDMX (Statistical Data and Metadata Exchange)

SDMX [SDM02] was established in 2002 by key players in the field of statistical data, such as the World Bank[14], the IMF[15] and the European Central Bank[16]. The focus has been set on the ability to enable automatic machine-to-machine exchange of data, which requires a self-expressive and self-descriptive metadata model. SDMX defines representations of statistical data and the respective metadata annotations not only for single data items but also for full data sets. The SDMX information model is based on labelled concepts to which dimensions and attributes are assigned. Dimensions can be grouped into keys using code lists for available realizations. Data Structure Definitions (DSD) organize these components according to a specific topic or data source in a well-defined structure. In this way, multidimensional statistical data can be represented by the SDMX information model (see the paragraph on the RDF Data Cube vocabulary below). Parts of SDMX are reused in the definition of the Data Cube metadata model. There exists a SDMX-RDF [CDR10] representation, which features only very basic concepts of SDMX. SDMX-RDF is the predecessor of the RDF Data Cube vocabulary [CRT14], which is presented below.

SCOVO (Statistical Core Vocabulary)

SCOVO [HHR$^+$09] is an RDFS-based, lightweight vocabulary for representing statistical data. SCOVO provides a basic core of classes and properties for representing data sets, dimensions and statistical items. For additional elements, extensions of other RDF vocabularies are fostered on both, the schema and the instance level. Another important design issue for SCOVO has been – in line with SDMX features – the ability to handle as many dimensions as necessary (supporting a multidimensional model). Compared to SDMX's focus on generic and efficient data exchange, SCOVO has weaknesses in this respect. Being part of the web of Data and complying with RDF standards as message format enables both self-descriptive data items and generic data exchange. Item classes of SCOVO describe single observations or events. [VLH$^+$10] extends the SCOVO vocabulary by SCOVOLink, an 'ontology that enables us to state the link between the data and the described entities explicit by grounding the statistical data in existing vocabularies and appropriate mathematical functions' [VLH$^+$10]. The main intention underlying this approach is to add more specific and meaningful information to statistical data sets as well as to enable the use of such data by, e.g. mathematical functions.

Listing 4.1 presents an excerpt of a data set represented in SCOVO. The excerpt includes one single observation of the National Radioactivity Statistics[17] measured in Japan from March 2011 to March 2012. The measure is represented as a simple `rdf:value`. Two

[14] http://www.worldbank.org/

[15] http://www.imf.org/

[16] https://www.ecb.europa.eu/

[17] http://www.kanzaki.com/works/2011/stat/ra/

dimensions are included, `ev:place` and `ev:time`. The property `scv:dataset` of SCOVO exposes the relationship of this observation to the data set `<ra/set/moe>`.

Listing 4.1: Example of an statistical observation in SCOVO format.

```
<ra/20110315/p02/t18>
  rdf:value "0.022"^^ms:microsv ;
  ev:place <http://sws.geonames.org/2130654/> ;
  ev:time <dim/d/20110315T18PT1H> ;
  scv:dataset <ra/set/moe> .
```

RDF Data Cube Vocabulary

An established RDF metadata vocabulary, which seems to be very similar to DDI-RDF at first glance, is the RDF Data Cube vocabulary [CRT14]. It has evolved from the combination of SDMX-RDF with SCOVO [CFG^{+}10] in order to overcome the limitations of both approaches. The vocabulary maps the SDMX information model to an ontology and is therefore compatible with the cube model that underlies SDMX. It can be used to represent aggregated data, such as multi-dimensional tables. A data set presented using the Data Cube vocabulary consists of a set of values organized along a group of dimensions, which is comparable to the representation of data in an OLAP cube [KOH12].

The vocabulary is a recommendation of the W3C since January 2014. It is already accepted and applied for modelling statistical data due to its various advantages. In particular, Data Cube (shortened as QB) incorporates all the features of SCOVO, but resolves some of SCOVO's limitations. It uses relevant parts of the SDMX information model. Due to the organization in a multidimensional cube, slices can be cut through the cube to get cross-sectional and low-dimensional data views. QB also has components like *dimension*, *measure* and *attribute*, which are all set up in a *Data Structure Definition* class. This has also been adopted from SDMX. The dimensions describe a specific space in which a single observation is measured (e.g. a temporal dimension with various observation points like different years). A measure describes the overall phenomenon that is being observed or represented, e.g. the unemployment rate. Furthermore, dimensions or statistical concepts can be defined and assigned to a SKOS concept class and even a complete `skos:ConceptScheme`, which allows, e.g. the inclusion of code lists. Details on SKOS will be given in Section 4.4.1. Metadata to all data sets or specific items can be added using Dublin Core terms [Ini] or by using the attribute component. The actual values are organized as single observations, which are enriched by dimensions, measures and attributes. According to [CRT14], Data Cube is unique in its features compared to SCOVO.

Listing 4.2 presents an excerpt that is respectively an observation of Eurostat[18]. The Linked Data representation has been implemented via a RDF wrapper by Ontology

[18] http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/

Central[19]. The data set contains statistics on the *Population at January 1st, by sex and age, at Territorial level 2*[20] *(source: OECD)*[21]. Each observation is organized in a blank node. Besides the dimensions, e.g. `sex`, `age`, `geo`, `unit` and `dcterms:date`, it can be observed that there is again the relation to the data set itself via the property `qb:dataset`. The property `sdmx-measure:obsValue` holds the value of the statistical observation as an object.

Listing 4.2: Example observation of a data set of Eurostat represented in QB.

```
_:node171gd7ng0x1583
  a qb:Observation ;
  qb:dataSet <http://estatwrap.ontologycentral.com/id/demo_r_d2janoecd#ds>
    ;
   :sex <http://estatwrap.ontologycentral.com/dic/sex#F> ;
   :age <http://estatwrap.ontologycentral.com/dic/age#Y15-64> ;
   :unit <http://estatwrap.ontologycentral.com/dic/unit#PERS> ;
   :geo <http://estatwrap.ontologycentral.com/dic/geo#US19> ;
  dcterms:date "2006" ;
  sdmx-measure:obsValue "970870" .
```

This data set is defined in Listing 4.3, where a reference to the Data Structure Definition can be identified in the property `qb:structure`. Then, this DSD describes the complete structure of the data set, including used code lists (see an excerpt in Listing 4.4). The dimensions referred to in the data set in Listing 4.2 are resolved in code lists inside the Data Structure Definition. This is clearly depicted in the last triple of Listing 4.4 by the property `qb:codeList` with the object `<http://ontologycentral.com/2009/01/eurostat/ns#cl_age>`.

Listing 4.3: Reference from the data set to the corresponding Data Structure Definition.

```
<http://estatwrap.ontologycentral.com/id/demo_r_d2janoecd#ds>

 a qb:DataSet ;
 rdfs:label "Population at 1st January by sex and age, at Territorial
    level 2 (source: OECD)" ;
 rdfs:comment "Source: Eurostat (http://epp.eurostat.ec.europa.eu/) via
    Linked Eurostat (http://estatwrap.ontologycentral.com/)." ;
 foaf:page <http://estatwrap.ontologycentral.com/> ;
 qb:structure
    <http://estatwrap.ontologycentral.com/../dsd/demo_r_d2janoecd#dsd> .
```

Listing 4.4: Excerpt of the corresponding Data Structure Definition.

```
<http://estatwrap.ontologycentral.com/../dsd/demo_r_d2janoecd#dsd>

  a qb:DataStructureDefinition ;
  foaf:page <unknown:namespace> ;
```

[19] http://estatwrap.ontologycentral.com/

[20] Eurostat uses three territorial levels, organized by the NUTS nomenclature system. This system is described in detail in section 4.1 Statistical Linked Data.

[21] http://estatwrap.ontologycentral.com/page/demo_r_d2janoecd

```
  qb:component _:node171gdac8dx1 ;
  qb:component _:node171gdac8dx2 ;
  qb:component _:node171gdac8dx3 ;
  qb:component _:node171gdac8dx5 .

_:node171gdac8dx1 qb:measure sdmx-measure:obsValue .

sdmx-measure:obsValue a rdfs:Property .

_:node171gdac8dx2 qb:dimension dcterms:date .

dcterms:date
  a rdfs:Property ;
  rdfs:range <http://www.w3.org/2001/XMLSchema#date> .

_:node171gdac8dx3 qb:dimension
    <http://ontologycentral.com/2009/01/eurostat/ns#geo> .

<http://ontologycentral.com/2009/01/eurostat/ns#geo>
  a rdfs:Property ;
  rdfs:range <http://rdfdata.eionet.europa.eu/ramon/ontology/NUTSRegion> .

_:node171gdac8dx5 qb:dimension
    <http://ontologycentral.com/2009/01/eurostat/ns#age> .

<http://ontologycentral.com/2009/01/eurostat/ns#age>
  a rdfs:Property ;
  qb:codeList <http://ontologycentral.com/2009/01/eurostat/ns#cl_age> .

 [...]
```

4.3.2 Data Documentation Initiative (DDI) Metadata Model

The following subsection summarized work presented in [BCWZ12, BZWG13].

DDI supports technological and semantic interoperability in enabling and promoting international and interdisciplinary access to and use of research data. Structured high-quality metadata enable secondary analysis without the need to contact the primary researcher who collected the data. Comprehensive metadata along the whole data life cycle are crucial for the replication of analysis results. DDI enables the reuse of metadata of existing studies (e.g. questions, variables) for designing new studies, an important ability for repeated surveys and for comparison purposes. Public accessible metadata of good quality are important for finding the right data. This is especially the case if access to micro data is restricted because a disclosure risk of the observed people exists. DDI is currently specified in XML Schema, organized in multiple modules corresponding to the individual stages of the data life cycle and comprehends over 800 elements.

The DDI metadata specification supports the entire research data life cycle. DDI metadata accompanies and enables data conceptualization, collection, processing, distribution, discovery, analysis, re-purposing, and archiving. Data documentation is seen as a process that begins early on in a project. The metadata could then be reused along the data life cycle. Such practices incorporate documenting as part of the research method according to [JH04]. Items like questionnaires, statistical command files and web documentation can be generated easily if metadata creation is begun at the design stage of a study (e.g. survey) in a well-defined and structured way.

But DDI does not represent another model for statistical data. Rather, it formalizes state-of-the-art concepts and common practice in this domain. Its strength is in micro data, which are opposed to aggregated data, also known as macro data (likewise covered by DDI). Aggregated data provides a summarized version of this information in the form of statistics like means or frequencies. A specific DDI module uses DCMI Metadata Terms [Ini] for references or as descriptions of a particular set of metadata. In DDI, Dublin Core is not used as the primary citation mechanism – this module is included to support applications that can process the Dublin Core XML, but not DDI. This module is used wherever citations are permitted within DDI. DDI is aligned with other metadata standards as well, such as SDMX [SDM02] (time-series data) for exchanging aggregate data, ISO/IEC 11179 (metadata registry) for building data registries like question, variable and concept banks [ISO04], and ISO 19115 (geographic standard) for supporting GIS (geographic information system) users [ISO03].

According to [All], multiple institutions can be involved in the data life cycle, which is an interactive process with multiple feedback loops. Figure 4.1 displays the data life cycle, which is described in greater detail on the DDI Alliance website[22].

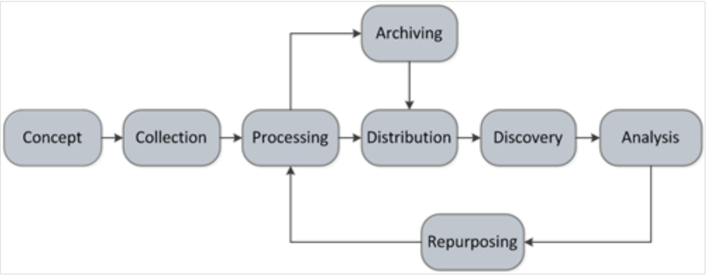


Figure 4.1: DDI Data Life Cycle [BCWZ12].

A large community of data professionals, including data producers (e.g. of large, academic international surveys), data archivists, data managers in national statistical agencies and other official data-producing agencies, and international organizations use the DDI metadata standard. Academic users include the UK Data Archive[23] at the University of Essex, the DataVerse Network[24] at the Harvard-MIT Data Center, and the Inter-University Consortium for Political and Social Research (ICPSR)[25] at the University of Michigan. Official data producers in more than 50 countries include the Australian

[22] http://www.ddialliance.org/
[23] http://www.dataarchive.ac.uk/
[24] http://thedata.org/
[25] http://www.icpsr.umich.edu

Bureau of Statistics (ABS)[26] and many national statistical institutes of the Accelerated Data Program for developing countries[27]. Examples of international organizations are UNICEF, for the Multiple Indicator Cluster Surveys (MICS)[28], The World Bank[29], and The Global Fund to Fight AIDS, Tuberculosis and Malaria[30]. Some of these organizations have already started activities involving LOD like the World Bank[31], the UK Data Archive, which aims to apply the Humanities and Social Science Electronic Thesaurus (HASSET)[32] to the SKOS format and GESIS, whose Thesaurus for the Social Sciences has been converted to the SKOS format (see Section 4.4 for details).

Ongoing work focuses on the early phases of survey design and data collection as well as on other data sources like register data. DDI has its strength in the domain of social, economic and behavioural data. The next major version of DDI will incorporate the results of this work. It will be opened to other data sources and to data of other disciplines.

4.3.3 DDI as Linked Data

There are several requirements and design goals for converting the DDI XML metadata standard to Linked Data. The Linked Data community benefits from a DDI representation, because there is currently no such ontology with a comparable level of detail for representing complex entities and relations regarding the complete life cycle of research data. But the publication of research data on the web has become popular and important in various domains besides the social sciences; hence, a valuable role has been played by the introduction of DDI-RDF. The benefits for the DDI community are being able to publish DDI data as well as metadata in the LOD cloud as RDF data. As a consequence, DDI instances can be processed by Semantic Web applications without supporting the DDI-XML Schemas' data structures. After publishing publicly available structured data, DDI data and metadata may be linked with other data sources of multiple topical domains. With the possibilities of Semantic Web technologies, querying multiple, distributed and merged DDI instances is possible.

For some parts of DDI, there already exist possible standard vocabularies in the Semantic Web, but most of them cover aspects insufficiently. For example, the different authorities that participate in a data life cycle, like principal investigators, collectors, distributors or producers cannot be represented by the Dublin Core standard [Ini] in a way that their function (e.g. a data archive) can be clearly be understood and distinguished. In Dublin Core, there are no producers and distributors, but there are contributors, publishers

[26] http://www.abs.gov.au/
[27] http://www.ihsn.org/adp
[28] http://www.childinfo.org/mics3_surveys.html
[29] http://data.worldbank.org/
[30] http://www.theglobalfund.org/
[31] http://worldbank.270a.info/.html
[32] http://www.data-archive.ac.uk/about/projects/skos-hasset

and rightsholders. These terms do not fit adequately with the requirements of the DDI authorities.

Nevertheless, classes and properties of existing vocabularies should be used for a DDI-RDF representation when possible because it is recommended and given as a best practice in Linked Data guidelines due to interoperability issues [HB11]. The DCMI Metadata Terms [Ini] have been applied for representing basic information about publishing objects on the web, for citation purposes and for `dcterms:hasPart` relationships. For representing concepts that are organized similar to thesauri and classification systems, classes and properties of Simple Knowledge Organization System (SKOS) [MB09b] are used. Furthermore, some aspects of DDI-RDF are already similarly represented in other metadata vocabularies, e.g. data management and documentation. The Vocabulary of Interlinked Datasets (VoID) [CZAH11] represents relationships between multiple data sets, while the vocabularies on provenance [HZ12, LSM13] provide the possibility to describe information on ownerships and terms of use. In this context, a study can be seen as a data-producing process and a logical data set as its output artefact.

A major design goal for DDI-RDF has been that not every element and not every relationship of the complex data documentation from the original DDI should be represented as Linked Data. Instead, DDI-RDF aims to enable the dissemination of research data in the Semantic Web especially for discovery purposes, i.e. for finding and exploring data as well as sharing and interlinking it with other data sets or researchers.

4.3.4 Model of the DDI-RDF Discovery Vocabulary (Disco)

In this section, we present the development process from the DDI-XML metadata standard to the DDI-RDF ontology for exposing DDI data according to Semantic Web standards. It summarizes the previous publications [BCWZ12, BZWG13] and the technical specification available at `http://rdf-vocabulary.ddialliance.org/discovery.html`. In accordance to its namespace prefix, the vocabulary is also shortened as Disco.

Figure 4.2 visualizes the conceptual model, including the DDI elements (subset of the whole DDI model) that are considered most relevant for disseminating DDI data on the web. This model is based on XML Schema describing the DDI domain data model with extensions that partly use existing vocabularies and partly reside in a new DDI vocabulary. Only relations between exactly two DDI elements and not between one DDI element and an instance of an XML Schema datatype are displayed in the figure in order to reduce the complexity of the overall conceptual model. The three components of the DDI conceptual model – *Study*, *Variable* and *LogicalDataSet* – are seen as the most important parts of the data model.

Some features of DDI can be addressed through other vocabularies, such as: describing metadata for citation purposes using Dublin Core [Ini], describing aggregated data like multi-dimensional tables using the RDF Data Cube Vocabulary [CRT14], and describing code lists, category schemes, mappings between them, and concepts like topics using SKOS [MB09b]. Object and datatype properties from Dublin Core and the SKOS have been used

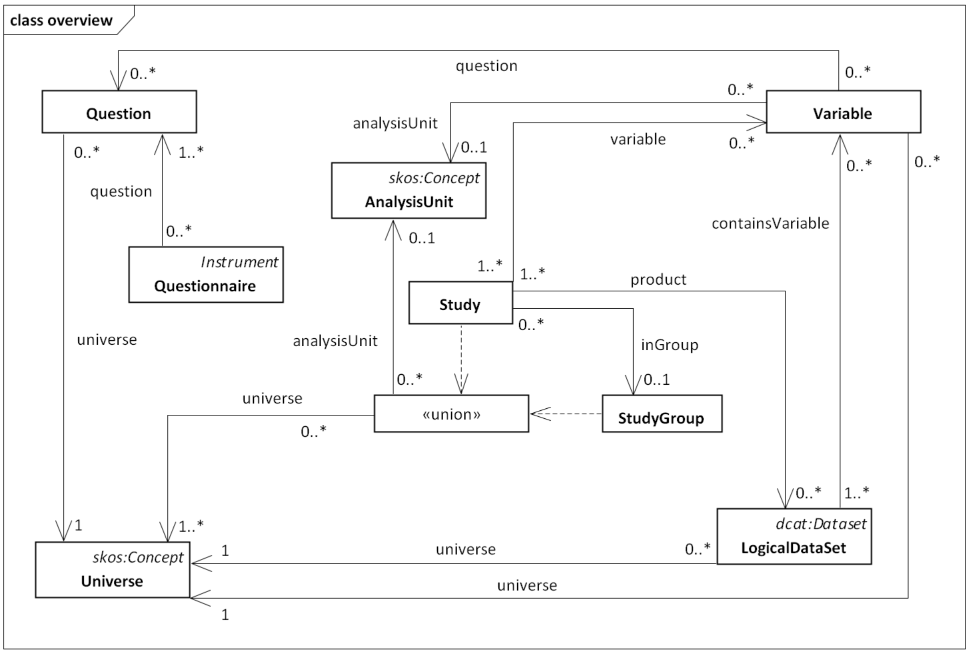


Figure 4.2: Overview of DDI-RDF model.

to represent various relations between DDI elements (e.g. `dcterms:hasPart`), between classes defined in other namespaces (e.g. `skos:inScheme` or `skos:hasTopConcept`), and between DDI elements and XML Schema datatypes (e.g. `dcterms:identifier` or `skos:definition`). Overall, two object properties and 13 datatype properties are reused from the Dublin Core Metadata Terms [Ini].

The class *Study* supports the stages of the full data life cycle in a modular style. This does not comprehend groups of studies (like repeated annual surveys). The key criteria for a study are: a single conceptual model (e.g. survey research concept), a single instrument (e.g. questionnaire) made up of one or more parts (ex. employer survey, worker survey), and a single logical data structure of the initial raw data (multiple data files can be created from this, such as a public use micro data file or aggregate data files) [All09]. The Dublin Core datatype properties `dcterms:abstract`, `dcterms:title`, and `dcterms:identifier` are used to describe studies.

The classes *Concept*, *Universe*, and *Coverage* define a *Study.* SKOS is used to define the class *Concept*, which is a unit of knowledge created by a unique combination of characteristics [ISO00]. In the context of statistical (meta)data, concepts are abstract summaries, general notions, knowledge of a whole set of behaviours, attitudes or characteristics that are seen as having something in common. Concepts may be associated with variables and questions. A *ConceptScheme*, also defined within the SKOS namespace, is a set of metadata describing statistical concepts. *Universe* is the total membership or population of a defined class of people, objects or events. There are two types of population – target population and survey population. A target population is the population outlined in the survey objects regarding which information is to be sought, while a survey population (also known as the coverage of the survey) is the population from which information can be obtained in the survey. *Coverage* comprehends the key features of the scope of the data (e.g. geographic product occupation). The *Coverage* has the datatype properties `dcterms:subject` and `dcterms:temporal` and the object property `dcterms:spatial` pointing to `dcterms:Location`.

The data for the study are collected by an instrument. The purpose of an *Instrument*, i.e. an interview, a questionnaire or another entity used as a means of data collection, is, in the case of a survey, to record the flow of a questionnaire, its use of questions, and additional component parts [All09]. A questionnaire contains a flow of questions. A *Question* is designed to gather information about a subject, or a sequence of subjects from a respondent. The *Variable* is a characteristic of a unit being observed. A variable might be the answer of a question, have an administrative source, or be derived from other variables. Two DCMI datatype properties `dcterms:identifier` and `dcterms:description` have the same domain class *Variable*. The *Representation* of a variable is the combination of a value domain, datatype, and, if necessary, a unit of measure or a character set [ISO04]. *Representation* is one of a set of values to which a numerical measure or a category from a classification can be assigned (e.g. income, age, and sex: male coded as 1). *DataElement* encompasses study-independent, reusable parts of variables like occupation classification. *DataElements* can be further described using the datatype property `dcterms:description`.

Each study has a set of logical metadata (*LogicalDataSet*) associated with the processing of data, at the time of collection or later during cleaning and re-coding. This includes the definition of variables (paired code and category schemes). The property `dcterms:title` specifies the title of a *LogicalDataSet*. The collected data result in the micro data represented by the *DataFile*. Four DCMI datatype properties share the same domain `dcterms:identifier`, `dcterms:description`, `dcterms:format` and `dcterms:provenance`. An overview of the micro data can be given either by the descriptive statistics or the aggregated data. *DescriptiveStatistics* may be minimal, maximal and mean values, and absolute and relative frequencies. *DataSet* originates from the RDF Data Cube Vocabulary [CRT14]. A *DataSet* represents aggregated data (also known as macro data) such as multi-dimensional tables. Aggregated data is derived from micro data by statistics on groups or aggregates, such as counts, means or frequencies.

4.3.5 Conversion Process of DDI-XML to DDI-RDF

The following subsection summarized work presented in [BCWZ12].

Figure 4.3 depicts an excerpt of the representation of DDI-RDF in RDFS and OWL derived from the underlying model. The classes and properties defined in the DDI domain are represented using the constructs specified by RDFS and OWL. The DDI element *Question* is both, an `rdfs:Class` as well as an `owl:Class`. Relations between two DDI concepts are expressed by new properties defined as the types `rdf:Property` as well as owl:ObjectProperty, since rdfs:Class and owl:Class are connected. According to the example visualized in the figure, the DDI property disco:hasQuestion is specified as an rdf:Property and an owl:ObjectProperty with the domain *Variable* and the range class *Question*. Relationships between DDI elements and XML Schema datatypes are realized by properties that are linked to rdf:Property and owl:DatatypeProperty via rdf:type. The relation disco:literalText, for instance, is an rdf:Property and an owl:DatatypeProperty with the domain class *Question* and the range xsd:string. In the RDFS and OWL representation of the DDI model, the namespace prefixes disco, skos and dcterms are used. The disco namespace prefix refers to a permanent URL.

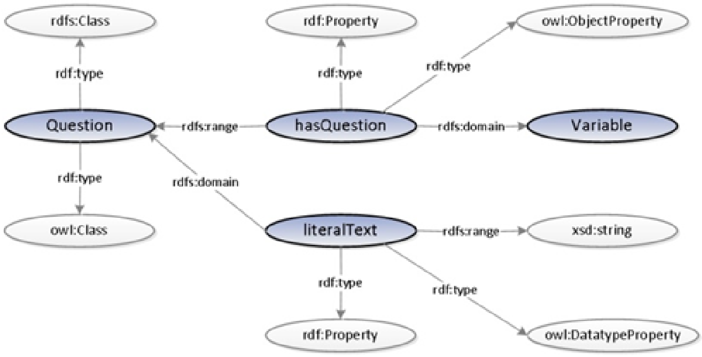


Figure 4.3: Excerpt from the RDFS- and OWL-based representation of DDI-RDF [BCWZ12].

We have implemented a direct and, in parallel, a generic mapping between DDI-XML and Disco. In the direct mapping, different versions of DDI XML documents (as defined in the DDI Specification) can be transformed automatically into an OWL A-Boxes corresponding to the Disco vocabulary. The mappings are implemented as XSLT stylesheets. This transformation is useful for existing DDI XML data and enables easy publication of this data as RDF. Moreover, regardless of different input formats, i.e. different DDI versions, the same Disco output is generated. The XSLTs are available at https://github.com/linked-statistics/DDI-RDF-tools. Bosch et al. [BM11] developed a generic multi-level approach for designing domain ontologies based on XML Schema. XML Schemas are converted to OWL-generated ontologies automatically using XSLT transformations. All information located in the underlying XML Schema of a specific domain is also stored in the generated ontologies. OWL domain ontologies can be inferred completely automatically out of the generated ontologies using SWRL rules [HPSB$^+$04]. On the instance level, XML document instances can be translated automatically into the RDF representation of the generated ontologies by means of Java code. Individuals of domain ontologies can relate to resources of generated ontologies using equivalence relationships [BM12].

4.3.6 Relationships to Other RDF Vocabularies

The following subsection summarized work presented in [BZWG13].

Widely accepted and adopted vocabularies are reused to a large extent. Some features of DDI can be addressed through other vocabularies, such as: representing detailed provenance information of web data and metadata using the PROV Ontology (PROV-O) [LSM13], describing catalogues of data sets using the Data Catalog Vocabulary (DCAT) [ME14], describing aggregate data like multi-dimensional tables using the RDF Data Cube Vocabulary [CRT14], describing formal statistical classifications using the SKOS Extension for Statistics (XKOS) [Cot14], delineating code lists, category schemes, mappings between them, and concepts like topics using the Simple Knowledge Organization System (SKOS) [MB09b], and the Asset Description Metadata Schema (ADMS) [AS13] for representing persistent identifiers. Furthermore, we reuse the external vocabularies Friend of a Friend (FOAF) [Bri10] to describe data about persons, the Organization Ontology (ORG) [Rey14] to model organization-related information, and the DCMI Metadata Terms [Ini] to describe the general metadata of Disco constructs.

In order to represent detailed provenance information of web data and metadata, the classes and properties of PROV-O [LSM13] can be used. Thus, it can be used as a natural vocabulary to attach provenance information to Disco metadata. The terms of PROV-O are organized among three main classes: `prov:Entity`, `prov:Activity` and `prov:Agent`. While classes of Disco can be represented as either as entities or agents, particular processes for, e.g. creating, maintaining and accessing data can be modelled as activities. Properties like `prov:wasGeneratedBy`, `prov:hadPrimarySource`, `prov:wasInvalidatedBy` and `prov:wasDerivedFrom` describe the relationship between classes for the generation of data in more detail. In order to link from a `disco:Study` to its original DDI XML file, the property `prov:wasDerivedFrom` can be used. Moreover, PROV-O allows for representing versioning information by, e.g. using the terms `prov:Revision`, `prov:hadGeneration` and `prov:hadUsage`. PROV-O can also be used to model information and relationships that are relevant for determining the accuracy, quality and comparability of a data set with others. By utilizing the properties `prov:qualifiedInfluence` or `prov:wasInformedBy`, qualified statements can be made about a relationship between entities and activities, e.g. that and how a particular method influenced a particular data collection or data preparation process.

DCAT [ME14] is a W3C recommendation for describing catalogues of data sets. DCAT makes few assumptions about the kind of data sets being described, and focuses on general metadata about the data sets (mostly using Dublin Core) as well as on different ways of distributing and accessing the data set, including availability of the data set in multiple formats. Combining terms from both DCAT and Disco can be useful for a number of reasons:

- Describing collections (catalogues) of research data sets
- Providing additional information about the physical aspects (file size, file formats) of research data files

- Providing information about the data collection that produced the data sets in a data catalogue
- Providing information about the logical structure (variables, concepts, etc.) of tabular data sets in a data catalogue

The *LogicalDataSet* is an extension of the `dcat:DataSet`. Physical, distributed files are represented by the *DataFile*, which is itself an extension of `dcat:Distribution`.

The RDF Data Cube Vocabulary [CRT14] is a W3C recommendation for representing data cubes, i.e. multidimensional aggregate data. A *DataSet* represents aggregate data such as multi-dimensional tables. Aggregate data is derived from micro data by statistics on groups, or aggregates such as counts, means, or frequencies. Data cubes are often generated by tabulating or aggregating unit-record data sets. For example, if an observation in a census data cube indicates that the population of a certain age group in a certain region is *12,653*, then this fact was obtained by aggregating that number of individual records from a unit-record data set. Disco contains a property disco:aggregation, which indicates that a Cube data set was derived by tabulating a unit-record data set. Data Cube provides for the description of the structure of such cubes, but also for the representation of the cube data itself, i.e. the observations that make up the cube data set. This is not the case for Disco, which only describes the structure of a data set, but is not concerned with representing the actual data in it. The actual data are assumed to sit in a data file (e.g. a CSV file or in a proprietary statistical package file format) that is not represented in RDF.

The class `skos:Concept` is reused to a large extent to represent DDI concepts, codes, and categories. SKOS defines the term `skos:Concept`, which is a unit of knowledge created by a unique combination of characteristics. In the context of statistical (meta)data, concepts are abstract summaries, general notions, knowledge of a whole set of behaviours, attitudes or characteristics which are seen as having something in common. Classes of `skos:Concept` may be associated with variables, variable definitions, and questions and are reused to a large extent to represent DDI concepts (`skos:prefLabel`), codes (`skos:notation`), and category labels (skos:prefLabel). *Concepts* may be organized in *ConceptSchemes* (`skos:inScheme`), sets of metadata describing statistical concepts. Hierarchies of DDI concepts can be built using the object properties `skos:broader` and `skos:narrower`. Topical coverage can be expressed using `dcterms:subject`. Disco foresees the use of `skos:Concept` for the description of topical coverage. Spatial, temporal and topical coverage are directly attached to *Studies*, *LogicalDataSets*, and *DataFiles*. *Universes* and *AnalysisUnits* are also a `skos:Concept`. Therefore, the properties defined for `skos:Concept` can be reused. *KindOfData*, pointing to a `skos:Concept`, describes, with a string or a term from a controlled vocabulary, the kind of data documented in the logical product(s) of a Study. Using `dcterms:format`, *DataFiles* formats can be defined.

The use of formal statistical classifications is very common in research data sets; these are treated in Disco as SKOS concepts, but in some cases, those working with formal statistical classifications may desire more expressive capability than SKOS provides. To support such users, the DDI Alliance also develops XKOS [Cot14], a vocabulary that extends SKOS to allow for a more complete description of such classifications. While the use of XKOS is not required by this vocabulary, the two are designed to work in complementary fashion. SKOS properties may be substituted by additional XKOS properties.

Persons and organizations, in particular, may hold one or more persistent identifiers of particular schemes and agencies (e.g. ORCID[33], FundRef[34]) that are not considered by the specific IDs of Disco. In order to include those identifiers and to distinguish between multiple identifiers for the same class, ADMS [AS13] is utilized. As a profile of DCAT [ME14], ADMS aims to describe semantic assets, i.e. reusable metadata and reference data. The class `adms:Identifier` can be added to a rdfs:Resource by using the property `adms:identifier`. This identifier class can contain properties that define the particular identifier itself, but also its scheme, version and managing agency. However, although utilized primarily for describing identifiers of persons and organizations, it is allowed to attach an `adms:Identifier` class to all classes in Disco.

4.3.7 Discussion and Limitations

During the development of DDI-RDF, we faced limitations regarding the interplay with other vocabularies. This discussion has been presented in [BZWG13].

The interplay of Data Cube [CRT14], Disco and PROV-O [LSM13] needs further exploration with respect to the relationships shared by of aggregate data, aggregation methods and the underlying micro data. The goal would be to drill down to the related micro data based on a search resulting in aggregate data. A researcher could then analyse the micro data often, only with the constraints of access restrictions to the data (i.e. access only in the closed shop of research data centres or anonymization methods to assure confidentiality). On the one hand, aggregate data are often easily available and provide a quick overview. On the other hand, micro data enable more detailed analyses.

4.4 Publishing a Domain-specific Thesaurus with Linked Data

A major aspect of publishing data as Linked Data is providing links to other data sources. Such links can be built upon terms of library vocabularies like thesauri, taxonomies, etc., because they provide a common vocabulary of terms that are suggested for the use in document indexing. Thesauri are crucial instruments for information retrieval in big document databases containing, e.g. bibliographical information. The terms of a thesaurus are typically connected to each other by equivalence, associative or hierarchical relations. Networks spanned by those term relations can serve as interlinking hubs between Linked Data sources [Neu09]. Thesauri have originally been standardized by ISO 2788 [ISO86] and ISO 5964 [ISO85]. These standards have recently been revised,

[33] http://orcid.org/

[34] http://www.crossref.org/fundref/

merged and extended to the new standard ‘ISO 25964: Thesauri and interoperability with other vocabularies’ [NIS11]. It aims to support ‘the development and application of thesauri in today’s expanding context of networking opportunities’.

In most cases, thesauri have been designed exclusively for specific domains or document collections in order to cover a specific knowledge field as extensively as possible. For information retrieval on documents indexed by multiple thesauri, mappings between the participated thesauri have to be defined. This occurs when multiple document collections or documents of different or related disciplines are used during a retrieval process. Moreover, in recent years, additional document types have been included within information retrieval that uses different thesauri or classification systems, e.g. the ELSST (European Language Social Science Thesaurus)[35] for survey data in the social sciences.

For social science literature, the Thesaurus for the Social Sciences[36] – shortened TheSoz – serves as a crucial instrument. It is applied among other disciplinary information systems in the databases, SOLIS (Social Science Literature Information System)[37] and SOFIS (Social Science Research Information System)[38], both of which are owned and maintained by GESIS - Leibniz Institute for the Social Sciences[39]. It is available in four languages (German, English, French and Russian) and contains over 12,000 keywords, out of which 8,000 are so-called descriptors, i.e. preferred terms for indexing documents, and 4,000 non-descriptors, i.e. non-preferred terms, for which preferred terms are recommended to be used. The thesaurus covers all topics and sub-disciplines of the social sciences. Additionally, terms from associated and related disciplines are included in order to support an accurate and adequate indexing process of interdisciplinary, praxis-oriented and multi-cultural documents. The thesaurus is owned and maintained by GESIS, the largest infrastructure organization in Germany, which provides research-based infrastructure services for the social sciences.

Following recent developments of the Semantic Web that promise, according to [Kra08, Sve07, Vat10], large potential for library vocabularies, the TheSoz has been made available on the web in a compatible and machine-readable format for providing and sharing its relevant information with a greater community. With SKOS (Simple Knowledge Organization System) [MB09b] it uses a ‘standard way to represent knowledge organization systems using the Resource Description Framework (RDF)’[40]. With RDF, information can be parsed and reused in an interoperable way, which allows easy application of the TheSoz by other systems. First attempts [ZS09a] for modelling the TheSoz with SKOS were made in 2009, when SKOS was announced as a standard by the W3C. Many organizations and libraries have begun bringing their thesauri and vocabularies to the web in SKOS format since then [Mal08, SIRK08, De 09]. These developments can also be observed in Germany, where the Thesaurus for Economics (STW) of the ZBW, the

[35] http://elsst.esds.ac.uk/

[36] http://www.gesis.org/en/services/research/thesauri-und-klassifikationen/social-science-thesaurus/

[37] http://www.gesis.org/en/services/research/solis-social-science-literature-information-system/

[38] http://www.gesis.org/en/services/research/sofis-social-science-research-information-system/

[39] http://www.gesis.org/en/home/

[40] http://www.w3.org/2004/02/skos/intro

Leibniz Information Centre for Economics, has been published as Linked Data [Neu09] as well as the Subject Headings Authority File (SWD) of the German National Library (DNB) [HK10] and UMTHES[41], the environment thesaurus of the German Federal Environmental Agency .

Section 4.4.1 provides an overview of the SKOS format. The transformation process of the TheSoz is described in detail in Section 4.4.2, which is based on [ZS09a, MZS10a] and has been revised and extended in [ZSMM13]. The generation of mappings and links to other thesauri, which are necessary for complex information retrieval purposes and for the inclusion into the LOD cloud diagram, are presented in Section 4.4.3, as published in [MZS10b, MZS10a, MZJ$^+$11, AEE$^+$12, KZ13]. In Section 4.4.4, we discuss the observed obstacles to the transformation and modelling process. Finally, in Section 4.4.4.3, we show how semi-automatic matching tools can support and enhance the detection and evaluation of vocabulary cross-walks. This section has been published in [KREZ14].

4.4.1 SKOS – The Simple Knowledge Organization System

With SKOS (Simple Knowledge Organization System) [MB09b], a standard was declared in August 2009 by the W3C, which allows the representation of indexing vocabularies like thesauri, classification systems or taxonomies in a machine-processable standard format for the Semantic Web. Thus, this standard enables the sharing of such data sets on the web and linking to other data sets on the web.

> ‘The SKOS data model provides a standard, low-cost migration path for porting existing knowledge organization systems to the Semantic Web’ [IS09b].

Based on the fundamental Semantic Web standard format RDF, SKOS provides a high interoperability in connection with other standards, formats and applications. Term relations, hierarchies and the overall structure and semantics of vocabularies can be represented through the use of specific classes and properties. In the SKOS data model, a vocabulary is typically represented as a *ConceptScheme* that holds multiple *Concepts*. In contrast to traditional vocabularies, which are defined by the ISO standards 2788 [ISO86] and 5964 [ISO85], SKOS holds a concept-centric organization of terms instead of a term-centric point of view. Each of these concepts can be labelled by multiple labels in multiple languages in order to express the *preferred labels* of a concept, i.e. those which are recommended for use, or *alternative* or *hidden labels* that are related to the concept. SKOS allows only one preferred label per language for each concept. SKOS concepts can be linked to other concepts by associative and hierarchical properties that are oriented on relations of the ISO norms for thesauri e.g. *broader*, *narrower* and *related* relations. Furthermore, editorial and lexical attributes and notes can be modelled as well as mapped to other concept schemes (e.g. to other thesauri) by the mapping properties of SKOS. Table 4.4 depicts an overview of basic SKOS classes and properties. Not all classes and properties of SKOS are described below; we present only the ones that are

[41] http://data.uba.de/umt/de/hierarchical_concepts.html

used often and particularly for the representation of the TheSoz. A complete and detailed description of all classes and properties of SKOS can be found at [MB09b, IS09b].

Thesauri organize complex relations between terms even on a lexical level. The SKOS eXtension for Labels (SKOS-XL) [MB09a] presents an optional extension for SKOS, which allows the identification and description of single lexical entities as well as relations between them. This provides a complexity in term relationships, which is needed by several vocabularies. Table 4.5 presents an overview of classes and properties of SKOS-XL. The use of SKOS-XL for the TheSoz will be described in detail in the following sections.

4.4.2 Modelling the Thesaurus for the Social Sciences (TheSoz) Using SKOS

The transformation process of a thesaurus into the SKOS format has been split up into three steps. Hence, it follows the structured method introduced in [vAMMS06], which consists of the following steps: (1) analysis of the structure, the extent and the complexity of the thesaurus, including contained terms and relations between terms, (2) a mapping of all detected terms and relations to adequate SKOS classes and properties and (3) the technical conversion of the thesaurus according to the defined mapping. This method aims to ensure the quality and utility of the resulting conversion and focuses on two goals: the interoperability and completeness of the converted thesauri.

Thesauri are mostly collections of terms that stand in specific relation to each other, typically in hierarchical or associated relations. While thesauri are more term-centralized instruments that have traditionally been designed and maintained in libraries over decades, the major design aspect of SKOS is a concept-centralized view on the thesaurus or classification system. In SKOS, there are concepts (`skos:Concept`) that represent a semantic concept. Therefore, each concept can hold more than one label. Exactly one label serves as the preferred label (`skos:prefLabel`), while additional labels can be included as alternative or hidden labels (`skos:altLabel` and `skos:hiddenLabel`). These labels describe, e.g. variants, additional expressions or unusual variants of the preferred terms, which are indicated as non-preferred terms in thesauri.

This design issue proves to be an obstacle when there are multiple relationships between preferred and non-preferred terms for a single term concept. These relationships have to be concentrated and rearranged as properties under a SKOS concept. Figure 4.4 presents a concept-based term relation in SKOS, where a single concept consists of three terms. One of the terms is the preferred term that is recommended for use; the two other terms depict the non-preferred terms. Most traditional thesauri have no such concept-based representation. Instead, the terms and their type (preferred or non-preferred) are organized via direct relations between the terms, e.g. *USE* and *USED FOR* (see Figure 4.4).

URI	Definition
skos:Concept	Builds the main class of SKOS and defines a suggestive semantic concept, thought or idea.
skos:ConceptScheme	skos:Concepts can be organized in skos:ConceptScheme, which build an aggregation of multiple concepts. There can be relations between multiple skos:Concept inside one skos:ConceptsScheme or between different skos:ConceptScheme. Regarding the representation of Knowledge Organization Systems, a skos:ConceptScheme typically comprises one terminology system or thesaurus.
skos:inScheme	Describes the relation of a skos:Concept to a skos:ConceptScheme.
skos:prefLabel	skos:prefLabel describe the preferred lexical term for a specific skos:Concept. Each skos:Concept can hold only one skos:prefLabel.
skos:altLabel	Alternative labels can be defined in order to provide alternative or non-preferred terms of a skos:Concept.
skos:hiddenLabel	skos:hiddenLabel is defined for describing, e.g. unusual spellings or terms for a skos:prefLabel.
skos:notation	skos:notation provides the possibility to connect a skos:Concept with entries from a notation system.
skos:note	This property enables the inclusion of notes associated with a skos:Concept. Subclasses of skos:note include, e.g. skos:editorialNote, skos:example, skos:definition and skos:scopeNote.
skos:broader, skos:narrower and skos:related	These semantic relations can be defined between multiple classes of a skos:Concept. They enable the construction of hierarchies of concepts.
skos:broadMatch, skos:closeMatch, skos:exactMatch, skos:narrowMatch and skos:relatedMatch	The mapping properties of SKOS are similar to the semantic relations described above. In contrast to them, the mapping properties can only be applied between classes of a skos:Concept of different skos:ConceptScheme.

Table 4.4: Overview of SKOS classes and properties.

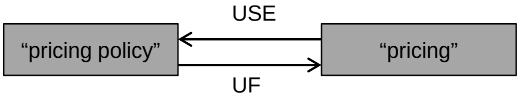


Figure 4.4: Example of a USE/USEDFOR relationship.

URI	Definition
skosxl:Label	This class contains lexical entities in a plain literal format.
skosxl:literalForm	This property describes the precise literal inside a skosxl:Label class. There can only be one skosxl:literalForm inside a skosxl:Label class. This restriction includes different languages of a lexical entity, which have to be modelled in another skosxl:Label class.
skosxl:prefLabel, skosxl:altLabel and skosxl:hiddenLabel	These properties are treated analogous to skos:prefLabel, skos:altLabel and skos:hiddenLabel and refer to a skosxl:Label class, which contains the stated label as skosxl:literalForm.
skosxl:labelRelation	The skosxl:labelRelation enables the modelling of specific relationships between skosxl:Label classes. This property allows the representation of complex relations between literals and terms.

Table 4.5: Classes and properties of SKOS-XL.

Thesaurus Analysis

The basis for a transformation of a thesaurus to the SKOS format is a detailed analysis of the thesaurus. Attention is not only paid to terms and the existing associative and hierarchical relations between them, but also to the general structure and design issues of the thesaurus, e.g. the existence of additional classification systems or how far the examined thesaurus conforms to established ISO norms. The Thesaurus for the Social Sciences contains about 12,000 keywords, of which more than 8,000 are *Descriptors* (authorized keywords) and about 4,000 are *Non-Descriptors*. The relationships between these keywords are expressed as *broader*, *narrower* or *related* terms and there are also *USE* (see Figure 4.4 above) and *USE COMBINATION* (see Figure 4.5 below) relations and their counterparts (*USED FOR* and *USED FOR COMBINATION*). Additionally, a classification hierarchy is provided and each thesaurus term is dedicated to one or more classification terms.

The TheSoz contains a special type of non-descriptor called AD (*Alternative Non-Descriptor*) that differs from the international standard norms for thesauri. An alternative non-descriptor in the TheSoz is used to describe ambiguities in relations between terms. Such descriptors hold more than one *USE* and/or *USE COMBINATION* relation at the same time. There are about 216 of such *AD* terms in the TheSoz.

Example The term ‘committee’, which is classified as an AD term, holds *USE* relations to the preferred terms ‘working group’, ‘parliamentary committee’, ‘Wirtschaftsausschuss’

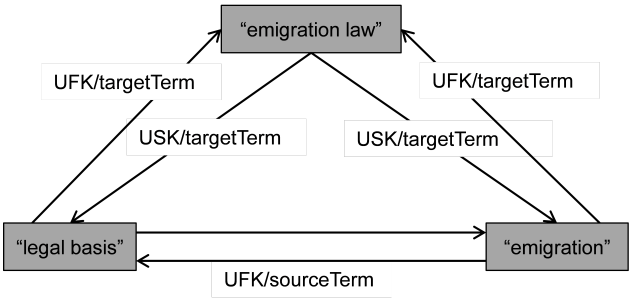


Figure 4.5: Example of a USE COMBINATION relationship.

(no English translation available; means the 'Standing Committee on Industry and Trade') and 'advisory panel' at the same time. Additionally, it contains the *USE COMBINATION* relation to the combined use of the terms 'product' and 'quality'. Terms of the type AD describe generic and ambiguous terms that have different concrete meanings in specialized sub-contexts. This is expressed through the use of more than one *USE* and/or *USE COMBINATION* relation for only one term. In this case, it means that the term 'committee' is semantically so general and ambiguous that it is recommended to use a more precise term to describe the intended semantics. Figure 4.6 depicts this example of an AD term.

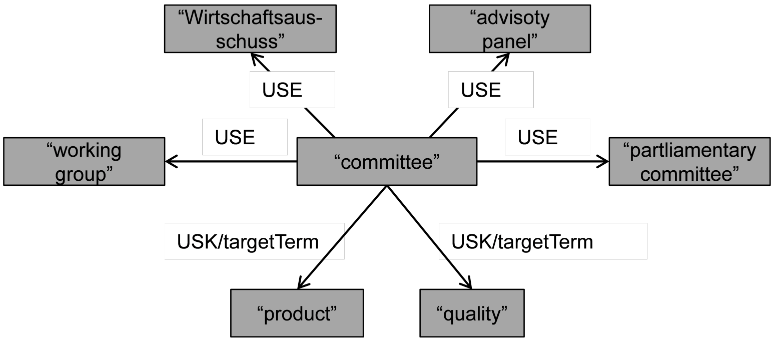


Figure 4.6: Example of an Alternative Non-Descriptor (AD).

An alternative to represent ambiguities is to transform the AD term to multiple non-descriptors that are extended with specific context information either in their label itself (e.g. 'committee (working group)') or in a note, e.g. 'used in the context of a working group'. But this solution omits the technical processability and detection of the ambiguity because the terms would be identified as different ones.

Mapping to SKOS

For most of the thesaurus items, i.e. terms and relations, adequate SKOS properties and classes can be identified easily because TheSoz conforms broadly to the standard norms for thesauri. Problems occur when mapping special data items and/or relations that do not conform to thesauri standards like the AD terms of the TheSoz described in the previous subsection. However, since SKOS is based on RDF, it is possible to define additional relations without greater effort. Therefore, a precise mapping to SKOS is more complex than a simple mapping [ZS09a]. In order to obey the concept-based structure of SKOS, but without risking the loss of relevant relations between preferred and non-preferred terms, classes and properties of SKOS-XL have been used [MZS10a]. For this reason, SKOS-XL has also been used for the conversion of the EUROVOC thesaurus [De 09]. Properties of SKOS-XL have been developed explicitly for the representation of lexical issues and provide the possibility to model relations between multiple terms inside one SKOS concept. These label relations allow the definition of own relations between lexical labels, such as typical equivalence relationships like *USE* or compound equivalence relationships like *USE COMBINATION* and their counterparts, which are necessary components of the TheSoz.

Table 4.6 presents the mapping from terms and relations of the TheSoz to adequate SKOS classes and properties. As described above, personal classes and properties have been defined in order to represent additional semantics as well as complex relations of the TheSoz.

Extensions have necessarily been defined for representing complex and relevant relations in the TheSoz correctly. They are described in detail using RDF Schema in order to ensure further processing and interoperability with other data sets on the web. Table 4.7 provides an overview of the SKOS extensions defined for the TheSoz.

Figure 4.5 outlines the *USE/USED FOR* term relations within a concept, where the term ‘pricing policy’ is the preferred one and is recommended for use instead (see the `thesoz:use` and `thesoz:usedFor` relations in the figure) of the non-preferred term (depicted as well, as `skosxl:altLabel`). This modelling approach provides more semantic information than the single use of `skosxl:prefLabel` and `skosxl:altLabel` allows.

These relations could also be modelled by only using `skos:altLabel` and `skos:prefLabel`, but the addition of personal relations provides more semantics about the relationship. Furthermore, it builds the basis for distinguishing *USE* and *USED FOR* relations from *USE COMBINATION* and *USED FOR COMBINATION* relations, which are introduced below. Distinguishing between these relations is necessary because a term of the TheSoz can hold together multiple such relations (see the example of the AD term). A representation of the relations *USE COMBINATION* and *USED FOR COMBINATION* is depicted in Figure 4.4 below.

In the example above, the term ‘university ranking’ is a non-preferred term in the TheSoz, i.e. it is not recommended to use this term. Instead, it is advised that a combination of the terms ‘university’ and ‘ranking’ be used to index documents because both are preferred terms. In order to represent this special relationship in SKOS, the property `skosxl:labelRelation` is used to define personal semantical relations. Thus, the term ‘university ranking’ gets the relation `thesoz:compoundNonPreferredTerm` and the two other preferred terms are extended by the relation `thesoz:preferredTermComponent`.

Thesaurus Element	Description	SKOS Class / Property
DD	Descriptor	skosxl:prefLabel
ND	Non-Descriptor	skosxl:altLabel
AD	Alternative Non-Descriptor	skosxl:altLabel
NT	Narrower Term	skos:narrower
BT	Broader Term	skos:broader
RT	Related Term	skos:related
USE	Use (Example: For X, use Y)	thesoz:use in conjunction with the class thesoz:EquivalenceRelationship
UF	Used For (Example: Y is used for X)	thesoz:usedFor in conjunction with the class thesoz:EquivalenceRelationship
USK	Use Combination (Example: For X, use Y and Z in combination)	thesoz:compoundNonPreferredTerm and thesoz:preferredTermComponent in conjunction with the class thesoz:CompoundEquivalence
UFK	Used For Combination (Example: Use Y in combination with Z for X)	thesoz:compoundNonPreferredTerm and thesoz:preferredTermComponent in conjunction with the class thesoz:CompoundEquivalence
translation	Translation of the terms via language tags	thesoz:hasTranslation and thesoz:isTranslationOf
scope	Scope Notes	skos:scopeNote
notationcode	Numerical code of the systematic classification, to which terms are assigned	skos:notation

Table 4.6: Mapping of TheSoz elements to classes and properties of SKOS.

All three relations hint at a new class, which is defined as a `thesoz:CompoundEquivalence`. In contrast to the *USE* relation in the former example (see Figure 4.5), it is now clear that both preferred terms have to be used together, while this information is omitted in a single *USE* relation.

Furthermore, the label relations of SKOS-XL allow a consistent and correct representation of the alternative non-descriptors of the TheSoz, where a non-preferred term holds relations to multiple preferred terms. A modelling example of such a term is depicted below in Figure 4.6. The term ‘committee’, which is classified as an AD term, holds *USE* relations to the preferred terms ‘working group’, ‘parliamentary committee’, ‘Wirtschaftsausschuss’

Extension	Description
thesoz:Classification	Element of the classification hierarchy of the TheSoz, which is defined as a subclass of a SKOS Concept.
thesoz:Descriptor	Descriptors of the TheSoz represented as a concept and defined as subclasses of a SKOS Concept.
thesoz:Equivalence Relationship	An equivalence relationship between two terms, where the terms are assigned via thesoz:use and thesoz:usedFor properties. This is a subclass of skosxl:Label.
thesoz:Compound Equivalence	A compound equivalence between terms. For constructing USE COMBINATION and USED FOR COMBINATION relations between terms. The non-preferred term is assigned via the compoundNonPreferrdTerm property. The preferred terms are modelled via the preferredTermComponent property. This is a subclass of skosxl:Label.
thesoz:use	Use relation, which is defined as a subproperty of skosxl:labelRelation.
thesoz:usedFor	Used for relation, which is defined as a subproperty of skosxl:labelRelation.
thesoz:hasTranslation	Relation between different languages of a term, which is defined as a subproperty of skosxl:labelRelation.
thesoz:isTranslationOf	Inverse property of thesoz:hasTranslation.
thesoz:preferredTerm Component	A preferred term as a component for a USE COMBINATION or USED FOR COMBINATION relation. This property is defined as a subproperty of skosxl:labelRelation.
thesoz:compoundNon PreferredTerm	The non-preferred term as a component for a USE COMBINATION or USED FOR COMBINATION relation. This property is defined as a subproperty of skosxl:labelRelation.
thesoz:isPartOf EquivalenceRelationship	This property serves as counterpart for thesoz:use and thesoz:usedFor.
thesoz:isPartOf CompoundEquivalence	This property serves as counterpart for thesoz:preferredTermComponent and thesoz:compoundNonPreferredTerm.

Table 4.7: SKOS extensions for TheSoz.

(i.e. ‘Standing Committee on Industry and Trade’) and ‘advisory panel’ at the same time. Additionally, it contains the *USE COMBINATION* relation to the combined use of the terms ‘product’ and ‘quality’. These multiple and complex relations cannot be modelled without personal defined relations of SKOS-XL because the ambiguities of the

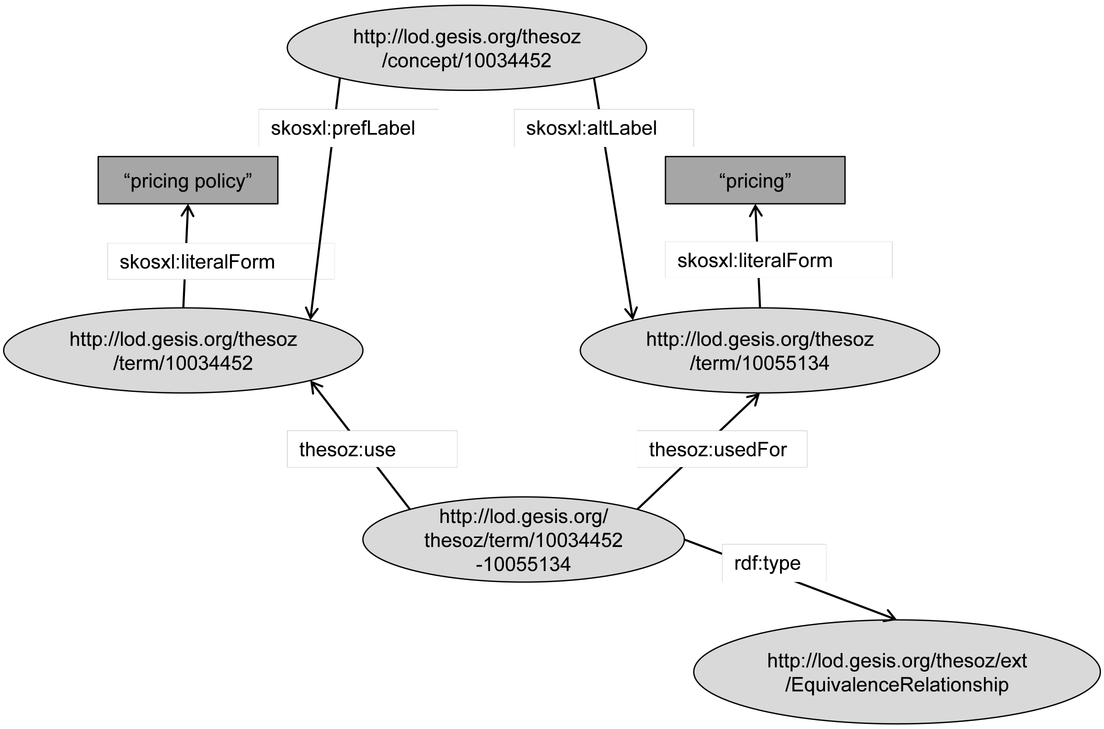


Figure 4.7: Example of a USE/USED FOR relation using SKOS extensions.

term would be lost.

The classification hierarchy of the TheSoz can be mapped to SKOS without further problems (see Table 4.6). In order to distinguish between concepts of terms and the concepts of classification terms, the classes `thesoz:Descriptor` and `thesoz:Classification` have been defined as subclasses of skos:Concept. The numerical code of each classification term, which appears in the URI as well as in `skos:notation`, is the same code that is included in the `skos:notation` of each concept containing descriptors (see Table 4.6). By referencing these notation codes via URIs, a connection between descriptors and their according classification terms is established. Each classification element holds all associated thesaurus terms as narrower concepts via the property `skos:narrower`. Each thesaurus term holds the associated classification element as a broader concept via the relation `skos:broader` backwards. Multiple assignments are possible, i.e. a term can be assigned to multiple classification elements. In addition, the hierarchy of the classification terms themselves, which is denoted as child and parent relations in the source, is modelled by `skos:narrower` and `skos:broader` relations as well.

Technical Conversion

Based on the defined mapping, the conversion program has been developed. This is typically a script that has to be executed on the dedicated thesaurus. In case of the

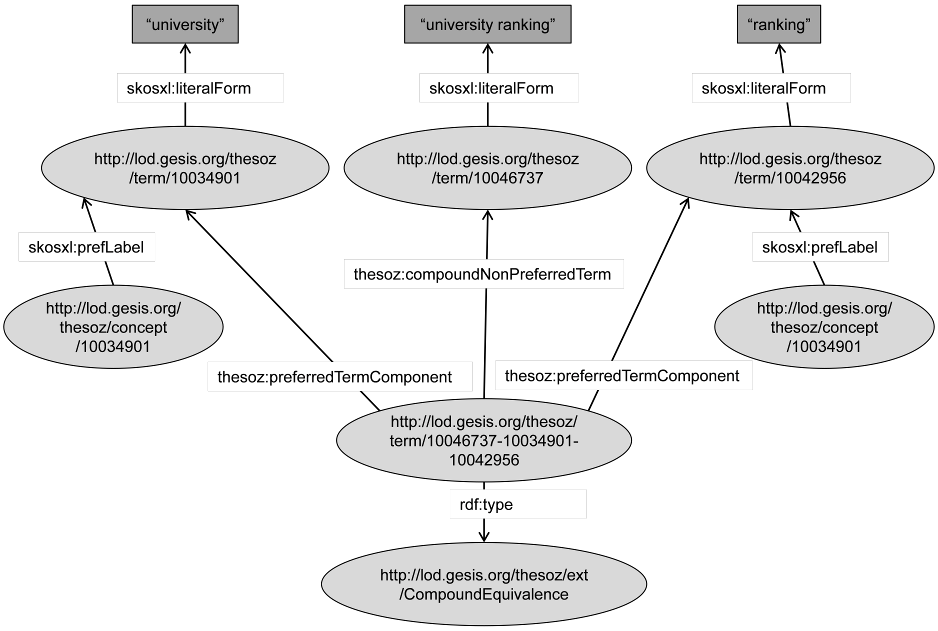

"university"
"university ranking"
"ranking"
skosxl:literalForm
skosxl:literalForm
skosxl:literalForm
http://lod.gesis.org/thesoz /term/10034901
http://lod.gesis.org/thesoz /term/10046737
http://lod.gesis.org/thesoz /term/10042956
skosxl:prefLabel
thesoz:compoundNonPreferredTerm
skosxl:prefLabel
http://lod.gesis.org/ thesoz/concept /10034901
http://lod.gesis.org/ thesoz/concept /10034901
thesoz:preferredTermComponent
thesoz:preferredTermComponent
http://lod.gesis.org/thesoz/ term/10046737-10034901- 10042956
rdf:type
http://lod.gesis.org/thesoz/ext /CompoundEquivalence

Figure 4.8: Example for a USE COMBINATION relation using SKOS extensions.

TheSoz, the technical conversion process is carried out by XSL transformations. The original digital format of the TheSoz, which was already encoded in XML, was converted to SKOS in RDF/XML format. The use of XSLT makes it easy to adjust or extend the mapping in case of later revisions or to implement additional or new mappings. Additionally to the mapping, each defined concept as well as each term itself received its own URI, which provides a persistent and unique identification. This is a very important aspect for reuse and links on the web, e.g. links from and to other data sets. All URIs are defined in the context path `http://lod.gesis.org/thesoz/`, which serves as the base URI. The URI has been chosen according to the naming conventions of web addresses of GESIS and in order to leave room for the publication of further data sets as Linked Data. The namespace of the personal classes and properties is defined at `http://lod.gesis.org/thesoz/ext/` and is shortened by the prefix `thesoz`. The SKOS version of the thesaurus contains three types of URIs, one for the terms, i.e. the descriptors and non-descriptors, one for the concepts summarizing descriptors and non-descriptors, and one for the labels of the classification hierarchy.

- URI scheme for Descriptors: `http://lod.gesis.org/thesoz/concept/########`
- URI scheme for Terms: `http://lod.gesis.org/thesoz/term/########`
- URI scheme for Classification Terms: `http://lod.gesis.org/thesoz/classification/#.#.##`

This allows an easy distinguishing between descriptors and classification terms by only knowing the concept URI. After the transformation process, the resulting SKOS version

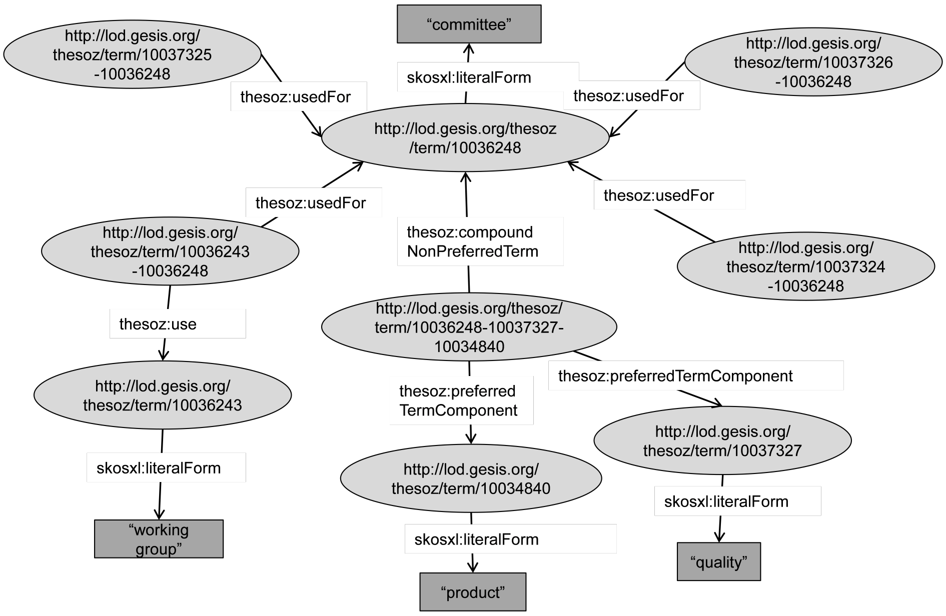

"committee"
http://lod.gesis.org/thesoz/term/10037325-10036248
skosxl:literalForm
thesoz:usedFor
thesoz:usedFor
http://lod.gesis.org/thesoz/term/10037326-10036248
http://lod.gesis.org/thesoz/term/10036248
thesoz:usedFor
thesoz:usedFor
thesoz:compound NonPreferredTerm
http://lod.gesis.org/thesoz/term/10036243-10036248
http://lod.gesis.org/thesoz/term/10037324-10036248
thesoz:use
http://lod.gesis.org/thesoz/term/10036248-10037327-10034840
thesoz:preferredTermComponent
http://lod.gesis.org/thesoz/term/10036243
thesoz:preferred TermComponent
http://lod.gesis.org/thesoz/term/10037327
skosxl:literalForm
http://lod.gesis.org/thesoz/term/10034840
skosxl:literalForm
"working group"
skosxl:literalForm
"quality"
"product"

Figure 4.9: Example of an AD term using SKOS extensions.

of the TheSoz has been tested and validated by various established validation services for RDF and SKOS. It is available via a SPARQL endpoint[42], as a HTML representation[43] and as a dump file in the RDF/XML and RDF/Turtle formats[44].

4.4.3 Establishing Links to Other Thesauri Using Semi-automatically Link Detection

While Section 4.4.2 covered the transformation process of a domain-specific thesaurus to the SKOS format, this section focuses on the application of existing cross-walks to the SKOS mapping properties [MZS10b, MZS10a] and the semi-automatic detection of links between two thesauri in a case study, which has been presented in [MZJ+11]. Links between three thesauri have been detected and established: (1) the TheSoz[45], held and maintained by GESIS, (2) STW[46], the Standard Thesaurus for Economics of the Leibniz Information Centre for Economics (ZBW)[47] and (3) AGROVOC[48], the agricultural thesaurus, held and maintained by the Food and Agriculture Organization of the United Nations (FAO)[49]. All thesauri are available in the SKOS format, but

[42] http://lod.gesis.org/thesoz/sparql
[43] http://lod.gesis.org/thesoz/
[44] http://www.gesis.org/en/services/research/thesauri-und-klassifikationen/social-science-thesaurus/
[45] http://www.gesis.org/en/services/research/thesauri-und-klassifikationen/social-science-thesaurus/
[46] http://zbw.eu/stw/versions/latest/about.en.html
[47] http://zbw.eu/
[48] http://aims.fao.org/website/AGROVOC-Thesaurus/sub
[49] http://www.fao.org/

Thesaurus Element	Description	SKOS Representation
code	Numerical code of the classification item	skos:notation
description	Title of the classification item	skos:prefLabel
child	Child nodes of the current classification item	skos:narrower
parent	Parent node of the current classification item	skos:broader

Table 4.8: Mapping of TheSoz classification hierarchy to SKOS.

have specific differences in their structure. The resulting links are available within the participating thesauri, which builds the technical foundation for Linked Data applications using links between connected thesauri.

4.4.3.1 The Semantics of Links between Thesauri

While thesauri by themselves are established retrieval instruments for searches in big document collections, they have mostly been designed and developed separately from other terminologies and exclusively for a specific domain or context. However, when searching through information collections over the web, especially with an interdisciplinary and cross-domain information interest, the simultaneous use of different vocabularies gains importance. To support such a search, the connection between the terms of thesauri is a good solution [Kra08]. The project KoMoHe (Kompetenzzentrum Modellbildung und Heterogenitätsbehandlung) [MP08a] funded by the DFG began to connect thesauri of different domains with each other via mappings, the so-called cross-concordances or cross-walks. These cross-walks have been used for information retrieval purposes. They are currently held in a relational database at GESIS and are used as a terminology hub for search and query expansion in the information portal sowiport[50].

Cross-walks are defined in [Kra03a] as intellectually and manually detected connections between vocabularies that describe equivalent, hierarchical or associative relationships between terms of the participating vocabularies. Typically, those connections are established in both directions, i.e. bilaterally. Bilateral cross-concordances do not have to be symmetrical, i.e. the term ‘computer’ of a vocabulary A is mapped onto the term ‘information system’ in vocabulary B, while the latter term can be mapped to the term ‘database’ back in vocabulary A, because it seems to be the more suitable term from the context of vocabulary B. These characteristics of cross-walks are the result of a domain-specific and intellectual detection of those connections that is usually carried out by domain experts. To ensure a high-quality mapping result, it is important that the meaning and the semantics of terms as well as their shared relationships are understood

[50] http://www.gesis.org/sowiport/en/home/overview.html

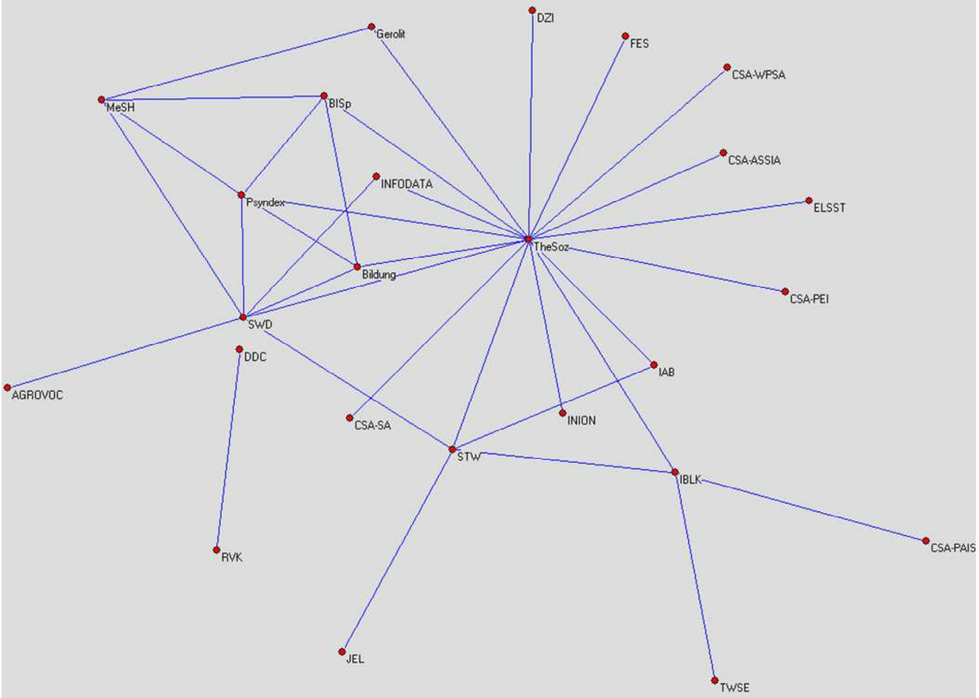


Figure 4.10: Network of controlled vocabularies built in KoMoHe [MP08a].

completely and correctly. This is the main aspect where intellectual and manual methods differ from automatic algorithms [MP08b].

The cross-concordances of the project KoMoHe have involved full or, at least, extensive parts of the participating vocabularies (see Figure 4.10). Before the intellectual detection of mappings of terms, possible topical and syntactical overlaps between the vocabularies have been examined. The mapping process itself is based on practical rules and guidelines [MP08a], where all relations (incl. scope notes) inside a thesaurus are examined and used. Recall and precision of the defined mappings are then evaluated manually in domain-specific databases by domain experts. This is especially relevant for multiple mappings like 1:n relations or mapping combinations, where one term is mapped to a combination of two or more terms. 1:1 relations are preferred. Finally, the detected cross-walks are evaluated by domain experts regarding their semantical correctness, which is supported by the empirical evaluation of samples on precision and recall.

Thesauri in the SKOS format and Semantic Web applications, which use those data sets, benefit from the links between them as well. Such links are also one of the basic requirements for inclusion in the LOD cloud diagram [SBJC14]. The SKOS mapping properties (see Table 4.4 in Section 4.4.1) provide standardized relations in order to link the SKOS concepts of different concept schemes, which are represented in this scenario by the three participating thesauri. When modelling cross-concordances in the SKOS format, inconsistencies and problems can occur that are caused by idiosyncrasies in the thesauri. The obstacles observed during the case study are described in Section 4.4.4 in detail.

4.4.3.2 Establishing Links Based on Existing Mappings

This section describes the process of converting existing cross-concordances between two thesauri to the SKOS format. The cross-concordances have been built as part of a major terminology mapping initiative in the KoMoHe project. The resulting links serve not only as a connection between the thesauri but also as an entrance in the LOD cloud. The work in this section has been published in [MZS10b, MZS10a].

The Standard Thesaurus Wirtschaft (STW) of the European Library of Economics (ZBW) consists of 5,800 descriptors and 17,000 non-descriptors. It covers economical topics as well as related disciplines like politics, sociology and justice [Neu09]. Furthermore, all descriptors are organized in a taxonomy of about 500 items. The STW was converted to SKOS in 2009, becoming one of the first thesauri to be transformed to the SKOS format in Germany. The mapping to classes and properties of SKOS is straightforward. There is only a minor extension by two subclasses of `skos:Concept` in order to distinguish between descriptors of the thesaurus (`zbwext:Descriptor`) and items of the taxonomy (`zbwext:Thsys`).

Cross-concordances between the STW and the TheSoz have been established during the project KoMoHe. From TheSoz to STW, 7,729 mappings have been defined, while 6,651 mappings have been identified for the other direction, from STW to TheSoz. The number of mappings differ because they are not required to be symmetrical. Several different types of mappings have been identified. Examples for each mapping type are presented in Table 4.9 below.

If precise mappings between two thesauri are available and the participating thesauri are available in the SKOS format as well, then the mappings between them can easily be represented using the SKOS mapping properties. For the cross-walks between STW and TheSoz, only simple mappings like equivalent, broader, narrower and related mappings could be applied to SKOS because of the restriction in the available mapping properties. This means that only mappings corresponding to `skos:exactMatch`, `skos:closeMatch`, `skos:broadMatch`, `skos:narrowMatch` and `skos:relatedMatch` can be modelled. Multiple mappings of one term (1:n relations) can also be represented. But, mappings to term combinations as well as information on the relevance of the mapping have been omitted because there is no suitable solution in representing them in SKOS yet.

In contrast to the mapping file with the cross-concordances, an SKOS mapping is only established between the corresponding URIs of the terms (see Figure 4.11), not between the terms themselves. Since bilateral cross-concordances do not have to be symmetrical, the mapping in SKOS between two terms should be defined in both directions. This ensures that no valuable mapping information is lost. Listing 4.5 presents an example mapping of two concepts.

Mapping Type	Description	Example
0	No term of the target vocabulary can be mapped to the source term.	Childhood (TheSoz) 0
=	Describes an equivalent mapping of a source term to a target term.	Taxes (TheSoz) = Tax (STW)
<	Describes that the source term is a narrower term according to the target term.	Media Technology (TheSoz) < Technology (STW)
>	Describes that the source term is a broader term according to the target term.	Disaster (TheSoz) > Natural Disaster (STW)
^	Indicates that source term and target term are associated (somehow) with each other.	Bundesrat ('Upper House of German Federal Parliament', TheSoz) ^ Parliament (STW)
+	The + is added to one of the mapping types above (except for 0), if the source term is mapped to a combination of at least two terms.	Electronic Government (TheSoz) =+ Public Administration + Internet (STW)
o	The o is added to one of the mapping types above (except for 0), if an alternative mapping of the source term is available. This alternative mapping must not be of the same type and can differ in its relevance.	Municipal Taxes =o Local tax (STW) Municipal Taxes ^o User Charge (STW)

Table 4.9: Overview on mapping types between TheSoz and STW.

Listing 4.5: Example for a simple cross-concordance in SKOS between TheSoz and STW.

```
<http://lod.gesis.org/thesoz/concept/10035317>
<http://www.w3.org/2004/02/skos/core#exactMatch>
<http://zbw.eu/stw/descriptor/16526-3> .
```

While the simple cross-concordances between STW and TheSoz have been transformed to SKOS easily, minor difficulties regarding the complexity of some mappings (e.g. mappings to combination of terms via the mapping type '+') and the restricted SKOS mapping properties could be observed. These observations are discussed in detail in Section 4.4.4.

4.4.3.3 Initial Case Study for Semi-automatic Link Detection

This section, parts of which have been presented in [MZJ$^+$11], covers how thesauri from different domains can be mapped automatically. Therefore, we reprise approaches made

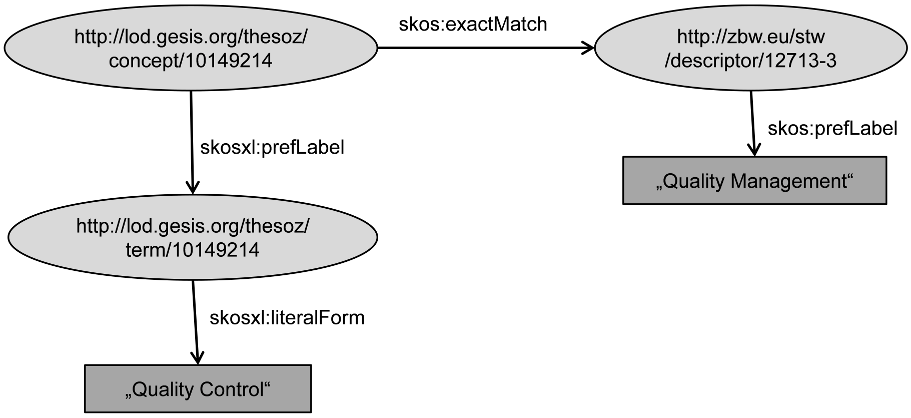

http://lod.gesis.org/thesoz/concept/10149214
skos:exactMatch
http://zbw.eu/stw /descriptor/12713-3
skosxl:prefLabel
skos:prefLabel
„Quality Management“
http://lod.gesis.org/thesoz/term/10149214
skosxl:literalForm
„Quality Control“

Figure 4.11: Example mapping between TheSoz and STW in SKOS.

in [LJC+08], where intellectual and automatic mapping approaches are compared and evaluated. In contrast to [LJC+08], we apply approaches on thesauri that are available in the SKOS format. Furthermore, we do not intend to evaluate the automatic approaches like, e.g. the Ontology Alignment Evaluation Initiative [OAE] already does. We aim to gain conclusions on the modelling of thesauri and mappings between them in SKOS as well as the quality and extent of identified mappings according to mappings done by domain experts. The Thesaurus for the Social Sciences (TheSoz) and the AGROVOC thesaurus are established knowledge organization systems (KOS) in their domains and by their scope, but they seem to have very little conceptual overlap.

The AGROVOC thesaurus contains more than 40,000 concepts in up to 21 languages covering topics related to food, nutrition, agriculture, fisheries, forestry, environment and other related domains. Currently, no direct links between both thesauri are available, but there are links via the Subject Headings Authority File (SWD) of the German National Library (DNB)[51]. It serves as a bridging vocabulary for available evaluation. These cross-walks are used in order to evaluate the results.

Both thesauri are available in the SKOS format and are freely available on the web. However, in order to detect possible direct links between both thesauri and to expose them into the LOD cloud diagram, this section intends to examine whether there are any good approaches for finding conceptual overlaps in thesauri from remote domains (semi-)automatically. Most efforts in developing and evaluating automatic alignment techniques have focused on ‘application-independent settings, where, typically, manually-built gold standards are created and used. Such gold standards are actually biased towards, at best, a single usage scenario (e.g. vocabulary merging), and can be of little use for other scenarios (e.g. query reformulation)’ [IMvdM+08].

Therefore, different approaches for aligning ontologies and linking data sources on the web are performed on both SKOS thesauri without processing or converting the thesauri

[51] http://www.dnb.de/EN/Standardisierung/Normdaten/SWD/swd_node.html

beforehand. The automatically generated matches, which should preferably be statements with properties `skos:exactMatch`, `skos:closeMatch` or `owl:sameAs`, are then evaluated by domain experts. Regarding possible matches between SKOS vocabularies (e.g. exact, close, related, broader and narrower matches),we focus on `skos:exactMatch` statements. Exact matches are considered at first because they deliver trustful links that can be reused by others for their applications. The matching results of all syntactic and semantic approaches are intellectually assessed concerning their mapping quality. Overlaps in the matching results of the different approaches are identified and interpreted. Furthermore, these mapping links extend the communication among the thesauri and bootstrap the linked data vision.

Our initial matching approach between TheSoz and AGROVOC is based on a syntactic algorithm, which uses the Levenshtein distance [Lev66] with a threshold of 0.21. This value has been chosen after preliminary tests with different values, where the number of detected links and the correctness of the links have been evaluated empirically. We can adapt it through the following the steps (see also Figure 4.12 for an overview on the workflow):

1. The selected thesauri are downloaded as SKOS resources from their respective websites.
2. A single triple store is created, with all SKOS triples coming from the thesauri. We use a Sesame[52] triple store since it is lightweight and open-source application.
3. Only (AGROVOC -> TheSoz) is considered.
4. For all possible pairs of concepts formed (the first concept coming from AGROVOC, the second one, from the other thesaurus), the following steps are carried out:
   a) only the preferred label is considered;
   b) the above similarity measure is applied;
   c) a threshold is applied for tuning the measure to find the matches
   d) mostly `skos:exactMatch`, and `skos:closeMatch` are considered at the initial stage in order to produce the trusted links.
5. All resulting candidate matches are loaded into a relational database and are then manually evaluated by a domain expert.
6. Candidate matches that are confirmed by the domain expert are then loaded in the sesame triple store.

The results have been evaluated by a domain expert and serve as a benchmark for testing the other approaches. There have been 1,613 alignments in our sample, of which 840 have been evaluated as correct `skos:exactMatch` statements and an additional six alignments as `skos:closeMatch`. This corresponds to a precision of 0.524. It was not possible to

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

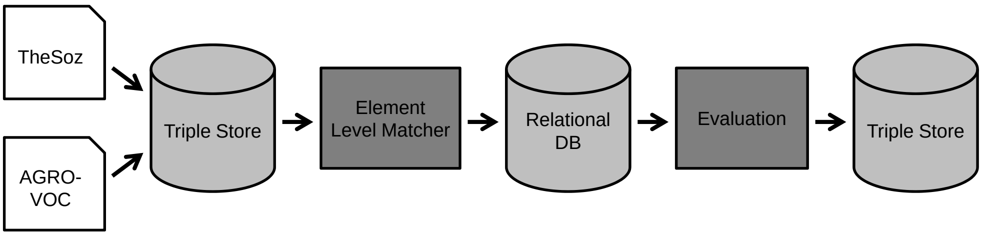


Figure 4.12: Matching process workflow.

Approach	Detected Matches	Correct Matches	Precision
Initial Approach	1613	846 (840 Exact Matches)	0.524
Silk (Levenshtein)	288	288	1
Silk (Norm. Levenshtein)	660	372	0.564

Table 4.10: Matching results generated by our initial approach and Silk.

compute the recall, since we are not aware of all the correct correspondences between both thesauri. Thus, no gold standard exists.

In order to compare the results of the initial algorithm, we also evaluate an existing link discovery tool. Link discovery tools have become popular in recent years because they identify links between the same instances in different Linked Data sets. Currently, several approaches are available, such as Amalgame [vOHdB11], SERIMI [AHSdV11] and Silk [VBGK09]. While SERIMI considers entity labels and structural context in order to detect links between instances [AHSdV11], no knowledge of the data set is required. Silk [VBGK09] detects links based on manually constructed rules that are described in a link specification language. This requires an extensive understanding of the data sets to be matched. Amalgame [vOHdB11] focuses on large SKOS-like vocabularies and aims to include the domain expert into an iterative alignment process, which has clearly increased precision and recall results. The task of instance matching has also been identified to be a relevant topic for the ontology alignment community [BMR11, SE11, EFvH^{+}11]. An instance matching task has been offered at the OAEI in recent years. However, only a few of the tools can be directly executed on mapping data in the SKOS format. Since most approaches were originally developed for aligning ontologies, especially those participating in the OAEI, the approaches either have to be adjusted or at least the thesauri have to be converted into OWL. The efforts involved in these adjustments have to be taken into account. For our experiment, we use Silk, because it offers various possibilities for configuring similarity measures and link discovery algorithms by the user. Also, Silk is able to process RDF without any preprocessing, either as a dump file or via a SPARQL endpoint.

In Silk, two approaches have been conducted. The results are presented in Table 4.10.

The first one used the Levenshtein distance like the initial approach, while the second one used the Normalized Levenshtein distance as a comparison operator. Both approaches in Silk have been performed with a threshold of 0.21 as the maximum distance. This makes the results comparable to the initial matching. The results above indicate a much lower amount of detected matches in Silk. But in contrast to the initial matching approach, all results delivered by the Levenshtein distance in Silk have been classified with a score of 100% and all of them could be evaluated as being correct (precision of 1). The Normalized Levenshtein distance delivered mixed results, which have been classified by Silk with more varying scores. 372 of the results could have been evaluated as being correct, which corresponds to a precision of 0.564. In Silk, we only defined exact matches to be detected. However, mappings with a lower score could theoretically be considered as close or related matches. This would require an additional evaluation by a domain expert. To identify relations other than equivalence ones is not that trivial for automatic approaches as [LJC$^+$08] has observed.

The results of this initial case study have encouraged our research on semi-automatic matching of thesauri.

4.4.3.4 The Library Track of the Ontology Alignment Evaluation Initiative

In order to further investigate the mapping results of the previous subsection, we decided to examine links between thesauri that are detected by ontology matching tools. This subsection summarizes the work presented in [AEE$^+$12, KZ13, KREZ14].

A major evaluation initiative in this regard is the Ontology Alignment Evaluation Initiative (OAEI), which started in 2004. Spanning various tracks from a wide range of different scientific disciplines, the main goal of this campaign is to improve ontology matching in general, by comparing and evaluating the different matching systems and algorithms. Participating in either a specific track or all tracks, these matching systems and algorithms are evaluated according to special criteria, e.g. the time spent in developing a set of mappings. Between 2007 and 2009, the OAEI included a so-called Library Track, directed towards KOS specifically applied in libraries (Isaac 2009). In 2012, the OAEI again offered a Library Track focused on the automatic matching of different domain-specific thesauri, co-organized by the authors of [KREZ14].

A key enabler for the OAEI Library Track was the availability of two considerably overlapping domain-specific thesauri in this case, the Thesaurus for the Social Sciences (TheSoz) and the Thesaurus for Economics (STW). Both thesauri are commonly used for indexing by domain-specific libraries and institutions providing information infrastructure, and so can be regarded as a real world-data set.

To make evaluation of the results possible, however, the organizers needed a reference set of mappings. During an earlier major terminology mapping initiative, a bilateral reference alignment between both thesauri was created manually by domain experts [MP08a]. It contains about 3,000 exact equivalences, 1,500 narrower and approximately 150 broader term relations. Since its initial creation in 2006, this reference alignment had not been updated. In recent years, however, the source thesauri have evolved and the changes were not reflected in the reference alignment. For the evaluation exercise, accordingly, an updated alignment would have been useful but, in its absence, only the established equivalence relations were used for validating the correspondences detected. This need, however, motivated subsequent investigation of whether the results could be used to update the existing alignment. In view of the large number of concepts, semantic relations and synonyms, the overriding aim of the evaluation was to show whether and to what extent the alignment of the two thesauri could be generated automatically. The question was whether current state-of-the-art matching systems developed for ontologies would be able to deal effectively with thesauri – the so-called ‘lightweight ontologies’ [UG04] that are widely used in practice. For the automatic creation of cross-correspondences, both thesauri needed to be available in a machine-readable format. Since OWL is used by almost all ontology matching systems, both thesauri had to be converted from their existing SKOS formats into OWL. General differences between ontologies and thesauri and a detailed description of difficulties, including the transformation from SKOS into OWL can be found in [AEE^{+}12].

Automatic Creation of Correspondences

For the automatic creation of correspondences all matching systems participating in the OAEI 2012 were applied: AROMA, ASE, AUTOMSv2, CODI, GO2A, GOMMA, Hertuda, HotMatch, LogMapLt, LogMap, MaasMatch, MapSSS, MEDLEY, OMR, Optima, ServOMapL, ServOMap, TOAST, WeSeE, Wmatch and YAM++ (see [AEE^{+}12] for more details). They match the ontologies and generate the resulting alignment by a fully automatic process. Our existing reference alignment made it possible to measure the quality of the alignments created. The results were evaluated by means of precision, recall and F-measure, where precision measures the correctness of the returned correspondences (i.e. the rate of all correct returned correspondences in regard to all returned correspondences), recall the completeness of the correspondences (i.e. the correct returned results in regard to all correct correspondences that should have been returned), and F-measure is the harmonic mean of both. An overview of the results can be found in Table 4.11 (matchers are sorted in descending order of their F-measure values). Altogether, 13 of the 21 submitted matching systems were able to create an alignment. Three matching systems (MaasMatch, MEDLEY, Wmatch) did not finish within the time frame of one week while five exited with an error.

This evaluation is based on the original reference alignment. It can safely be assumed that if the reference alignment had been up-to-date, many more correct correspondences would have been identified by each of the matchers. GOMMA performs best in terms of F-measure, closely followed by ServOMapL and LogMap. However, the precision and recall measures vary considerably across the top three systems. The choice of matcher for a given application would depend on whether high precision or high recall is preferred. If the focus is on recall, the alignment created by GOMMA is probably the best choice, with a recall of about 90%. Other systems generate alignments with higher precision,

Matcher	Precision	Recall	F-Measure	Time(s)	# of detected Mappings
GOMMA	0.537	0.906	0.674	804	4712
ServOMapL	0.654	0.687	0.670	45	2938
ServOMap	0.717	0.619	0.665	44	2413
LogMap	0.688	0.644	0.665	95	2620
YAM++	0.595	0.750	0.664	496	3522
LogMapLt	0.577	0.776	0.662	21	3756
Hertuda	0.465	0.925	0.619	14363	5559
WeSeE	0.612	0.607	0.609	144070	2774
HotMatch	0.645	0.575	0.608	14494	2494
CODI	0.434	0.481	0.456	39869	3100
MapSSS	0.520	0.184	0.272	2171	989
AROMA	0.107	0.652	0.184	1096	17001
Optima	0.321	0.072	0.117	37457	624

Table 4.11: Results of the OAEI Library Track 2012 [AEE+12].

e.g. ServOMap with over 70% precision, but most give lower recall values (except for Hertuda). Concerning the run-time, LogMapLt as well as ServOMap were quite fast with a run-time below 50 seconds. These systems are even faster than a simple Java program comparing the preferred labels of all terms. Thus, they are very effective in matching large ontologies while achieving very good results. Other matchers take several hours or even days and do not produce better alignments in terms of F-measure. A detailed discussion of the results can be found in [AEE+12].

Intellectual Evaluation of Automatically Created Correspondences

The use of a partial reference alignment to identify a good matcher is interesting, but does not solve the problem of updating and extending the reference alignment in an efficient way. Manually evaluating new correspondences took up to several minutes for each mapping established. Therefore, a good strategy is needed to maximize the number of new correct correspondences while minimizing the tedium of evaluating the matcher results. Unsurprisingly, the matching tools were easily able to detect matches based on the term alone, even in cases of small variations in the character string. For example, useful matches were often found between geographical and ethnographical terms. But the tools were less effective when taking the term's context into account. Incorrect matches were often generated when:

- The lexical value of the term was the same but broader and narrower terms showed the underlying concept to be different;
- The lexical value of the term was the same but the scope note in one thesaurus indicated an exclusion not valid in the other;

System	New Correspondences	Correct Correspondences
AROMA	15179	215
CODI	1756	162
GO2A	867	213
GOMMA	2180	246
Hertuda	2974	269
HotMatch	886	194
LogMapLt	1587	235
LogMap	818	201
MapSSS	475	63
Optima	424	38
ServOMapL	1016	251
ServOMap	682	230
WeSeE	1077	223
YAM++	1425	249

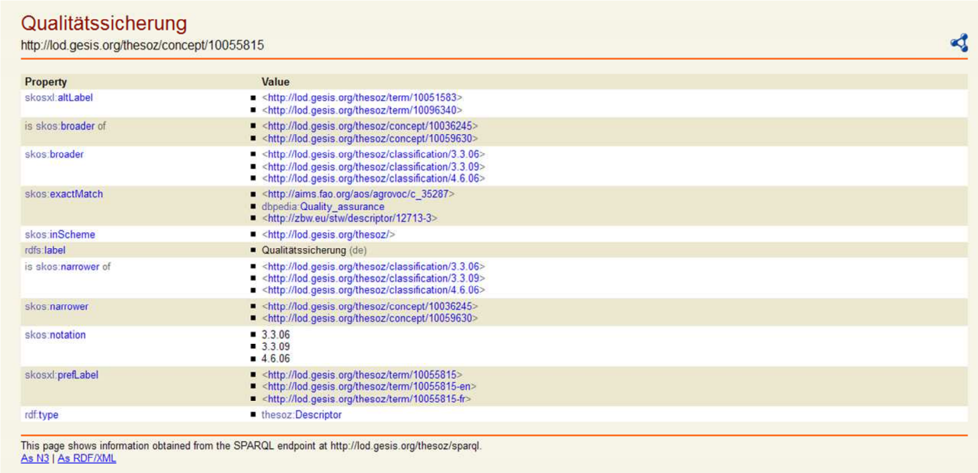
Table 4.12: Results of the manual evaluation [KREZ14].

- Terms in different domains looked similar, but their meanings were different;
- The presence of a synonym matching a preferred term in the other thesaurus caused an incorrect equivalence to be generated.

To sum up, the overall intellectual evaluation results of the newly established vocabulary mappings vary greatly between the different matching tools as shown in Table 4.12. The number of successfully established equivalence mappings ranged (approximately) between 40 and 270, i.e. between 6% and roughly 54% of the total correct number. Despite these promising results, it was judged that the alignments obtained were not precise enough for immediate use, since in a live situation every single cross-concordance has to be totally correct. Nevertheless, given the large number of matching systems and their fast, automated execution, they can be used to support domain experts in the creation of cross-concordances. Integrated in a semi-automatic workflow, they can serve as a recommender system, showing a domain expert the most probable cross-concordances and, hence, saving a huge amount of time.

4.4.3.5 Technical Implementation of Links

In order to provide interoperability between the participating thesauri and the external data sets, the thesauri and the cross-concordances between them have to be made accessible on the web in an integrated way. In [MZS10a], such a multi-thesauri setting has been established. The most common way is to publish them via multiple SPARQL endpoints according to assumed physically different storage locations, because thesauri are often held by different organizations. A Linked Data interface, e.g. the Pubby linked data frontend [CB07], is set upon these endpoints to generate a combined html representation of the different thesauri via dereferencing the URIs of the participating thesauri and their cross-concordances. The inclusion of mappings represents a stronger interlinking between the thesauri, which is not only based on a term or lexical level, i.e. established via `owl:sameAs`, but also on precise mappings between the concepts, e.g. exact, related, broader or narrower matches. Figure 4.13 depicts a screenshot of the TheSoz implemented in Pubby with exposed links to STW, AGROVOC and DBpedia [BLK+09]. The modelling of these links in TheSoz is shown in Figure 4.14.



Qualitätssicherung

http://lod.gesis.org/thesoz/concept/10055815

Property	Value
skosxl:altLabel	■ <http://lod.gesis.org/thesoz/term/10051583> ■ <http://lod.gesis.org/thesoz/term/10096340>
is skos:broader of	■ <http://lod.gesis.org/thesoz/concept/10036245> ■ <http://lod.gesis.org/thesoz/concept/10059630>
skos:broader	■ <http://lod.gesis.org/thesoz/classification/3.3.06> ■ <http://lod.gesis.org/thesoz/classification/3.3.09> ■ <http://lod.gesis.org/thesoz/classification/4.6.06>
skos:exactMatch	■ <http://aims.fao.org/aos/agrovoc/c_35287> ■ dbpedia:Quality_assurance ■ <http://zbw.eu/stw/descriptor/12713-3>
skos:inScheme	■ <http://lod.gesis.org/thesoz/>
rdfs:label	■ Qualitätssicherung (de)
is skos:narrower of	■ <http://lod.gesis.org/thesoz/classification/3.3.06> ■ <http://lod.gesis.org/thesoz/classification/3.3.09> ■ <http://lod.gesis.org/thesoz/classification/4.6.06>
skos:narrower	■ <http://lod.gesis.org/thesoz/concept/10036245> ■ <http://lod.gesis.org/thesoz/concept/10059630>
skos:notation	■ 3.3.06 ■ 3.3.09 ■ 4.6.06
skosxl:prefLabel	■ <http://lod.gesis.org/thesoz/term/10055815> ■ <http://lod.gesis.org/thesoz/term/10055815-en> ■ <http://lod.gesis.org/thesoz/term/10055815-fr>
rdf:type	■ thesoz:Descriptor

This page shows information obtained from the SPARQL endpoint at http://lod.gesis.org/thesoz/sparql.

As N3 | As RDF/XML

Figure 4.13: TheSoz implemented in Pubby.

4.4.4 Discussion and Limitations

The heterogeneous environment of various vocabularies worldwide can technically be harmonized by the use of SKOS and especially the content of traditional databases can be made accessible and connectible for applications of the Semantic Web, i.e. as LOD. Vocabularies in the SKOS format and the mappings between them can play a relevant role in this context by serving as a bridging hub for the inter-linking of different published and indexed data sets [Neu09].

The results of this section have revealed several limitations when modelling and converting thesauri to SKOS as well as while converting or identifying links between multiple thesauri. However, there are also benefits for domain experts who maintain or generate links between thesauri.

4.4.4.1 Transformation to SKOS

Even if a thesaurus meets established ISO norms, a conversion to SKOS is not always as trivial as expected, which is proved by several case studies [vAMMS06, De 09, Neu09, PZ09, MZS10a, Mal07]. All of these studies have encountered one of the main obstacles

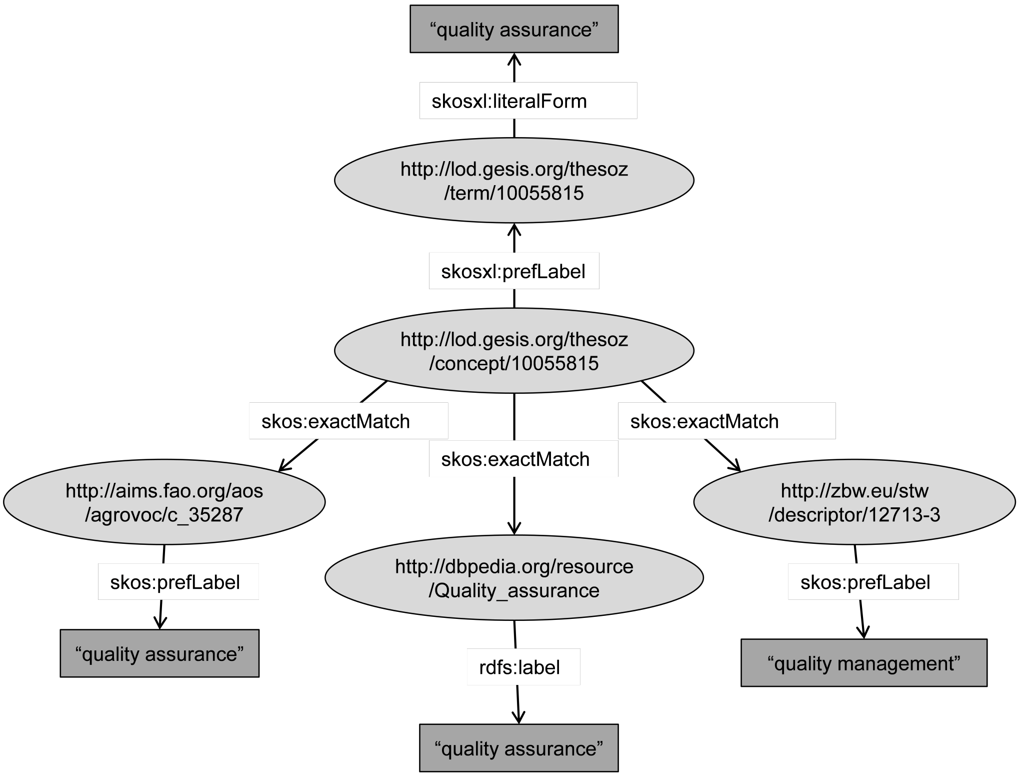

"quality assurance"
skosxl:literalForm
http://lod.gesis.org/thesoz /term/10055815
skosxl:prefLabel
http://lod.gesis.org/thesoz /concept/10055815
skos:exactMatch
skos:exactMatch
skos:exactMatch
http://aims.fao.org/aos /agrovoc/c_35287
http://dbpedia.org/resource /Quality_assurance
http://zbw.eu/stw /descriptor/12713-3
skos:prefLabel
rdfs:label
skos:prefLabel
"quality assurance"
"quality assurance"
"quality management"

Figure 4.14: Links from TheSoz to STW, AGROVOC and DBpedia.

raised by the conversion of the TheSoz: Some of the relations of the original thesaurus cannot be modelled adequately in SKOS. For TheSoz, this is the representation of compound relations and concepts, which is also an issue with [vAMMS06] and [De 09]. This obstacle can be observed regularly because compound concepts are part of the ISO 2788 standard.

For some modelling issues, the SKOS Primer provides an overview of correspondences between the ISO norms 2788 and 5964 and SKOS [IS09a]. Regarding the compound equivalences stated as syntactical composition of terms, it is suggested to define the personal extensions of either `skos:Concept` or `skosxl:Label`. We have defined subclasses of the latter class and defined extensions of the `skosxl:labelRelations`, which allow the representation of compound equivalences. De Smedt [De 09] applied these relations similarly for the EUROVOC thesaurus. Neubert [Neu09] has defined compound equivalences as an additional construct called `zbwext:useInsteadNote` as a subproperty of `skos:note`, which holds information about what term to use instead.

Since SKOS is based on RDF, it is easy to define additional classes and properties, but such self-defined structures can lead to inconsistencies and incompatibilities with respect to other SKOS data sets or with applications for processing data in the SKOS format, e.g. using SKOS thesaurus management tools. Therefore, extensions should be described using standard classes and properties, e.g. with RDF Schema, so that the data is at least processable in a minimal way to ensure maximum compatibility.

Additional challenges of SKOS that TheSoz has not addressed have been observed during various case studies. Malaisè [Mal07] has identified drawbacks by modelling special types of relations like a 'linked term relationship', by defining subconcepts of concepts and by adding qualifiers to concepts. Adding multiple and alternative notations as well as including a semantically meaningful order of concepts has been observed by [PZ09] as a challenge because they are crucial elements for thesauri.

The SKOS standard has been well received in the community and many open issues are currently being discussed. The work on 'ISO-Norm 25964: Thesauri and interoperability with other vocabularies' [NIS11] implies a step in the right direction in order to cover more specific issues of thesauri structures.

4.4.4.2 Links between Thesauri

Although SKOS provides a standard model for representing vocabularies, transformed or converted thesauri can be quite different due to varying complexity and heterogeneous structure. Modelling mostly term-based thesauri in a concept-based way can be implemented differently. One reason for inconsistencies is that the given cross-concordances were defined on term-based thesauri, but the SKOS versions of those thesauri are concept-based. Therefore, cross-walks between traditional thesauri cannot simply be adapted to the SKOS mapping properties under certain conditions. It has to be examined whether the two terms of a given cross-walk represent adequate concepts in the corresponding SKOS versions by e.g. being used as `skos:prefLabel` in a concept. In the case of the cross-concordances defined in the KoMoHe project, they were defined only between preferred terms, which means that a conversion to SKOS should be feasible without further complications. In general, if the above-described requirements are met, it should be relatively easy to transform existing cross-concordances to SKOS.

For the case that there are cross-walks between non-preferred terms, each participating SKOS vocabulary has to be checked regarding how non-preferred terms are modelled because the mapping properties of SKOS can only be applied between concepts. If non-preferred terms are not represented as concepts, cross-walks between them cannot directly be modelled in SKOS. This is usually the case because these terms are typically modelled as `skos:altLabel`. Although ISO 5964 allows relations between non-preferred terms, it is not possible in SKOS unless SKOS-XL extensions are defined and used.

Another aspect that cannot be represented in SKOS, – except by adding a comment `rdfs:comment` or `skos:note` –, is information on the relevance of a mapping. Although this is very specific information derived from the mappings of the KoMoHe project, it is relevant, especially when dealing with 1:n relations.

Domain-specific differences in thesauri can also cause conversion problems. For example, a term or concept in one thesaurus can correspond to a combination of two terms or concepts in another thesaurus. The mapping properties of SKOS do not allow such single-to-multiple relations (neither for one language nor for multiple languages). Therefore, personal extensions to the standard are required once again.

Cross-concordances can occur in a complex manner as associate relations between the terms of one vocabulary. But the mapping properties of SKOS are too restrictive in their current definition; hence, alternative possibilities, i.e. defining personal extensions, have to be defined on how to deal with these special use cases. Thus, the mapping between the terms of TheSoz and STW shown in Listing 4.6 cannot be directly represented in SKOS. A `skos:labelRelation` between two or more labels of different classes of `skos:ConceptScheme` would have to be defined.

Listing 4.6: Compound mapping between TheSoz and STW.

```
Electronic Government (TheSoz)
=+ Public Administration + Internet (STW)
```

Transforming existing vocabularies and thesauri to SKOS remains a complex issue according to the heterogeneous structure of the involved vocabulary. Especially the SKOS conversion of given cross-walks that have a term-based origin can bear major problems when the participating terms are not the preferred terms that would usually be represented as concepts in SKOS. In that case and in case of the requirement of semantically more complex relations, i.e. *USE COMBINATION* relations between the terms of different vocabularies, extensions have to be defined if the relevant information of the cross-walks is to be preserved.

Semi-automatic approaches for identifying mappings between thesauri deliver promising results as the complexity of possible algorithms increases beyond calculating lexical similarities [SE11]. However, regardless of whether two thesauri share an overlap in concepts or not, a manually evaluation by domain experts remains necessary. This is also required if mappings other than exact matches are desired to be identified. Then, at least the identified matches with a lower confidence have to be considered additionally. We will continue research on mapping thesauri in the future. Since 2012, the TheSoz has been participating in the Library Track at the OAEI together with the STW thesaurus.

4.4.4.3 Benefits of Using Semi-automatic Matching Procedures for Building up Vocabulary Cross-walks

Based on our findings in Section 4.4.3.4, we can conclude that thesauri published as Linked Open Social Science Data and the execution of ontology matching tools can be used to support the work of domain experts. By optimizing the workflow, these methods promise to facilitate sustained updating of high-quality vocabulary cross-walks. The following subsection summarizes the results presented in [KREZ14].

In an experiment [KREZ14], we investigated whether the effort of a domain expert during manual evaluation can be reduced and optimized. The underlying assumption of this approach is that the more matching systems have found a certain correspondence,

	All correspondences (including duplicates)	De-duplicated correspondences
Total number	55466	22592
of which are correct	21541	2484 (11%)

Table 4.13: Number of correspondences: total; de-duplicated and correct [KREZ14].

the more likely it is correct. Additionally, we investigated whether a reorganization of the results presented for manual evaluation had an impact on the time spent by domain experts. We tested this assumption on the results of the OAEI Library Track 2012 [AEE+12]. This experiment addressed the order and the number of detected correspondences that the domain expert had to consider. Any duplicate correspondences (i.e. correspondences generated by more than one matcher) were removed. After de-duplication, the correspondences were grouped according to the number of matchers detecting them. This resulted in a group containing correspondences that were found by all 13 matching systems, a group with correspondences found by 12 matchers and so on. The last group contained correspondences found by only one matcher.

In the experiment, the groups were presented to the domain expert for evaluation in descending order, i.e. the expert began with the group of correspondences found by all the matching systems. From the total numbers of correspondences and of those that turned out to be correct, we can observe the rate of finding correct correspondences and compare that with the rate when no reordering of the results was done. In other words, the calculation shows how many correct correspondences would be found after evaluating the same number of correspondences as before.

In Table 4.13, the results of the manual evaluation are summarized. For our experiment, only the de-duplicated correspondences were considered.

In Figure 4.15, we illustrate the percentage of correct correspondences (y-axis) found by a certain number of matching systems (x-axis). For example, x = 9 means that these correspondences were identified by nine matching systems, regardless of which nine systems found them. Above the graph, the total number of detected correspondences for x systems is indicated (71). Altogether, 71 correspondences were found by all matching systems, of which ~99% proved correct. Of the correspondences found by 12 matching systems (209), about 93% were found to be correct. The graph clearly shows a correlation between the number of matchers to identify a given correspondence and the likelihood of its being correct.

Table 4.14 shows the number of all correspondences and the numbers of all correct correspondences, grouped by the number of matchers that found these correspondences. For example, 506 correspondences were found by ten matching systems and 409 of them (80% approximately) were correct.

These numbers confirm our assumption that the more matching systems have found a certain correspondence, the more likely it is to be correct. This ‘majority vote’ method

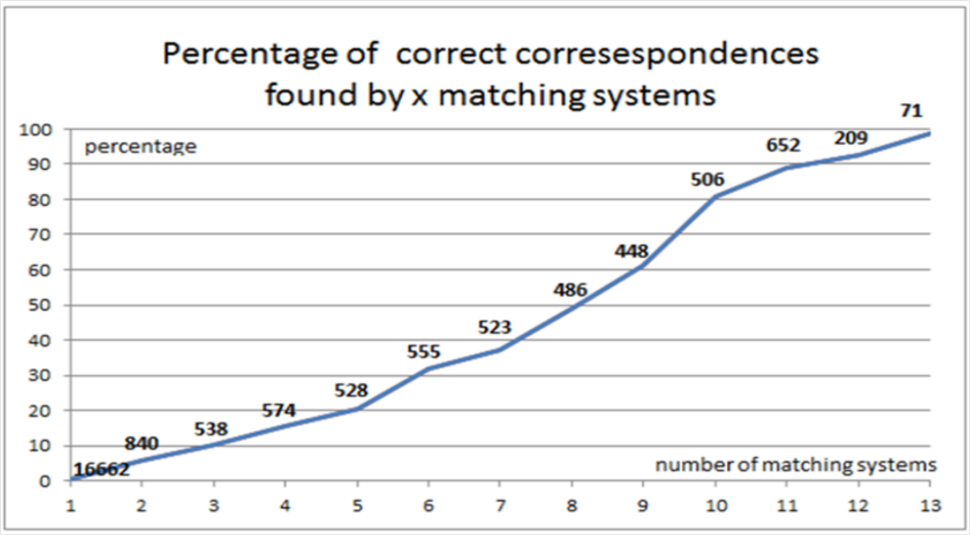


Figure 4.15: Percentage of correct correspondences found by # matching systems [KREZ14].

Number of corresponding matchers	Number of all correspondences	Percentage of correct correspondences	Number of correct correspondences
1	16662	0.27007562	50
2	840	5.71428571	48
3	538	10.4089219	56
4	574	15.6794425	90
5	528	20.4545455	108
6	555	31.8918919	177
7	523	37.0936902	194
8	486	48.8659794	238
9	448	61.3839286	275
10	506	80.8300395	409
11	652	89.1104294	581
12	209	92.8229665	194
13	71	98.5915493	70

Table 4.14: Results of the majority vote [KREZ14].

has already emerged as a promising technique, e.g. for combining different ontology matching systems [EMS09]. Regarding the time spent by users during manual evaluation, our results confirm that at least a certain number of correct correspondences can be found relatively quickly by optimizing the sequence of entries in the list of matches. To show the extent of the efficiency gain, the first five columns of Table 4.15 reverse the sequence of Table 4.14, beginning with those correspondences that were found by as many matchers as possible. This reveals how many correct correspondences can be found at each stage, if the list is reorganized. The percentages of correct correspondences are also shown for each group of matchers. Finally, in the last two columns, we compare these numbers to the numbers when the evaluation is not optimized. The number of corresponding matchers (column 1) was not taken into account. The overall correctness rate of 11% (see Table 4.13) was used to estimate the number of correct correspondences shown in Column 6. This shows the number of correct correspondences that would have been found after checking the same number of candidates as were checked at the corresponding stage of the optimized process.

In summary, a critical mass of correct correspondences can be detected faster by reordering the results for manual evaluation. For example, after having evaluated 1,886 correspondences a total of 1,529 correct correspondences were found in the optimized scenario (i.e. 61.5% of all correct correspondences), while only 207 correct correspondences would have been found without optimization (only 8.33% of all correct correspondences). Nevertheless, if it is necessary to find all the correct correspondences, all results of all matchers must eventually be evaluated.

Under the requirement of the availability of thesauri published as Linked Open Social Science Data, our study has shown that the use of ontology matching tools can greatly speed up the process, especially if the work is organized in the most time-efficient order. This enables automatic creation of an alignment between different thesauri that are available in machine-readable format. The most recent OAEI Library Track has shown significant differences between the performances of various ontology matching tools on offer. Some are rather promising. None of them, however, could alone prepare a high-quality vocabulary cross-walk. As a first conclusion, it was judged that the matching tools could be used in recommender systems. Secondly, the matches generated by a variety of different tools were combined and presented in the most time-efficient order, so as to speed up the intellectual evaluation of the matches. This proved to be highly effective. However, more research could be useful in the provision of automated support for intellectually verified matching procedures. Knowledge organization systems like thesauri are built with elaborate semantic content and structures. The challenge of achieving interoperability between them is an intellectual task that cannot easily be emulated by automatic means. That is why further research could usefully study the interplay between process-supporting technical solutions and intellectual demands.

			Optimized scenario		Normal evaluation	
Number of corresponding matchers	Number of all correspondences	Percentage of all correspondences (22,592 = 100%)	Number of correct correspondences	Percentage of all correct correspondences (2,484 = 100%)	Number of correct correspondences (estimated)	Percentage of all correct correspondences (2,484 = 100%)
13	71	0.31%	70	2.82%	8	0.32%
12	280	1.24%	264	10.63%	31	1.25%
11	932	4.13%	845	34.02%	103	4.15%
10	1438	6.37%	1254	50.48%	158	6.36%
9	1886	8.34%	1529	61.55%	207	8.33%
8	2372	10.50%	1767	71.14%	261	10.51%
7	2895	12.81%	1961	78.95%	318	12.80%
6	3450	15.27%	2138	86.1%	380	15.30%
5	3978	17.61%	2246	90.42%	438	17.63%
4	4552	20.15%	2336	94.04%	501	20.17%
3	5090	22.53%	2392	96.30%	560	22.54%
2	5930	26.25%	2440	98.23%	652	26.25%
1	22592	100%	2490	100%	2485	100%

Table 4.15: Comparison of different evaluation strategies [KREZ14].

4.5 Summary

In this chapter, we have applied Semantic Web standards and technologies for publishing Linked Open Social Science Data completely. We have investigated how such data can be published according to the Linked Data principles [BL06] and what special requirements have to be fulfilled for this purpose. In order to enable a complete publication of all aspects of Linked Open Social Science Data, i.e. processes, data and structures, we adopted and extended existing ontologies and vocabularies namely, SWRC [SBH$^+$05] and SKOS [MB09b] in order to be able to represent particular processes and structures of Social Science Data, like the inclusion of research data into the research process. We have also overcome current limitations and developed an ontology for representing person-level data, the DDI-RDF Discovery vocabulary. Until now, there was no vocabulary for representing this kind of data in such a detailed and complex way. With these three examples, we cover all the aspects of processes, data and structures of Linked Open Social Science Data.

The results of this chapter contribute to the fifth block of research question (see Section 1.3). The publication of data as Linked Data completely and to the fullest extent is a necessary requirement and a foundation for any further processing, use and consumption **(5b)**. Following this observation, the results of this chapter are also an essential step towards the investigation of methods for matching such data. The next chapter focuses on data matching, particularly on matching Statistical Linked Data. We will investigate whether Semantic Web technologies can be applied to data matching tasks and what limitations can be identified and addressed.

5 Data Matching for Published Linked Open Social Science Data

The publication of Linked Open Social Science Data is the first step to its meaningful consumption. According to the expert interviews in Section 3.1, data matching with two or more data sets is a typical processing step in social research and is necessary for integrating or merging data sets to ensure a combined or comparative analysis. With data matching, the required similarities between these data sets can be identified, e.g. corresponding schema elements that can serve as key variables for merging or instances like entries of code lists.

In the field of databases and data warehouses, schema matching has a long tradition [BBR11] and is also an active research field in the Semantic Web, jointly with ontology matching [ES07]. However, Semantic Web-based ontology and schema matching approaches have not yet been applied to Statistical Linked Data. Since this kind of data is typically used for scientific analysis and, therefore, the merging or integration of multiple data sets is often required, schema and ontology matching seem to be reasonable methods for supporting this task. However, matching systems are not trained against specific characteristics in the structure and semantics of Statistical Linked Data, since no similar data is currently being used in established benchmarks and evaluation campaigns [EFvH$^+$11]. Vice versa, popular statistical tools are yet not capable of processing Linked Data, although initial approaches have emerged recently [vHEM12].

In this chapter, we investigate methods and approaches for data matching with Linked Open Social Science Data, as it is the main objective of this thesis. Since the publication of Linked Open Social Science Data was enabled in the previous Chapter 4, we can now adopt the topic of interest, ‘data matching’, as determined in Section 3.3 according to our use case of ‘Analysing Research Data’ (see Section 1.2). We will focus on Statistical Linked Data, since it is a subpart of Linked Open Social Science Data (see Section 2.2). We will fill the gap between established ontology and schema matching approaches of the Semantic Web and the requirements of Statistical Linked Data. In Section 5.1, we outline the current situation and problems regarding the matching of Statistical Linked Data and discuss why it has still not been fully considered. We also introduce three matching methods that address particular challenges that accompany Statistical Linked Data. The three approaches are presented in the following sections in detail. We define assessment tests for Statistical Linked Data in Section 5.2, which is based on [ZM11a]. These tests provide researchers a first insight into whether the particular data set is suitable for matching according to particular basic requirements and whether two data sets can be matched. Section 5.3 proposes a method for instance-based schema matching using regular expressions that are reasonable for summarizing the instance values of schema elements. This method is based on [ZZS12]. In order to consider object properties that link to entries of code lists and classifications, we present a method for object property matching, where the overlap between these imported code lists and classifications is utilized. This approach is presented in Section 5.4 and is based on [ZM14]. Finally, we summarize the results of this chapter in Section 5.5

5.1 Matching Statistical Linked Data

When analysing research data, multiple data sets are often compared or merged in order to investigate particular correlations between indicators or variables [SHE05], e.g. the relationship between unemployment ratios and fear of losing one's job. In order to merge or integrate such heterogeneous data sets their schemata have to be matched. When considering Statistical Linked Data as a data source for this task, established Semantic Web-based schema and ontology matching approaches are theoretically a reasonable choice for conducting the matching process. However, Statistical Linked Data has previously not been considered in established matching benchmarks and campaigns, although large potential and challenges are seen in including new types of information resources and domain-specific constraints [SE11] for ontology matching in order to improve the performance and results of current tools. According to [Hal05], especially domain-specific values, significant occurrences of values and patterns of values are stated as relevant characteristics to be considered at the instance level, as are integrity constraints for schema elements and their instance values. In turn, the increasing amount of Statistical Linked Data sets determines a true value addition for users providing that these data sets can be matched and processed with other data sets. As presented in Section 2.2, Statistical Linked Data holds a distributed structure that originates from its source in data warehouses. We will analyse additional special characteristics and patterns of Statistical Linked Data in Sections 5.3 and 5.4, which are not covered in existing benchmarks and evaluation campaigns. Existing matching systems are not trained against these particular characteristics because there is no reason for developers to consider problems for matching that are not evaluated in the benchmarks. However, statistical tools that are programmed to process statistical data (including data matching) are still not able to process Statistical Linked Data, although first approaches in this direction exist [vHEM12].

In the following sections, we investigate the challenges of matching Statistical Linked Data and why these have still not been fully considered. We introduce three methods that address particular challenges of matching Statistical Linked Data. With the development of these methods, we refer to the research questions raised in Section 1.3, particularly to the first four questions. In addition, these methods are located in our use case 'Analysing Research Data' introduced in Section 1.2. Although motivated by and focused on Statistical Linked Data, these methods can be also be applied to other Linked Data sets that hold a similar composition (e.g. data originated from relational databases) or similar characteristics and patterns. The three methods developed in this chapter are the following:

Assessment Tests for Statistical Linked Data Before the actual data matching process is conducted, it may be helpful to know whether a Statistical Linked Data set is suitable for a scientific analysis as well as whether two data sets can be matched semantically and technically. Considering large data sets and users with less technical expertise, this first insight cannot be gained by inspecting the RDF representation of the data because RDF, e.g. serialized in XML, may become very extensive and confusing. With the definition of assessment tests for Statistical Linked Data, we allow for testing data sets according to particular requirements, i.e. in the case of statistical data, whether observation values, dimensions, etc. can be identified and whether they may correlate to those of another data set. The assessment tests are presented in Section 5.2.

Instance-based Schema Matching utilizing Regular Expressions In Statistical Linked Data sets, the instance values per schema elements can be very similar, e.g. numerical values describing temporal information and alphanumerical codes of particular code lists. This characteristic can be leveraged for matching the schema elements. In an instance-based schema matching approach, we utilize regular expressions for summarizing the instance values of a particular schema element. Although the focus is on the instance values of datatype properties, this method is generally able to consider classes of object properties by inspecting the URI paths. This approach is presented in Section 5.3.

Object Property Matching utilizing the Overlap between Imported Ontologies To achieve a complete consideration of characteristics of Statistical Linked Data, the third method supplements the instance-based matching approach and focuses on object properties in particular. This method has been motivated by the distributed structure of Statistical Linked Data, which holds many object properties with links to entries of code lists and classifications. For these entries and corresponding additional information stored inside the code lists, a summarization with regular expressions is insufficient. Hence, we present a method that computes the overlap between the code lists and classifications that are linked to by object properties. This approach is presented in Section 5.4.

5.2 Assessment Tests for Statistical Linked Data

Once potentially relevant research data has been found, a researcher has to decide whether the data is technically and semantically suitable for further scientific analysis, e.g. for processing using statistical tools or whether it has to be preprocessed first. Due to the complex structure of most statistics, comparing and integrating data from different data sources is a time-consuming task. Not only different schema elements, but also different code lists and their entries may have to be aligned with each other. While this alignment can be generated easily for schema elements when the data is stored in spreadsheets, it may get more complicated and confusing when only raw RDF/XML data is available, especially when the researcher is not familiar with the technical format. Statistical Linked Data sets can theoretically provide more meta-information about the data itself than a spreadsheet. There may be additional links to other data sets or further aggregated information about a single element may be available at its own URI. Thus, the complexity of structure and semantics enhances the analysis of such data and makes it more technical. In this section, we present assessment tests that analyse the chosen data regarding its validity in terms of Statistical Linked Data. Thereby, we focus on schema information and instance patterns. This provides an initial insight regarding the decision of whether the data is suitable for scientific analysis.

The tools currently in use are mainly validation services[1] for a general validation of RDF or OWL data concerning data modelling and logical aspects or of compliance with the Linked Data principles (e.g. the Vapour[2] [BFF08]). But when Linked Data is used for scientific analysis, further assessment tests are required than current approaches can carry out [GGSL12, MMB12, HZ09]. Of special interest in this regard is the comparability between heterogeneous data sets and the identification of common characteristics, such as the time range and geographical region of the data. The identification of provenance and other circumstances of the data are also relevant, such as base population, observation intervals and the nature of the sample used. All of this information supports researchers in making an educated decision about which data to use and how.

The assessment tests defined in this section support researchers during their decision process on how relevant and useful a specific Linked Data resource might be for a scientific statistical analysis and whether further technical preprocessing is necessary before using it. They provide insight regarding whether two data sets technically and semantically (in terms of their dimensional coverage, e.g. their temporal and geographical coverage) fit together. We have implemented these tests in a web-based prototype application that is capable of extracting information from Linked Data sets. We exploit both, domain knowledge and the inherent semantic annotations of data sets, by scanning them for known patterns that signify e.g. typical numerical data blocks or potentially temporal and geographical dimensions. Thus, researchers are supported in judging usage like detecting observation values, different dimension, etc. of the data. Additionally, two data sets can be analysed together in order to detect possible similarities or conflicts between them. The implementation is evaluated with real-world statistical data sets. The results provide not only information on potential usage of the data, but also on differences and difficulties in data modelling aspects related to the problem of schema matching.

Section 5.2.1 defines basic data requirements for valid statistical data that are necessary for the assessment tests. Out of these requirements, we formulate rules for assessing Statistical Linked Data in Section 5.2.2. The resultant assessment tests are presented

[1] see for instance validation services at http://www.w3.org/RDF/Validator/, http://www.mygrid.org.uk/OWL/Validator or http://owl.cs.manchester.ac.uk/validator/

[2] http://vapour.sourceforge.net/

in Section 5.2.3. In Section 5.2.4, we describe the technical implementation of these tests. The assessment tests are then evaluated with real-world Statistical Linked Data in Section 5.2.5. In Section 5.2.6, observations and limitations made during the tests are discussed in detail.

5.2.1 Basic Requirements for Valid Statistical Data

As part of the definition of assessment tests, basic requirements regarding particular data features have to be formulated. This is necessary in order to examine whether the chosen Linked Data is valid statistical data. For this purpose, we adopt the basic components of statistical data as defined in the guidelines of the SDMX model [SDM09], which is a widely accepted standard information model for statistical data.

According to [SDM09], the minimal components for a *statistical data* and a *data set* are the following:

> ‘Statistical data are data derived from either statistical or non-statistical sources, which are used in the process of producing statistical products.’
>
> ‘a data set can be understood as a collection of similar data, sharing a structure, which covers a fixed period of time. A data set is any permanently stored collection of information usually containing either case level data, aggregation of case level data, or statistical manipulations of either the case level or aggregated survey data, for multiple survey instances’

The term *data* itself in this context is defined as

> ‘Characteristics or information, usually numerical, that are collected through observation.’ [SDM09]

The data item inside each *observation* is, therefore, also called observation value. Additionally, inside each observation, a number of particular statistical concepts are used as dimensions to identify data. In [SDM09], a *statistical concept* is ‘a statistical characteristic of data’ and a dimension is defined as ‘a statistical concept used, in combination with other statistical concepts, to identify a statistical series or single observations’. Typical dimensions are e.g. a fixed period of time or a geographical area associated with the observation. Finally, the measure determines which phenomenon has been observed, e.g. the unemployment rate. Based on this foundation, we formulate the following basic requirements for valid statistical data.

Requirement 1: Observation Values According to [SDM09], the *observation value* is a ‘value of a particular variable at a particular period. The observation value is the field which holds the data.’ Since the data itself is the key part of statistical data, our first requirement is that the data set has to hold at least one observation value. Inside the RDF data, the observation values should be literals, since their representation is alphanumerical.

Requirement 2: Measure The *measure* of a data set is the 'phenomenon or phenomena to be measured in a data set' [SDM09]. The 'instance of a measure is often called an observation' [SDM09]. The measure is one of the key information of statistical data, since it determines the content of the data. Thus, the measure is the second requirement for our assessment tests. It can be expected that the measure is also a literal inside the RDF representation of the data set.

Requirement 3: Dimensions The third requirement for our assessment tests is knowledge about the dimensions of the data set. Since a *dimension* is 'a coded statistical concept used (most probably together with other coded statistical concepts) to identify a time series, e.g. a statistical concept indicating a certain economic activity or a geographical reference area' [SDM09], it is necessary information for scientific analysis. As the information on the time period or the geographical area associated with the data set or with single observations is relevant, we focus especially on these two dimensions. In the SDMX guidelines [SDM09], these dimensions are called *REF_AREA* and *REF_PERIOD*. The *REF_AREA* is 'The country or geographic area to which the measured statistical phenomenon relates' [SDM09], while the *REF_PERIOD* determines 'The period of time or point in time to which the measured observation is intended to refer' [SDM09]. According to [SDM09], both can be represented as free text inside an observation that would result in a literal in a RDF representation. However, *REF_AREA* can also be represented by a code list with respect to its entries and *REF_PERIOD* by a date/time stamp. For both dimensions, the latter solution is preferred to free text.

To summarize, we can conclude that the basic requirement for a valid statistical data set with regard to our purpose is the occurrence of 1 to n *observation values*, one *measure* and 1 to n *dimensions*. In the context of Statistical Linked Data, these requirements are independent of the vocabulary used to represent the data. Our requirements are kept at a minimum, which is justified by the current quality and extent of Statistical Linked Data. This is mostly a result of the extent of the openly published data source, which underlies the RDF representation of statistical data. Only a few data sets hold a very detailed description about the data itself, e.g. regarding its attributes, measures and dimensions, or information about acquisition and provenance. Especially the latter information as well as details about variance and bias in the data are highly relevant for judging the statistical quality and possible usage of the data. In this section, we do not address such issues, but will focus on the pragmatically relevant data items, i.e. observation values and dimensions.

5.2.2 Rules for Assessing Statistical Data

In order to test Statistical Linked Data according to our basic requirements technically, we formulate the following rules. Since we also plan on assessing two data sets regarding the possibility of whether they can be matched for a comparative data analysis, we also define rules for comparing Statistical Linked Data. The data will be retrieved with

SPARQL queries and analysed in the JSON format. Thus, we will address elements inside the JSON data structure of our rules.

Observation Values For detecting the observation values, we determine the following rule. All numerical values inside the data set have to be considered. There may also be special characters like `.`, `,` or `%`.

- Pattern to be inspected inside each value of the `o` element with the type `literal`: `^\\-*[0-9]+\\.*\\,*[0-9]*$`

Measure For identifying the measure of a data set, we formulate the following rule. Since this representation can be free text, all labels inside a data set have to be considered. This is done by focusing on suitable properties.

- Pattern to be inspected inside each value of the `o` element with the type `literal`: `[dc:title, rdfs:label, skos:prefLabel]`

Dimensions For detecting the dimensions of a data set, we determine the following rule.

- Each observation has been collected at one or more particular dimensions, i.e. a particular dimension is coupled with every single observation value inside a data set (occurring with different values). Together with each triple in JSON that contains the earlier detected observation value, there has to be another triple with the same subject element `s` and a particular element `p` with an value of the type `uri` or `literal`. A triple of this form has to be coupled with every triple containing the observation value.

Geographical Dimension In order to identify more information on a geographical dimension (if available), we formulate the following rule. The geographical dimension may be coded with an entry of a code list like ISO 3166[3], OECD countries[4] and NUTS[5] for coding regions within Europe. However, even if the value of the dimension is described as free text, the pattern of the entry may be similar to one of the code lists.

- For each triple that is assumed to be a dimension, the element `o`[6] is inspected in accordance to codes of the existing geographical code lists.

[3] http://www.iso.org/iso/home/standards/country_codes.htm

[4] http://stats.oecd.org/glossary/detail.asp?ID=461

[5] http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction

[6] Since entries of code lists may be represented as URIs, we consider both types of the element `o`, the type `literal` and the type `uri`.

Temporal Dimension In order to detect more information on a temporal dimension (if available), we determine the following rule. The temporal dimension may be coded with a date/time stamp. However, even if it is described as free text, the pattern of the entry may be similar to one of the code list containing the date/time stamps.

- For each triple that is assumed to be a dimension, the element `o` is inspected in accordance to patterns of existing date/time stamps.

These rules form the basic part of our assessment tests that will be defined in the next section.

5.2.3 Definition of Assessment Tests for Statistical Linked Data

Based on the rules for assessing Statistical Linked Data, we define the following assessment tests. These tests focus on extracting information from dedicated Statistical Linked Data sets and on detecting matching possibilities between two of them. They consist of two stages, each containing multiple packages with different tests. The first stage, A, focuses on the identification of elements inside a data set. Thus, the packages are developed directly from the rules formulated in the previous section. In the second stage B, it is inspected whether two data sets can be matched. This is done by transferring the results of Stage A and comparing particular elements of the two data sets with each other. Table 5.1 provides an overview of the defined assessment tests.

Stage	Package	Description
A		Identification of Data Elements
	A1	Observation Values
	A2	Measures
	A3	Dimensions
	A3.1	Geographical Dimension
	A3.2	Temporal Dimension
	A4	Refinement
B		Data Comparison
	B1	Similar Data Elements
	B2	Different Data Elements

Table 5.1: Overview of assessment tests.

In the following paragraphs, a detailed description of the two stages and their packages is given.

Stage A: Identification of Data Elements The first stage of the checks identifies the existence of necessary elements in the data set. The included packages are directly based on the rules for assessing Statistical Linked Data. Literals and URIs contained in the data set are extracted by an algorithm and analysed in Packages A1 to A3.2.

Package A1 covers the identification of observation values, where the data is searched for included numbers and digits that may be suitable as observation values. Package A2 identifies one or more measures and their labels, which fit the detected values. In Package A3, dimensions and their labels are detected from the data. Since geographical and temporal dimensions are frequently used in statistical data, their occurrence is inspected in Packages A3.1 and A3.2. Instances of the temporal dimensions, i.e. dates, may also be assumed as observation values because they may resemble single numbers, e.g. 2004, 1992. Thus, Package A4 refines the results of A1 based on the knowledge gained in Package A3.2. The results of Stage A are multiple one-dimensional arrays (one per package). An example output is depicted in Listing 5.1.

Listing 5.1: Example output of Stage A.

```
Following Observation Values have been identified:
[10839905, 7761049, 10532770, 64369147, ...]

Measure of data set: [Total population  ]

Following temporal dimension has been identified:
[http://purl.org/dc/terms/date]

Geographical Dimension:
[http://ontologycentral.com/2009/01/eurostat/ns#geo,
    http://lod.gesis.org/dic/geo#BE, http://lod.gesis.org/dic/geo#BG, ...]

Additional Dimensions:
[http://purl.org/dc/terms/publisher,
    http://ontologycentral.com/2009/01/eurostat/ns#indic_de,
    http://www.w3.org/1999/02/22-rdf-syntax-ns#type, ...]
```

Stage B: Data Comparison For comparing two data sets regarding their matching suitability, a more detailed examination of the detected values and information is necessary. These actions are executed on the arrays that have been created as a result in Stage A. In a preprocessing step, the characteristics of the detected dimensions are analysed, i.e. the instance values and URIs of these elements are de-duplicated and sorted in order to identify their range and scope. For the temporal dimension, the pattern of the instance values is detected and out of it the frequency of the observations is derived, i.e. the intervals at which the observation values have been collected (annually, monthly, quarterly, etc.). The results of this preprocessing step are assignments of the different arrays to particular characteristics of the dimensions. Listing 5.2 presents an example output of this step. The result enables the comparison between temporal and spatial coverage for both the data sets.

Listing 5.2: Example output of the preprocessing step.

```
Following time interval and observation points have been detected:
[annually, 2010, 2005, 2011, 2009, 2006, 2007, 2004, 2008]
```

In Packages B1 and B2, similarities and conflicts in terms of overlapping instance values between two data sets are inspected. A similarity has to be identified between, at least, one pair of dimensions that can then be used as a key variable for merging and integrating the data sets afterwards. This is necessary in order to match them for a combined analysis. The tests in these packages are conducted on the schema elements, i.e. the dimensions, and their instance values. The instance values of the corresponding dimensions of both data sets are compared. For the dimensions detected in Package A3 that are neither temporal nor geographical, these dimensions are compared pair-wise. With these comparisons, it is determined whether there are overlaps between the instance values of the dimensions of two data sets or not. Conflicts can arise through differing time ranges or intervals (observation frequencies, e.g. annually or monthly) or different geographical areas that cannot be compared with each other without further manual work. As already mentioned, if there are no similarities between two data sets, it does not mean that no scientific analysis is possible. This stage can been seen as a simple approach for schema matching between two data sets. Stage B provides, as a result, an overview of detected similarities between two data sets and the potential conflicts that can occur through a combined analysis. In Listing 5.3, an example output is shown where similarities between the dimensions of two data sets have been detected.

Listing 5.3: Example output of Stage C.

```
Similar Time Points between the datasets could be detected:
[2006, 2004, 2010, 2007, 2009, 2005, 2008]

Assumed Observation Intervall: annually

Similar Geo Points between the datasets could be detected:
[http://lod.gesis.org/dic/geo#LT, http://lod.gesis.org/dic/geo#IE, ... ]
```

In order to keep the tests as a simple as possible and to provide an initial insight into the data, the following issues are not covered because they would need a more specific and complex treatment. Distinguishing between incomparable data sets and incompatible data sets remains an open issue. Data sets are incomparable if they hold different dimensions that cannot be associated, e.g. one data set holds a temporal dimension and the other one holds a geographical dimension. Thus, there is a lack of additional dimensional information, which can result from an incomplete representation as Linked Data, but can also result from a marginally annotated data source. Incompatible data sets hold comparable dimensions, but the instance values do not fit together without preprocessing work, such as annual versus monthly observation frequencies. In that case, the data sets could be made comparable, but this requires additional input from the user. Technically, only suggestions can be made, such as leaving out certain data points or using averages. The data comparison in Stage B is kept at a minimum because of the scope of this approach. In Sections 5.3 and 5.4, we focus on the challenges of data matching with Statistical Linked Data in detail.

5.2.4 Technical Implementation

The assessment tests have been implemented prototypically in JAVA and are accessible in a web application at `http://lod.gesis.org/gesis-lod-pilot/stat/structure.jsp`. As we provide a basic framework for performing the tests, the prototype can be executed using Linked Data sets that are consistent with our definition of Statistical Linked Data in Section 2.2. The data is retrieved by an internal SPARQL query service, which loads data from the web that is addressed by `FROM` and `FROM NAMED` clauses in the query. An additional data set is retrieved by adding another `FROM` clause to the query. This is required for a combined assessment of multiple data sets as defined in Stage B. Only then it is executed. In order that the data can be retrieved correctly, specific namespaces that are used inside the data set, have to be added to the query as a `PREFIX` clause. As we want to provide the maximum possible information about the data sets, all triples from a source are queried. Since Statistical Linked Data sets can hold a lot of information and, therefore, contain a lot of triples, we decided not to follow URIs in order to enable a quick analysis. Thus, the algorithm works best if all the required items are encoded as literals inside the queried data. If the data is split into multiple files or if there is an extra schema or metadata file, these files can be included into the SPARQL query through the use of the `FROM` clause. Listing 5.7 presents the query for an assessment of a single data set.

Listing 5.4: Exemplary SPARQL query for the analysis of a single data set.

```
PREFIX sdmx-measure: http://purl.org/linked-data/sdmx/2009/measure#
PREFIX dcterms: http://purl.org/dc/terms/
PREFIX eus: http://ontologycentral.com/2009/01/eurostat/ns#
PREFIX rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
PREFIX qb: http://purl.org/linked-data/cube#
PREFIX rdfs: http://www.w3.org/2000/01/rdf-schema#

SELECT *
FROM http://estatwrap.ontologycentral.com/data/tps00001
WHERE {
  ?s ?p ?o .
}
```

Listing 5.5: Excerpt of a retrieved JSON result.

```
{
  "head": {
   "vars": [ "s" , "p" , "o" ]
  } ,
  "results": {
    "bindings": [
      ...
      {
        "s": { "type": "bnode" , "value": "b298" } ,
        "p": { "type": "uri" , "value": "http://www.w3.org/1999/02/22-rdf-syntax-ns#type" } ,
        "o": { "type": "uri" , "value": "http://purl.org/linked-data/cube#Observation" }
      } ,
      {
        "s": { "type": "bnode" , "value": "b298" } ,
        "p": { "type": "uri" , "value":
    "http://purl.org/linked-data/cube#dataset" } ,
        "o": { "type": "uri" , "value":
    "http://lod.gesis.org/id/tps00001#ds" }
      } ,
      {
        "s": { "type": "bnode" , "value": "b298" } ,
        "p": { "type": "uri" , "value":
    "http://ontologycentral.com/2009/01/eurostat/ns#indic_de" } ,
        "o": { "type": "uri" , "value":
    "http://lod.gesis.org/dic/indic_de#JAN" }
      } ,
      {
        "s": { "type": "bnode" , "value": "b298" } ,
        "p": { "type": "uri" , "value":
    "http://ontologycentral.com/2009/01/eurostat/ns#geo" } ,
        "o": { "type": "uri" , "value": "http://lod.gesis.org/dic/geo#NO" }
      } ,
      {
        "s": { "type": "bnode" , "value": "b298" } ,
        "p": { "type": "uri" , "value": "http://purl.org/dc/terms/date" } ,
        "o": { "type": "literal" , "value": "2011" }
      } ,
      {
        "s": { "type": "bnode" , "value": "b298" } ,
        "p": { "type": "uri" , "value":
    "http://purl.org/linked-data/sdmx/2009/measure#obsValue" } ,
        "o": { "type": "literal" , "value": "4920305" }
      } ,
        ...
    ]
  }
}
```

Listing 5.5 presents the retrieved results of the SPARQL query as an excerpt[7]. The data is retrieved in the JSON format and written into one table. Each column of the table depicts the value of the retrieved element p from the JSON result, except for the first column, which contains vertically all values from the element `s`. The rows of the table are then filled with the corresponding values of the element `o`. The structure of the table enables easy scanning of the results, but secures that, for each cell, the corresponding JSON binding is reproducible. The defined tests of Stage A are performed on this generated table. In a first step, information is extracted from the table as much as possible. In a second step, the extracted information is analysed and compared according to the corresponding test. Listing 5.6 presents an excerpt[8] of the result output of the algorithm. First, the indicator, i.e. the measure of the observation, has been detected.

[7] The '...' denote the places where content has been left out.

[8] The '...' denotes the place where content has been left out.

The analysis of the dimensions then delivers the observation frequency and the detected values of the temporal dimensions, i.e. the time points where the observation has been measured. With respect to a multidimensional data model, each observation value is enlisted with the detected values of its corresponding dimensions.

Listing 5.6: Excerpt of the result output for one data set.

```
Indicator: Total population

Interval of observations is: annually
Observations for the following time points are available:
2010 2006 2005 2011 2008 2009 2004 2007

Value: 414372
Dimension Time: 2010
Dimension Geo: http://lod.gesis.org/dic/geo#MT
Additional Dimension: http://lod.gesis.org/dic/indic_de#JAN

Value: 299891
Dimension Time: 2006
Dimension Geo: http://lod.gesis.org/dic/geo#IS
Additional Dimension: http://lod.gesis.org/dic/indic_de#JAN

Value: 21462186
Dimension Time: 2010
Dimension Geo: http://lod.gesis.org/dic/geo#RO
Additional Dimension: http://lod.gesis.org/dic/indic_de#JAN

Value: 9011392
Dimension Time: 2005
Dimension Geo: http://lod.gesis.org/dic/geo#SE
Additional Dimension: http://lod.gesis.org/dic/indic_de#JAN
...
```

An assessment of two data sets follows the same process except that Stage B is also executed subsequently. The different steps are presented in the following listings. Listing 5.7 depicts the SPARQL query and Listing 5.8 presents an excerpt[9] of the result output of the assessment of two data sets.

Listing 5.7: Exemplary SPARQL query for the combined analysis of two data sets.

```
PREFIX sdmx-measure: http://purl.org/linked-data/sdmx/2009/measure#
PREFIX dcterms: http://purl.org/dc/terms/
PREFIX eus: http://ontologycentral.com/2009/01/eurostat/ns#
PREFIX rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
PREFIX qb: http://purl.org/linked-data/cube#
PREFIX rdfs: http://www.w3.org/2000/01/rdf-schema#

SELECT *
FROM http://estatwrap.ontologycentral.com/data/tps00001
FROM http://estatwrap.ontologycentral.com/data/teicp000
WHERE {
```

[9] Again, the '...' denote the places where content has been left out.

```
  ?s ?p ?o .
}
```

Listing 5.8: Excerpt of the result output for two data sets.

```
Detected Indicators: [HICP - all items, Total population ]

Detected Observation Values:
[43758250, 4.6, 1.1, ..., 114.33, 2.8,  2.8, 129.57, 0.2, 4435056, 112.63,
    3.3, 106.6, 0.2, 63229635, 131.84, 0.3, 5326314, 2.8, 0.8, -0.8,
    111.33, 1.4, 2041941]

Detected Time Points:
2006 2010-10 2010-12 2004 2010-11 2011 2011-04 2011-05 2009 2011-01 2008
    2005 2010 2011-03 2007 2011-02

No similar time points between the data sets could be detected.

There might be differences in the observation intervals between the data
    sets.

Assumed Observation Interval: annually

Detected Geographical Dimensions:
[http://lod.gesis.org/dic/geo#ES, http://lod.gesis.org/dic/geo#IS,
    http://lod.gesis.org/dic/geo#HR, http://lod.gesis.org/dic/geo#SE,
    http://lod.gesis.org/dic/geo#PT, http://lod.gesis.org/dic/geo#LU,
http://lod.gesis.org/dic/geo#DE, http://lod.gesis.org/dic/geo#IE,
    http://lod.gesis.org/dic/geo#GR, http://lod.gesis.org/dic/geo#ME, ...]

Similar Geo Points between the data sets could be detected:
[http://lod.gesis.org/dic/geo#ES, http://lod.gesis.org/dic/geo#SE,
    http://lod.gesis.org/dic/geo#PT, http://lod.gesis.org/dic/geo#DE,
    http://lod.gesis.org/dic/geo#IE, http://lod.gesis.org/dic/geo#GR, ...]
```

In Listing 5.8, the detected measures in both data sets are listed followed by the identified observation values and values of the temporal dimensions. The detected time points indicate that there seem to be differences in the observation frequencies as one data set contains annual values and the other data set monthly values. As the 'Assumed Observation Interval', an annual frequency is detected. This results from a comparison between both frequencies, where the annual one is more generic and can therefore serve as a common frequency for both. However, this implies an aggregation of the values of the second data set in order for it to be comparable to the annual frequency. Finally, the geographical dimensions of both data sets are listed in the results, followed by the results of a comparison between both, which lists only those values of the geographical dimensions appearing in both data sets.

Data Provider	Number of Used Data Sets
Eurostat	10
World Bank Climates	10
data.gov	10
OECD	10
Global Hunger Index	1
EnAKTing Energy	1
ISTAT	1
data.gov.uk	1

Table 5.2: Data sets used in the evaluation.

5.2.5 Evaluation Setup and Results

We have evaluated the assessment tests using several Statistical Linked Data sets. We chose various different data sets in order to achieve significant results and to test the general applicability of our approach.

Evaluation Setup

We evaluated whether the assessment tests work with Statistical Linked Data and how they perform. Therefore, we have run each test on several Statistical Linked Data sets and computed precision, recall and F-measure scores as the results. For Stage A, it was evaluated whether the data elements of the packages could be identified correctly and completely. For Stage B, it was evaluated whether the dimension characteristics and the overlaps between instance values have been found. We have not evaluated the data sets themselves.

For the evaluation, we use real-world Statistical Linked Data sets from the web (for more information see also Section 2.2). We have chosen representative data sets of different data providers in order to achieve universal and significant results for statistical data in general. For this reason, we randomly chose ten data sets per data provider that has a large number of published data sets as well as single data sets from data providers with a smaller offering on Linked Data. Additionally, we ensured that all data sets are not modelled and represented in RDF the same way. This was necessary because the assessment tests rely heavily on the structure of the data.

Table 5.2 presents an overview on the chosen data sets. All data sets are available in either the OWL or RDF formats and are published as Linked Data. The correct data elements have been labelled by domain experts for each test. Additionally, for Stage B, the corresponding pairs of schema elements have been created as a reference alignment by domain experts as well. This information serves as the gold standard in our evaluation and is the basis for the computation of precision, recall and F-measure.

In accordance to [vR79], we define precision as the ratio of all detected data elements that are correct and all detected data elements. Recall is defined as the ratio of all detected data elements that are correct and all correct data elements (see Section 2.3.4 for more details).

Results

In the following tables, the detailed results for all packages of the Stages A and B are shown that have been conducted on all data sets. Tables 5.3 and 5.4 show the results for the packages of Stage A[10]. Since we evaluated ten data sets for each of the data providers, Eurostat, World Bank Climates, data.gov, and OECD, Table 5.3 depicts only the computed means for precision and recall. However, we have observed that, in most cases, precision and recall is either high, from about 0.75 to 1, or very low, from 0 to 0.25. In general for this evaluation, a low precision or recall means that the particular test retrieved no satisfying results, i.e. not all or incorrect data elements have been detected. A reason for this lies, in first place, in the prototypical implementation of the packages. But another reason may also be the modelling and structure of the data set, i.e. missing information or the insufficient annotation of data elements. In particular, the results of Packages A3.1 and A3.2, where the geographical and temporal dimensions had to be identified, show that the data can be modelled very differently. Although we have defined rules that are based on a standardized way for representing these dimensions, it is not the only way allowed for modelling this particular part of the data. This impression is supported by the results for all data sets in Table 5.6, where the mean values for each stage have been computed. When precision and recall values are 1, either all data elements have been identified correctly or no particular element occurred in the data set, which has been the case for some data sets during the evaluation of Package A2. However, in these cases the assessment test was successful because it was correct that the particular data element was not included into the data set and, therefore, could not be detected.

Package	Eurostat		World Bank Climates		data.gov		OECD	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
A1	0.75	1	1	1	0.695	0.833	1	0.931
A2	1	1	1	1	1	1	1	1
A3	0.794	1	0.724	1	0.898	1	0.712	1
A3.1	0.633	1	0.217	0.5	0.5	0.5	0.369	1
A3.2	0.5	0.5	1	1	1	0.972	0.95	1
A4	1	1	1	1	0.695	0.833	1	0.931
Stage A	0.779	0.917	0.824	0.917	0.798	0.856	0.839	0.977

Table 5.3: Results of Stage A.

[10] Prec. = Precision, Rec. = Recall

Package	Global Hunger Index		EnAKTing Energy		ISTAT		data.gov.uk	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
A1	0.333	1	1	1	1	1	0.5	1
A2	0.2	1	1	1	1	1	1	1
A3	0.778	1	0.625	1	0.833	1	0.875	1
A3.1	0	0	0	0	0	0	0.333	1
A3.2	0.25	1	0	0	0	0	0.5	1
A4	1	1	1	1	1	1	1	0.5
Stage A	0.427	0.833	0.604	0.667	0.639	0.667	0.701	0.917

Table 5.4: Results of Stage A (continued).

In Table 5.5, the results for the evaluation of Stage B are shown[11]. Since two data sets have been analysed together for this stage and the pair-wise combination of all data sets used in this evaluation would lead to a large number of tests, we decided to reduce the number of pairs examined to the data sets of the four larger data providers, Eurostat, Worldbank Climates, data.gov, and OECD. The results of stage B are independent of those of Stage A, i.e. for the case that no similar dimensions have been identified in Stage A, potential pairs of dimensions between both data sets may be identified according to their instance values in Stage B.

Between the data sets of one single data provider, the precision and recall is always 1, because the data is modelled in the same way, i.e. the same instance values are always represented equally. In general, the precision is always 1, since all detected similar instance values have been correct. However, the low recall for some tests prove that in a significant number of cases, not all correct overlapping instance values (B1) and no overlaps between instance values (B2) have been detected. We have observed that it is easy for the algorithm to detect similar or different instance values when the instances are represented using the same code or pattern. Most prominent cases in the evaluation have used different country lists and, in the case of World Bank Climates, a completely different coded representation of dates. If this is not the case, correct instance overlaps may be missed. This observation is valid for the detection of no overlaps between instance values. These have been missed when the instance values have been coded differently. The results in Table 5.5 prove that Stages B1 and B2 receive similar results and that it is reasonable to combine them into a single task.

In Table 5.6, we summarize the results of the evaluation. We have calculated the mean values for each particular stage and package. This highlights the strengths and weaknesses of the assessment tests. As already discussed, the identification and extraction of information on data elements works successfully in general. However, depending on the structure and semantic richness of single data sets, the tests may not achieve the desired results. While dimensions can be identified in general, the detection of temporal and geographical dimensions, in particular, can be improved (Packages A3.1 and A3.2),

[11] Prec. = Precision, Rec. = Recall

Package	Eurostat		World Bank Climates		data.gov		OECD	
	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
B1								
Eurostat	1	1	1	0.5	1	0.4	1	0.5
World Bank Climates	-	-	1	1	1	0.3	1	0.1
data.gov	-	-	-	-	1	1	1	0.25
OECD	-	-	-	-	-	-	1	1
B2								
Eurostat	1	1	1	0.5	1	0.4	1	0.5
World Bank Climates	-	-	1	1	1	0.3	1	0.1
data.gov	-	-	-	-	1	1	1	0.25
OECD	-	-	-	-	-	-	1	1
Stage B								
Eurostat	1	1	1	0.5	1	0.4	1	0.5
World Bank Climates	-	-	1	1	1	0.3	1	0.1
data.gov	-	-	-	-	1	1	1	0.25
OECD	-	-	-	-	-	-	1	1

Table 5.5: Results of Stage B.

since the underlying rules cannot be defined clearly. The use of several heterogeneous data sets in the evaluation prove that the approach can be generalized and that it is executable with different Linked Data sets. We will discuss the observed limitations in the following Section 5.2.6 in detail.

5.2.6 Discussion and Limitations

The evaluation has shown that detecting the data elements in Statistical Linked Data is challenging because there is no consistent labelling of data elements and no consistent patterns for instance values. This complicates the definition of assessment rules. However, the results are promising.

The complexity and extent of modelling data is often very different. Some providers deliver additional information about units, populations, provenance, etc., but this is not always the case. In most cases, this is not a problem of the RDF representation of the statistical data. It is often due to the original published data format, which often does not include such information directly. All examined data sets are – more or less accurately – modelled according to the Linked Data principles. Therefore, a lot of additional information about dimensions, etc. is encoded in the URIs. Currently, the implementation does not query URIs in a data set in order to retrieve more information. This hinders the full identification of data characteristics as intended in Stage B, thus complicating the data comparison in general.

Stage / Package	All Data Sets		
	Precision	Recall	F-Measure
A1	0.785	0.971	0.864
A2	0.9	1	0.947
A3	0.779	1	0.876
A3.1	0.257	0.5	0.339
A3.2	0.525	0.684	0.594
A4	0.961	0.908	0.934
Stage A	0.701	0.844	0.766
B1	1	0.605	0.754
B2	1	0.605	0.803
Stage B	1	0.605	0.803

Table 5.6: Summarized results for Stages A and B.

The results have revealed the challenge that there is sometimes more than one date in a single observation. For example, data about schools from data.gov.uk includes diverse dates like the 'opening date' or the 'date of the last welfare visit', among other things. This complicates the automatic detection of temporal dimensions because there might not be just one correct solution and because research interests are diverse. While in a well-formed data set all of these dates are accompanied with specific XML datatype properties, the semantics behind the dates, e.g. in the underlying schema, have to be taken into account.

In order to guess possible factors for making data sets comparable, the information on dimensions must be very detailed, e.g. the existence of hierarchical structures in a dimension. For example, the structure of NUTS levels may be useful in order to aggregate data between different levels. This can be a solution if one data set is available on NUTS level 2 and the other one on NUTS level 1. From a scientific point of view, this might represent a loss in data quality, but it can at least support researchers in getting an initial insight into the data.

Also important for the detection of values and dimensions is the naming of the property and class types in a data set. The more that standardized vocabularies (e.g. Data Cube vocabulary [CRT14], SCOVO [HHR$^+$09], Dublin Core [Ini]) are used or that the naming conventions of the URIs are generic and machine-interpretable, the easier is an automatic detection. A promising approach, especially as a preprocessing step for Packages B1 and B2, can be the use of link discovery tools (e.g. Silk [VBGK09], SERIMI [AHSdV11]) and ontology/schema matching systems, including the matching of instances like PARIS [SAS11] or COMA++ [EM07]. Such tools can detect links between dimensions or precise values of their instances. More powerful techniques for this purpose are discussed in Section 5.3 and 5.4.

With the execution of the Packages A1 to B2, all of our assessment tests are complete. After a successful execution, we are able to determine special characteristics and dimensional coverages of a data sets with little effort. We are also able to detect, whether two data sets are suitable for a comparative analysis or in which dimensions problems can occur. However, the detection and extraction of necessary information from data sets is not trivial and can be improved. Additionally, the results of the evaluation of Stage B have shown that it is relevant to consider how the instance values are coded, i.e. whether they are coded differently. Since information on dimensions lies in the properties of Linked Data sets (i.e. the schema elements of the data sets that are desired to be matched), we will focus on matching these properties in the following sections.

5.3 Instance-based Schema Matching

While Stage B in the assessment tests of the previous section followed a prototypical approach in testing whether two data sets can be matched, we investigate in this section what a schema matching approach focused on Statistical Linked Data would look like. Currently, there exist various approaches on schema matching that consider different characteristics of the schemata, including e.g. structure, schema elements and instances [SE05]. While existing schema matching approaches are evaluated among different kinds of data sets, it can be observed that they have not yet been evaluated in the context of Statistical Linked Data. In recent years, instance-based schema and ontology matching gain importance [BMR11] as similar schema elements can be derived out of the similarity of their instance values. An overview of existing approaches can be found in [IMSW07, ES07]. For the related field of ontology matching, [SE11] states that different domains and the inclusion of users into the matching process reveal new challenges like treating new types of information resources, e.g. spatial or temporal information and domain-specific constraints. According to [Hal05], especially domain-specific values, significant occurrences of values and patterns of values are stated as relevant characteristics to be considered at the instance level, as well as integrity constraints for schema elements and their instance values.

Continuing our insights from the previous section about considering datatype properties and object properties of Statistical Linked Data separately, we present in this section a novel approach for instance-based schema matching considering the patterns of the instances, i.e. the instances associated with datatype properties. Given two schema elements, we inspect their instances and analyse whether they can be expressed by regular expressions that are predefined and stored in different sets. Each set represents a specific statistical data element, e.g. date. Hereby, to every regular expression, a weight is assigned that expresses its adequacy to the data element. For the two given schema elements, matchings are calculated using particular weightings, if instances from both schema elements can be expressed by regular expressions from the same set.

This section is an extended version of [ZZS12], in which the approach has been initially introduced. In Section 5.3.1, we describe the problem of schema matching with Statistical Linked Data and statistical data in general. We present our approach on instance-based schema matching utilizing regular expressions in Section 5.3.2. We introduce the concept of pattern classes and propose how to utilize them to detect similar schema elements. The implemented algorithm is described in Section 5.3.3 in detail. In Section 5.3.4, we present the evaluation of our approach. Engaged by the results of the evaluation, we discuss further considerations and limitations of our approach in Section 5.3.5.

5.3.1 Challenges of Schema Matching with Statistical Linked Data

Statistical Linked Data differs in its structure and semantics from other Linked Data sets (see Section 2.2). It has not or rarely been considered by existing schema and ontology matching approaches so far, which focus commonly on hierarchical structures and semantics at the schema or element level (see Section 2.3.2).

The matching of the schemata of Statistical Linked Data is a challenging problem because schema elements are often named by simple and short labels (e.g., geo, date), sometimes even by abbreviated terms (e.g. refArea, ISIC). However, the structure and semantics of the instances differ in various aspects from text-heavy data. They are mostly numerical values, entries of code lists, or simple and short strings (see Section 2.2). Moreover, within a schema element, instances are described by a specific syntactical pattern, e.g. dates consist of numerical values divided by periods or slashes. These patterns can be leveraged for the schema matching of Statistical Linked Data because there may be no identical instances inside both of the desired schemata, e.g. when thinking of merging two data sets of unemployment rates from different countries or covering different time frames. Considering the patterns of instance values, instance-based schema matching approaches seem to be a more adequate approach for matching the schema elements of statistical data. Table 5.7 presents an overview of different schema elements and instance values encoded in different patterns, with which researchers are confronted when intending to match schema elements.

Different approaches in considering instance values for schema matching can be identified in current systems. An overview and evaluation of current matching functions can be found in [IMSW07]. In most cases, instance matching algorithms for the resolution of entities are used. This is grounded in the assumption that identical or similar instance values belong to the same or a similar schema element. Most similar to our approach are the following systems. In COMA++ [EM07], the matching process is enhanced by adding constraint-based matching. A similar approach is described in [ZSC09], where the use of regular expressions and catchwords is considered for instance-based schema matching. But in contrast to both systems, we focus on statistical data, where the potential of patterns and regular expressions can be fully exposed.

Other systems like the GLUE system [DMDH03] use similarity measures, the exploitation of domain constraints and heuristic knowledge for schema matching. Some schema matching systems use machine learning and rule-based approaches for detecting features within schema elements and instance values [BM02, JFNP12]. Features of instance values are equally well-computed in [ZC09]. Lexical similarities of instance values are computed in [HSNM11] in order to create object properties that can be seen as a similar approach

Statistical Element	Examples of Different Schema Elements	Examples of Different Instance Values

Temporal	time
	date
	year
	reporting year
	refPeriod
		`11.03.2004`
		`11/03/2004`
		`2004-11-03`
		`2004-03`
		`03-2004`
		`2004`

Geographical	geo
	refArea
	country code
		`US`
		`USA`
		`United States`

Gender	gender sex	`Female` `F`

Age Group	age
	ageRange
		`30-49`
		`From 30 to 49 years`
		`Age range 30 - 49`
		`Y30-49`

Table 5.7: Overview of typical instance values and patterns.

to the use of features. PARIS [SAS11], a relatively new approach, measures degrees of matchings between instances and schema elements based on probability estimates.

5.3.2 Schema Matching Using Regular Expressions

In this section, we introduce our approach towards instance-based schema matching utilizing regular expressions in pattern classes. In this approach, we define pattern classes for each statistical concept as introduced in Section 5.2.1 and defined in the SDMX Content-Oriented Guidelines [SDM09], i.e. one pattern class for the geographical dimension, one for the temporal dimension and so on. These classes are used as background knowledge during the matching process and contain multiple regular expressions for representing instance values of this particular statistical concept. A correspondence between two schema elements is considered if their instances can be expressed via a regular expression from one particular pattern class.

In the following subsection, we describe how such pattern classes containing multiple regular expressions are defined and present our approach for finding similar schema elements including the implemented algorithm.

Pattern Classes as Background Knowledge

For our approach, we assume that there exist several pattern classes that contain multiple patterns describing a particular statistical concept (see Section 5.2.1) like e.g. dates, age groups or geographical codes. Each pattern is described as a regular expression. Table 5.8 presents an excerpt of different instance values that can appear in a schema element with temporal coverage and their corresponding patterns. More patterns for this particular statistical concept can be found in [SDM09].

Instance Value	Regular Expression
`2010`	`[0-9]{4}`
`2010`	`[0-2][0-9]{3}`
`10-2010`	`[0-9]{2}-[0-9]{4}`
`28.10.2010`	`[0-9]{2}.[0-9]{2}.[0-9]{4}`
`2010-28-10`	`[0-9]{4}-[0-3][0-9].[0-1][0-9]`

Table 5.8: Overview of instance values and their corresponding regular expressions for temporal information.

In the case of geographical codes, the definition of patterns is more complicated than for dates. Although there are international standards like the ISO norms 3166-1 and 3166-2 or the NUTS classification, the derived patterns are of a very generic kind, like two characters for countries, e.g. `DE`, `FR`, `ES` or `US`. Patterns describing these instances may be derived from other entries of other schema elements as well. This implies that lower weightings have to be assigned to them. In Table 5.9, patterns for geographical codes are presented. The first three patterns refer to the ISO norms, while the other three patterns correspond to entries of the NUTS classification. `DEA` encodes the federal state of *North Rhine-Westphalia* at NUTS level 1, `DEA2` describes the administrative district of *Cologne* at NUTS level 2 and `DEA22` refers to the independent city of *Bonn*.

Table 5.10 presents possible instance values and derived patterns across age groups. This example illustrates the problem that occurs if there is no standardized way of encoding such an information in statistical data. The entries are very heterogeneous and the only element of a pattern that is certain in most cases is the encoding of a numeric range like `## - ##`, where the `##` determine a specific age. Nevertheless, as the example depicts, there can be other definitions like `Y_LT15`, which encodes the age group for all people younger than 15.

Instance Value	Regular Expression
`DE`	`[A-Z]{2}`
`DEU`	`[A-Z]{3}`
`DE-NW`	`[A-Z]{2}- [A-Z0-9]{0-3}`
`DEA`	`[A-Z]{2}[A-Z0-9]`
`DEA2`	`[A-Z]{2}[A-Z0-9]{2}`
`DEA22`	`[A-Z]{2}[A-Z0-9]{3}`

Table 5.9: Overview of instance values and their corresponding regular expressions for geographical code lists.

Instance Value	Regular Expression
Y_LT15	[A-Z]_[A-Z]{2}[0-9]{2}
Y30-49	[A-Z][0-9]{1,2}-[0-9]{1,2}
30-49	[0-9]{1,2}-[0-9]{1,2}

Table 5.10: Overview on instance values and their regular expressions for age groups.

Finding Similar Schema Elements

For our approach, we define two given data sets as M and N. A given set of pattern classes, we denote as C. For each pattern class $C_x \in C$, consisting of regular expressions, a match between two schema elements $S_M \in M$ and $S_N \in N$ is detected, if at least one instance from S_M and S_N can be expressed by a regular expression from the same pattern class C_x. Hereby, a weighting Ω expresses the probability of the match with a value between 0 and 1.

Let C be the set of pattern classes with

$$C = \{C_1, C_2, C_3, ..., C_n\}$$

then each class $C_x \in C$ is itself a set comprising tuples of regular expressions and an additional weighting ω. The regular expressions describe the patterns for representing a particular statistical concept x (e.g. age groups) and the weighting ω is a value between 0 and 1 determining how appropriately the regular expression represents x:

$$C_x = \{(regex, \omega) | regex \text{ matches } x,\ 0 < \omega < 1\}$$

This additional weighting ω was included for sorting multiple regular expressions regarding their appropriateness for representing the statistical concept of the class, e.g. for the concept of age groups. For example, $C = \{C_{\text{date}}, C_{\text{age}}, C_{\text{geo}}\}$ is a set of pattern classes that represents a date, an age reference and a geographical location. An example for such a pattern class is $C_{\text{date}} = \{([0-9]\{2\}.[0-9]\{2\}.[0-9]\{4\}, 0.9), ([0-9]\{2\} - [0-9]\{4\}, 0.8)\}$ for dates, as shown in Table 5.10.

As we intend to calculate a confidence value considering all instances within a schema element, for each $C_x \in C$ we calculate the average weighting for all schema elements $S_M \in M$ and $S_N \in N$. As soon as an instance of a schema element can be expressed by a $(regex, \omega) \in C_x$, the value of ω is added to the sum of all weightings whose regular expressions previously matched another instance, resulting in the final $\sum_0 \omega$ for all instances. The average is calculated by normalizing this sum regarding the total number of instances in this particular schema element. For each S_M and S_N, this is

$$avg(S_M) = \frac{\sum_0 \omega}{|\text{Instances in } S_M|}$$

$$avg(S_N) = \frac{\sum_0 \omega}{|\text{Instances in } S_N|}$$

For C_{date} from the example above, let a schema element $Date_M \in M$ have the instances `28.20.2010` and `10-2010`. Then the first instance can be expressed by the second regular expression in C_{date}, and the second instance by the first one. Accordingly, the average weighting is $avg(Date_M) = \frac{0.9+0.8}{2} = 0.85$.

All schema elements are collected in a set together with their average weighting if the average weighting is not 0. We define these sets as M_x and N_x for each C_x. All schema elements form a tuple with their aggregated weight and are denoted as

$$M_x = \{(S_M, avg(S_M)|\exists(regex, \omega) \in C_x : regex \text{ matches min. 1 instance of } S_M\}$$

$$N_x = \{(S_N, avg(S_N)|\exists(regex, \omega) \in C_x : regex \text{ matches min. 1 instance of } S_N\}$$

In the example, we can see, that the schema element $Date_M$ contains instances matched by a regular expression in C_{date}. Thus, it is an element of M_{date}, whereas another schema element like Geo_M, containing strings for country codes, would probably not be matched by any regular expression in C_{date}. Therefore, it would not be an element of M_{date}.

Finally, we calculate the Cartesian product $Matches_x = M_x \times N_x$, where a triple (S_M, S_N, Ω) defines a match between a S_M and a S_N with the probability of Ω computed from the average weightings.

$$Matches_x = \{(S_M, S_N, \Omega) \in M_x \times N_x | \Omega = avg(S_M) * avg(S_N)\}$$

Additionally to $Date_M \in M$ from our example, we assume that in data set N there exists a different schema element $Dval_N = \{19.11.2009\}$. Regarding the pattern class C_{date}, N_{date} contains this schema element with $(Dval_N, 0.9)$ analogous to $(Date_M, 0.85) \in M_{date}$. Consequently, $Matches_{date}$ would calculate the triple $(Date_M, Dval_N, 0.76)$. Thus, a match with a specific confidence has been found.

5.3.3 Algorithm and Implementation

Our algorithm as depicted in Listing 5.9 starts with two data sets M and N and one or more pattern classes comprised in C as input. According to our approach described in Section 5.3.2, the elements (S_M, S_N, Ω) of $Matches_x$ are enlisted as output. They contain a schema element of each data set together with a probabilistic value, which describes the confidence of a matching between both schema elements.

Each instance of each schema element S_M is examined regarding whether it can be expressed by one regular expression *regex* of the first pattern class C_x. If so, the weight ω of the regular expression is added to sum_M. If the instance value does not match with the current pattern, the next pattern of C_x is examined. After each match, the next instance value is analysed regarding a potential match with one of the regular expressions. After each instance of the current schema element has been tested, the overall weight $avg(sum_M)$ for the element is calculated. The element and its weight are then added as $(S_M, avg(sum_M))$ to the class M_x, which contains all analysed schema elements with their weightings for the pattern class C_x. These steps are repeated for each schema element.

This process is also conducted analogously with each instance of each schema element S_N of the second input data set N. The detected schema elements are added with their calculated weightings as $(S_N, avg(sum_N))$ to the class N_x. After this cycle, the Cartesian product $M_x \times N_x$ is computed and it delivers the matchings between the schema elements of each ontology with a calculated confidence value (S_M, S_N, Ω). After that, both ontologies are compared with the patterns of the next pattern class C_x according to the above described algorithm.

Listing 5.9: Pattern matching using regular expressions.

```
Input: $M,N,C$
Output: List of $(S_M,S_N,\Omega)$
for ($C_x$) {
  for ($S_M \in M$) {
    $sum_M = 0$
    while (instance $\in S_M$) {
      a: for (($regex,\omega) \in C_x$) {
        if ($regex$ matches instance) { // otherwise take
                                        // the next regex
          $sum_M = sum_M + \omega$
          break a: // continue with the next instance
        }
      }
    }
    if ($sum_M \neq 0$) { // at least one instance was matched
                        // by a regex
      calculate $avg(sum_M)$
      add $(S_M, avg(sum_M)) \rightarrow M_x$
    }
  }
  for ($S_N \in N$) {
    $sum_N = 0$
    while(instance $\in S_N$) {
      b: for (($regex,\omega) \in C_x$) {
        if ($regex$ matches instance) { // otherwise take
                                        // the next regex
          $sum_N = sum_N + \omega$
          break b: // continue with the next instance
        }
      }
    }
    if ($sum_N \neq 0$) { // at least one instance was matched
                        // by a regex
      calculate $avg(sum_N)$
      add $(S_N, avg(sum_N)) \rightarrow N_x$
    }
  }
  for ($(S_M, avg(sum_M)) \in M_x$) {
    for ($(S_N, avg(sum_N)) \in N_x$) {
      println $(S_M, S_N, \Omega)$ // compute the cartesian product
                           // and generate output
    }
  }
}
```

In order to consider all instances of all schema elements from both data sets and all regular expressions from all pattern classes, we have implemented several loops nested in each other. This implies that the complexity of our algorithm increases very quickly according to the size of the data sets and pattern classes. To avoid this, we have defined breaks in the loops after a regular expression matches an instance. Hereby, the sorting of the regular expressions ensures that, for each instance, the most promising and precise pattern is considered first as a match. This reduces the costs of the algorithm.

We have implemented our approach for matching schema elements using regular expressions as an Apache Maven project in Java. The ontologies are processed with the JENA API[12]. The source code and an executable jar file are available at `https://github.com/mazlo/smurf`.

5.3.4 Evaluation Setup and Results

Since schema matching systems have been evaluated extensively in the past [RB01], we focus on the performance of those systems for matching Statistical Linked Data. We assume that the specific structure and characteristics of instances of schema elements in statistical data hinder existing schema matching systems from delivering adequate results for this type of data.

Setup

The evaluation is conducted on real-world Statistical Linked Data sets for two reasons:

1. In current evaluation campaigns like the OAEI[13] no suitable benchmark or data set exists that addresses the investigated characteristics in the data.
2. Using real-world data is closer to our use case of 'Analysing Research Data'.

[12] http://jena.apache.org/

[13] http://oaei.ontologymatching.org/

For all evaluated pairs of data sets, the resulting correspondences are validated with their particular reference alignments that have been created by domain experts. We compute precision, recall and F-measure for each alignment task, since these are standard evaluation measures for ontology matching evaluation [ES07].

Data Sets and Matching Tasks

For our evaluation, we use real-world Statistical Linked Data sets from the web (for more information, see Section 2.2). We have chosen representative data sets of different data providers in order to achieve universal and significant results for statistical data in general. For this reason, we randomly chose ten data sets per data provider with a large number of published data sets as well as single data sets from smaller data providers.

Table 5.11 presents an overview of the chosen data sets. In the evaluation, all schema elements are considered for computing the results. All data sets are available in either the OWL or RDF formats and are published as Linked Data. The reference alignments between the schemata of these data sets which are necessary in order to calculate precision, recall and F-measure have been created manually by domain experts.

Data Provider	Number of Used Data Sets
Eurostat	10
World Bank Climates	10
data.gov	10
OECD	10
Global Hunger Index	1
EnAKTing Energy	1
ISTAT	1
data.gov.uk	1

Table 5.11: Data sets used in the evaluation.

In this evaluation, we have created multiple matching tasks. Each task consists of a pair of two data sets that have to be matched by the matching systems. Table 5.12 shows an overview of the defined matching tasks. We did not define matchings tasks between the data sets of Global Hunger Index, EnAKTing Energy, ISTAT and data.gov. Since we considered only one data set per each of these collections, the results would not have been significant enough.

Matching Systems

Apart from our implemented approach, we chose three representative matching systems for the evaluation from which we assume the best results regarding the problem statement. There have been four criteria for choosing the matching systems: (1) treatment of instance

Task	Data Set 1	Data Set 2
1	Eurostat	World Bank Climates
2	Eurostat	data.gov
3	Eurostat	OECD
4	Eurostat	Global Hunger Index
5	Eurostat	EnAKTing Energy
6	Eurostat	ISTAT
7	Eurostat	data.gov.uk
8	World Bank Climates	data.gov
9	World Bank Climates	OECD
10	World Bank Climates	Global Hunger Index
11	World Bank Climates	EnAKTing Energy
12	World Bank Climates	ISTAT
13	World Bank Climates	data.gov.uk
14	data.gov	OECD
15	data.gov	Global Hunger Index
16	data.gov	EnAKTing Energy
17	data.gov	ISTAT
18	data.gov	data.gov.uk
19	OECD	Global Hunger Index
20	OECD	EnAKTing Energy
21	OECD	ISTAT
22	OECD	data.gov.uk

Table 5.12: Matching tasks.

values for schema matching, (2) use of advanced operations and algorithms for matching, (3) performance during previous evaluations [EFM$^+$10, EFvH$^+$11] and (4) the capability of using RDF or OWL data without extensive data preprocessing. Especially the last criteria has been of relevance for our choice because we want to focus on real-world statistical data as it is being published on the web. Based on the criteria and the availability of various schema matching systems, we have decided to use the following three systems in the experiment.

FALCON-AO Falcon-AO [HQ08] is a matching system for ontologies expressed in RDF(S) and OWL, which has become popular in recent years. It can be used to consecutively apply different components to match two schemata, e.g. a linguistic, structural and partition-based matchers. We examine Falcon-AO since it has delivered very promising results in previous evaluations.

COMA++ COMA++ [EM07] is a generic and mature matching system, which is capable of several algorithms that can be flexibly combined with each other. A broad range of schema characteristics like schema structure or instance values is included into a matching process. For instance-based matching, there is a constraint-based and a content-based matcher. The constraint-based matcher determines constraints that describe the characteristics or patterns of instance values, and compares these constraints to each other. If they match, COMA++ matches the schema elements of the instances. The content-based matcher executes a pair-wise comparison of instances using a similarity function and stores the results in a comparison matrix. This matrix is aggregated to one value, which defines whether the two schema elements should be matched or not. In our experiment we consider both matching approaches.

PARIS PARIS [SAS11] is an alignment tool in the group of mixed, i.e. schema-based and instance-based matching systems. PARIS is capable of matching relations, instances and schemata. This is achieved by a holistic approach that measures degrees of matchings based on probability estimates. In preliminary tests with simple test cases, PARIS has been able to identify equal schema elements as well as literals on the instance-level correctly. This has also been observed at instance level even with different schema elements.

Our Approach Our approach towards schema matching utilizing regular expressions has been implemented and evaluated as described in Section 5.3.3. Since it is based on predefined pattern classes as background knowledge, such classes have been created for the statistical concepts *geographical* and *temporal dimension*s. We have focused on these two concepts because they are the most common ones and occur in all of the data sets chosen for evaluation. The patterns of the classes have been generated by domain experts based on the SDMX guidelines [SDM09], i.e. for the temporal dimension, the date patterns of SDMX were used and for the geographical dimension, the code lists of ISO 3166[14], OECD countries[15] and NUTS[16] have been used.

Results

In the following subsection, we present the results of our evaluation for each matching system and matching task. While all of the chosen matching systems have been capable of processing RDF or OWL data in general, some data sets had to be preprocessed. Since the used data sets mainly include instance data, we added the complete schema information into the data set as far as it was available. This extension lies at hand for the use of such data in schema matching systems. However, we have observed that, in some cases, the matching systems still had problems with parsing and loading some of the data sets. These parsing errors have been caused by inconsistent data modelling and had to be fixed first.

[14] http://www.iso.org/iso/home/standards/country_codes.htm

[15] http://stats.oecd.org/glossary/detail.asp?ID=461

[16] http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction

System	Precision	Recall	F-measure
COMA++	0.903	0.628	0.741
FALCON-AO	1	0.416	0.588
PARIS	1	0.481	0.649
Our Approach	0.895	0.713	0.794

Table 5.13: Summarized results of the matching systems for all tasks.

Table 5.13 shows the summarized results for all tasks by each matching system, i.e. the mean values for all tasks have been computed. The results indicate that our approach can achieve a significant improvement on the recall in comparison to the other matching systems, i.e. additional correct correspondences between schema elements could have been detected by considering the patterns of their instance values. Since, in COMA++, a constraint matcher and a content matcher are applied that also consider patterns of instance values, the results of COMA++ are most comparable to those of our approach. The precision for FALCON-AO and PARIS is always 1 because the detected correspondences were always correct, although in most cases, only a few correspondences have been found that results in a low recall. The loss in precision for our approach and for COMA++ can be explained by detected false correspondences, which have been found due to two reasons. First, there may be equal or similar patterns that refer to different statistical concepts, e.g. a pattern describing instance values for dates that occurs in a schema element containing other numerical instances. Second, in some of the evaluated data sets there have been multiple schema elements containing dates; i.e. the pattern assignment was correct, but the correspondence was semantically incorrect, e.g. `refPeriod` and `lastModified`, where the last element describes the date when the data set has been modified.

When examining the single results for each task in Table 5.14, we can observe that the loss in precision occurs in only a few cases (the same can be observed for COMA+ in Table 5.15), while the recall is nevertheless higher than the recall of FALCON-AO and PARIS (see Tables 5.16 and 5.17). Since we have used ten data sets per task for Eurostat, Worldbank Climates, OECD and data.gov, the values for precision, recall and F-measure are again the computed mean values.

COMA++ was able to deliver results similar to our approach as is shown in Table 5.15. Both, the constraint-based matcher and the content-based matcher found correspondences between schema elements based on their instance values or the patterns of their instance values. In tasks where the instance patterns were slightly different, correspondences could be detected, e.g. in Task 14, the instance values of two particular schema elements are numbers, but a few numbers were considered to be a subset of another, just as `1992` is a subset of `1992-03-04`. However, in COMA++, these values are assumed to be the same content. In Task 8, a correspondence has been detected, where the instance values of two schema elements were `1992-03-04` and `06/16/12`. Regarding the general and numerical constraints of COMA++, the constraints ‘average length’ and ‘special

Task	Precision	Recall	F-measure	Task	Precision	Recall	F-measure
1	0.75	0.6	0.667	12	1	0.6	0.75
2	1	1	1	13	0.8	0.8	0.8
3	1	0.8	0.889	14	1	0.75	0.857
4	0.6	0.6	0.6	15	1	0.75	0.857
5	1	0.6	0.75	16	1	0.5	0.667
6	1	0.6	0.75	17	1	0.75	0.857
7	0.66	0.5	0.569	18	0.8	1	0.889
8	0.5	0.4	0.444	19	0.75	0.6	0.667
9	1	0.8	0.889	20	1	0.8	0.889
10	1	1	1	21	1	0.8	0.889
11	1	0.6	0.75	22	0.833	0.833	0.833

Table 5.14: Results for our approach.

Task	Precision	Recall	F-measure	Task	Precision	Recall	F-measure
1	0.75	0.6	0.667	12	1	0.6	0.75
2	1	1	1	13	1	0.6	0.75
3	1	0.8	0.889	14	1	0.75	0.857
4	0.66	0.8	0.723	15	1	0.75	0.857
5	1	0.6	0.75	16	1	0.5	0.667
6	1	0.6	0.75	17	1	0.5	0.667
7	0.5	0.5	0.5	18	1	0.75	0.857
8	0.5	0.4	0.444	19	0.66	0.4	0.498
9	1	0.6	0.75	20	1	0.6	0.75
10	1	1	1	21	1	0.8	0.889
11	1	0.6	0.75	22	0.8	0.67	0.729

Table 5.15: Results for COMA++.

characters' like `/` or `-` are determined for every instance of an element. Therefore, the lengths of `1992-03-04` and `06/16/12` are defined by the same constraints. This means that the constraints are matched as well as the schema elements.

In Table 5.16, we show the results for FALCON-AO in detail. FALCON-AO has delivered the lowest recall values in the evaluation because the system does not consider instance values at all. In some of the tasks, data sets had to be matched where the labels of the schema elements differed clearly. The results for FALCON-AO are not comparable to those of our approach or of COMA++ though. However, the few detected correspondences were all correct, which results in high precision.

In Table 5.17, the evaluation results of PARIS are shown. PARIS was able to find correspondences between schema elements with equal or similar labels very easily. In this case, even the instance values do not have to be equal. However, when comparing

Task	Precision	Recall	F-measure	Task	Precision	Recall	F-measure
1	1	0.4	0.571	12	1	0.4	0.571
2	1	0.66	0.795	13	1	0.4	0.571
3	1	0.6	0.75	14	1	0.5	0.667
4	1	0.4	0.571	15	1	0.25	0.4
5	1	0.4	0.571	16	1	0.25	0.4
6	1	0.4	0.571	17	1	0.25	0.4
7	1	0.5	0.667	18	1	0.25	0.4
8	1	0.2	0.333	19	1	0.4	0.571
9	1	0.4	0.571	20	1	0.4	0.571
10	1	0.8	0.889	21	1	0.4	0.571
11	1	0.4	0.571	22	1	0.5	0.667

Table 5.16: Results for FALCON-AO.

Task	Precision	Recall	F-measure	Task	Precision	Recall	F-measure
1	1	0.4	0.571	12	1	0.4	0.571
2	1	0.67	0.802	13	1	0.4	0.571
3	1	0.6	0.75	14	1	0.25	0.4
4	1	0.6	0.75	15	1	0.25	0.4
5	1	0.4	0.571	16	1	0.25	0.4
6	1	0.4	0.571	17	1	0.5	0.667
7	1	0.5	0.667	18	1	0.5	0.667
8	1	0.4	0.571	19	1	0.4	0.571
9	1	0.4	0.571	20	1	0.6	0.75
10	1	0.8	0.889	21	1	0.8	0.889
11	1	0.4	0.571	22	1	0.67	0.802

Table 5.17: Results for PARIS.

different labelled schema elements, PARIS detects only correspondences between those schema elements that contain at least one equal instance value in both data sets. The only constraint seems to be the equality for either instance values or schema elements.

Significance of Results In order to test the significance of the evaluation results, we have conducted a two-sample t-test. These tests can be carried out when it is desired to detect whether the averages of two data sets are significantly different [Bor93]. Since the sample size of our evaluation is rather small, Bortz [Bor93] recommends to perform the t-test instead of other tests. We apply the test on the computed F-measures, since they represent the harmonic mean of precision and recall. We computed the test between our approach and the approach that performed best, which was COMA++. The null hypothesis of our test is that there is no significant difference between the performances (represented by the F-measures of both approaches), while the alternate hypothesis states that there is a significant difference.

We chose $\alpha = 0.05$ as confidence interval, which means that there is 95% confidence that the conclusion of the test will be valid [Bor93]. In accordance to the sample sizes, the degrees of freedom are $df = 42$. The t-statistic value is computed as follows [Bor93].

Theorem 5.1. *T-statistic Value*

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

where

$\bar{x}$ is the mean of each data set;

S is the variance of each data set;

n is the sample size of each data set.

For our evaluation, we have computed a t-statistic value of 5.7395, which is much higher than the corresponding critical t value of 1.684 in the t distribution table [Bor93]. That means that the difference between the F-measures of both approaches is significant. As a result, the alternate hypothesis is correct and our approach performs significantly better than the second-best approach, COMA++.

5.3.5 Discussion and Limitations

The evaluation proves that our approach retrieves better results in this specific regard than other existing matching systems. Since we are covering various patterns for describing instances of typical statistical concepts, our approach can be generalized for statistical data, which may not necessarily be available as Linked Data. However, the implementation of the proposed algorithm and its evaluation in comparison to other matching systems have revealed several aspects for discussion.

Creation of Pattern Classes In our implementation, we have used patterns generated either out of existing definitions from code lists and classifications or manually by domain experts. With respect to the evaluation of our approach, this has been sufficient for investigating its feasibility. But an automatic extraction of patterns is desirable in order to achieve adequate patterns for the specific data sets and to consider unknown patterns. Manually created patterns deviate to include the domain knowledge of the user, which can be interpreted positively or negatively. For an entry like `sex-F`, a user might consider the pattern `[A-Za-z]{3}-[F|M|T|N|U]` because he is aware of other logical entries. An automatic extraction, however, creates `[A-Za-z]{3}-[A-Za-z]`. One can argue that both patterns should be included with a different weighting. However, extracting patterns automatically and allocating them into a pattern class requires a more powerful similarity mechanism for regular expressions than currently implemented. Currently extracted patterns are either added to a new pattern class or discarded if they are already contained in a pattern class. An advanced regex similarity algorithm can enable and support the decision regarding whether an extracted pattern fits an existing class because of its similarity to contained patterns or not. Suitable approaches have been presented in [Pow92, CS11, TQJ11].

Assignment of Weightings A manual assignment of weightings to patterns is a very subjective process. In particular, fine-grained differences between patterns can be determined only empirically. An alternative approach is to calculate the weightings by their appearance in the instance values inside a particular schema element according to the total number of instances in that element. While this process delivers a more objective weighting, the value is always very specific to a particular data set and might be improper for another data set. A solution can be to compute all weightings before each matching process. Conflicts can also be unfolded by calculating different weightings for the same patterns because of their simultaneous occurrence in multiple schema elements of a single data set.

Feature Extraction Currently, our approach is limited to the treatment of patterns of instance values. The results of the evaluation indicate that as the patterns become more generic and string-based, the approach of considering only the patterns may no longer be sufficient. Extracting and analysing features from the instance values promises even better results. Approaches in this direction have been presented in [ZC09]. Another approach is to include schema level information into the matching process. While our implementation tackles a small and very specific subset of relevant elements for schema matching, it is worth considering whether existing matching systems might benefit from our approach and vice versa.

Benchmark and Gold Standard There is no suitable data corpora that can serve as a gold standard for Statistical Linked Data sets. This is why we have chosen to use various real-world data sets published as Linked Data for our evaluation. This has been the only possibility for considering different statistical standards for representing schemata, their elements and instance values as well as different data modelling approaches at the same time. While respecting the popularity of the RDF Data Cube vocabulary [CRT14], the challenge for covering all relevant and typical standards of representing statistical elements still remains.

Instance Values and Datatype Properties Overall, the evaluation has shown that our approach receives good results for instance values. These are typically part of datatype properties. However, we have also observed that entries of code lists and classifications are often linked by object properties. In that case, our approach considers only the URIs of the linked classes, which may not be expressive enough to be used for matching and especially not for being summarized with regular expressions. Additional information within these linked classes like labels, notations, etc. are not considered, even hough they may represent relevant information for finding corresponding schema elements.

In order to complete matching of Statistical Linked Data, we will address the object properties of such data in the following section. Additionally, we will introduce a benchmark for Statistical Linked Data in order to evaluate our approach.

5.4 Object Property Matching Utilizing the Overlap between Imported Ontologies

In the previous section, the focus with respect to matching Statistical Linked Data was on the instance values of datatype properties. The evaluation has revealed that the consideration of URI paths of the object properties within Statistical Linked Data is insufficient for matching object properties, since neither URI paths nor the instances of the linked classes can be summarized adequately by regular expressions (e.g. country names). However, existing matching systems retrieve a rather low recall as well, which we will demonstrate in Section 5.4.7. Finding all correspondences manually is much harder than dismissing the wrong ones.

This systematic shortcoming is due to a high occurrence of heterogeneously labelled object properties, e.g. `ex1:geo` and `ex2:location`. The individuals that are linked to by object properties are not considered to the fullest extent during ontology matching when they are part of external or separate ontologies like, e.g. code lists of country names maintained by a particular authority. Ontologies and instance data that are aligned in current benchmarks and alignment tasks of the Ontology Alignment Evaluation Initiative (OAEI) [OAE] do not yet address this problem. This is verified in Section 5.4.3 by comparing a large number of statistical data sets and the OAEI data sets. This critique on the current limitation on domains for ontology matching is not new. Shvaiko et al. [SE11] suggests that the consideration of new domains will reveal new challenges. Also, according to [Hal05], domain-specific values, significant occurrences, patterns and constraints of values should also be considered.

Based on these ideas and our own findings, we develop a novel ontology matching method to improve the matching of object properties. The method utilizes an instance-based matcher as a core, but refines the results by matching the imported ontologies as well. The similarities between these imported ontologies are computed as an overlap score. This overlap score indicates whether a new correspondence between object properties is added to the generated correspondences between the input ontologies. This method allows us to detect additional correspondences between object properties like `ex1:geo` and `ex2:location` based on the individuals of imported ontologies. Thus, recall is increased. The approach is independent of the matching algorithm employed and may utilize any instance-based approaches or algorithms that consider extensional techniques and object similarity techniques [ES07]. We test different methods to calculate the overlap score: the Jaccard Coefficient and three variants of it, finding that although some of the variants show clearer distinction between correct correspondences and false positive correspondences, the improvements are statistically not significant, particularly when comparing it to the influence of the matcher used.

We formulate our problem statement and distinguish it from current related work in Section 5.4.1. The problem is then adopted to the use case of Statistical Linked Data in Section 5.4.2 and is validated by an analysis of Statistical Linked Data in Section 5.4.3. In Section 5.4.4, we present our proposed method and the applied similarity measures in detail. The evaluation setup is described in Section 5.4.5 and complemented by Section 5.4.6 where a benchmark for Statistical Linked Data is introduced. The results of the evaluation are presented in Section 5.4.7. Finally, we discuss our method and its limitations in Section 5.4.8.

5.4.1 Challenges in Object Property Matching

In the context of Linked Data, the matching of properties is not a trivial task, as [RNX$^+$12] argues, because the instances of two properties are typically described in ontologies that differ from those defining the properties. This observation can be adopted to ontologies when object properties are used to link to classes or individuals of another, imported ontology.

In this section, we focus on the instance-based matching of object properties. Many established methods perform instance-based matching and apply extensional techniques like object similarities. Both, OLA [EV04] and Similarity Flooding [MGMR02], process input ontologies as graph structures and compute proximities between all elements of two graphs. These proximities are propagated throughout the graph structure. However, Similarity Flooding only detects correspondences between nodes of a graph, i.e. classes of an ontology, and does not perform property matching. COMA++ [EM07] contains two instance-based matchers that consider the similarities and patterns of instance values. Pereira Nunes et al. [PNMC$^+$13] presents an approach for matching RDF datatype properties based on the construction of a matrix of the property values. In [LCBF09], the domains and ranges of object properties, the property characteristics and the cardinality restrictions are considered for computing similarities among properties. ASMOV [JMSK09] computes several similarities between properties like internal and extensional similarities. The instance values are part of an overall similarity measure consisting of four calculations. RiMOM [LTLL09] combines multiple strategies for ontology matching automatically and also considers instances for property matching. Detecting correspondences between attributes is also a traditional part in the domain of schema matching [RB01]. In the context of Linked Data, BLOOMS+ [JYV$^+$11] uses contextual information from the input data for matching and a rich knowledge source. While BLOOMS+ focuses on linking classes only, ObjectCoref [HCCQ10] and

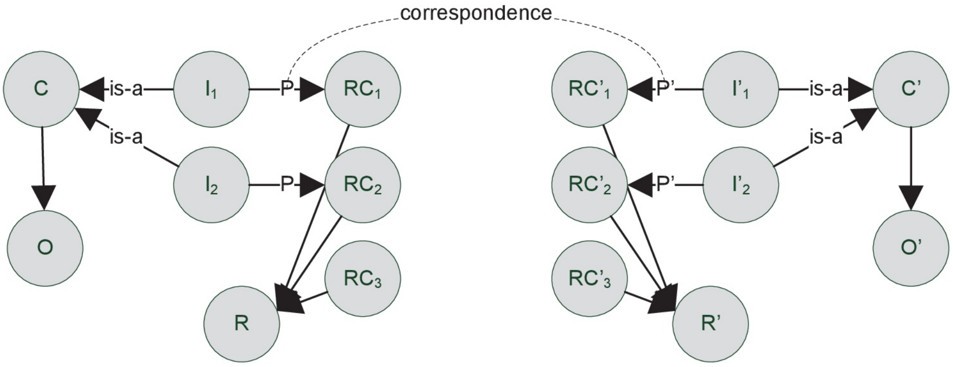


Figure 5.1: Matching of object properties linking to individuals of imported ontologies.

RAVEN [NLAH11] also detect similarities between property values. Additional prominent matching approaches are FALCON-AO [HQ08], AgreementMaker [CAS09], Semint [LC00], GLUE [DMDH03] and Dumas [BN05].

In all the above approaches, only those individuals are considered for matching that are linked in the object properties of the input ontologies. In contrast, our approach identifies and considers additional individuals of an imported ontology that are not linked to in the object properties of the input ontologies. Another specific point of our approach is that we assume the imported ontologies to be sets of homogeneous entities like authority or code lists. This assumption will be verified in Section 5.4.2.

The problem we address in this section is illustrated in Figure 5.1. We assume two ontologies O and O' that hold classes C and C' with individuals I_n and I'_n. We also assume that R and R' are ontologies with homogeneous entities RC_n and RC'_n of the same type, e.g. authority or code lists. The individuals of the ontologies O and O' contain object properties P and P' that link to entities of the ontologies R and R'. When matching ontologies O and O', correspondences between semantically similar object properties P and P' could be missed when they are of different names and structures. This occurs even though both object properties link to individuals of similar ontologies e.g. like `ex1:geo` and `ex2:location` linking to entities of country lists R and R'.

Current approaches consider the individuals linked by object properties for ontology matching; these include ASMOV [JMSK09], RiMOM [LTLL09] and others. However, these referenced individuals play only a subsidiary role in the computation of a correspondence between the linking properties. Also, individuals of such imported ontologies that are not linked to are not considered. Thus, correspondences between object properties can be missed.

5.4.2 Use Case: Statistical Linked Data

The use case, which supplements our problem statement, centers on Statistical Linked Data. Scientists often integrate and merge two or more of these data sets in order to

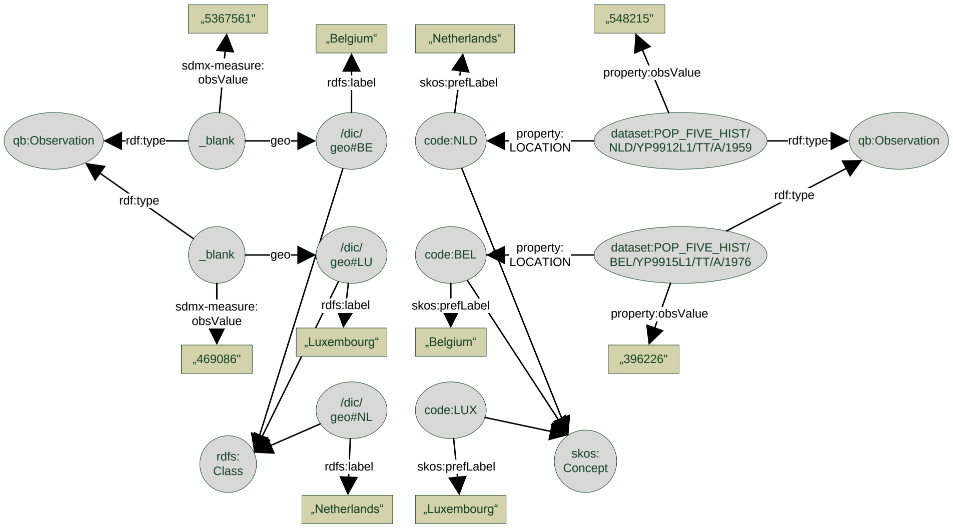


Figure 5.2: Example of statistical data represented as Linked Data.

conduct comparative data analysis. In theory, ontology matching is the ideal method for this task; however, in practice, we show how matchers can be improved to give better results for this scenario.

In Figure 5.2, the problem of object property matching stated in the previous section is illustrated using Statistical Linked Data as it is defined in Section 2.2. Excerpts of two data sets from Eurostat[17] and OECD[18] are shown. The instances of both data sets hold an object property indicating some geographical information (`geo` and `property:LOCATION`). Other object properties are omitted here. The object properties link to other individuals of code lists, which are indicated by a different URI path and a different namespace. In Figure 5.2, the referenced data sets also contain the individuals `/dic/geo#NL` and `code:LUX`, which are not linked to by object properties. In our tests involving matching systems, the object properties `geo` and `property:LOCATION` are not matched because they are labelled differently and belong to different data sets. Even when matching the individuals of the referenced code lists, there is no inference on the referencing object properties.

5.4.3 Analysis of Patterns of Statistical Linked Data

In order to validate whether our problem statement is reasonable for the domain of statistical data by affecting a large number of data sets, we verify our assumptions on the patterns of statistical data, which are:

1. Data entries are modelled as individuals and are accompanied by various named

[17] http://estatwrap.ontologycentral.com/
[18] http://oecd.270a.info/

object properties linking to classes and individuals of external code lists or lightweight ontologies. Rather than form a network or tree connected with homogeneous object properties, the data model is similar to a star schema[Inm05].

2. Classifications and code lists[19] are often used in statistical data sets in the described way.

3. These code lists are referenced by object properties and are identifiable as additional ontologies or data sets by inspecting namespaces and URIs.

We verify our assumptions by analysing and comparing data from three sources. Real-world data sets are considered from two of the main repositories for Open Data: Data Hub[20] (DH)[21] and the wiki of Planet Data[22](PD). They are compared to data sets used in previous campaigns of the OAEI [OAE] to show that this is, in fact, a novel problem, not one that has been investigated. Within this third set, we examine data sets of the instance matching (IM) tracks separately due to major differences between ontologies and data sets containing mostly instance data. Duplicate data sets – e.g. ontologies that have been used for several years in the OAEI or in multiple tracks – have been omitted. Due to the diversity of the data sources, the data analysis was done manually with the help of standardized SPARQL queries and scripts. Depending on how the data sets are published, we have either performed SPARQL queries over a data set or executed the `listStatements()` method of the JENA API package on a dump file of the data set.

Criteria	DH	PD	OAEI	IM
Number of all examined data sets	49	22	54	15
Data structure	93,8	95,4	0	13,3
Presence of classification references	91,8	95,4	3,7	13,3
OWL/RDF data set	0	0	90,7	40
Other RDF-based data set	100	100	24,1	73,3

Table 5.18: Comparison of Statistical Linked Data and OAEI data (as of December 2013).

We investigate our data structure hypothesis by examining whether the data set is organized similar to our assumed pattern. The structure is detected by analysing and counting the links inside a data set and out to other data sets. The results in Table 5.18 show that most of the examined data sets from Planet Data and Data Hub reflect this typical structure of statistical data, but almost none from the OAEI and IM challenges. Based on the identified schema structure, we then investigate whether links to ontologies similar to code lists can be identified. References to code list entries could be observed

[19] With regard to their similar function for statistical data, classifications and code lists are summarized as code lists for the entire paper.

[20] http://thedatahub.io/

[21] Due to the quantity of data sets, Data Hub has been analysed by sampling. Data sets have been examined that are tagged with 'format-rdf', 'format-qb', 'format-scovo' as well as 'statistics', 'government', 'census', or 'lod' and similar spellings. Duplicates have not been counted.

[22] http://wiki.planet-data.eu/web/data sets

Criteria	Percentage
Number of all examined data sets (sample from DH and PD)	40
Different NS for input and referenced ontologies	67,5
URI path of linked individuals equal for particular object properties	100
Individuals of a referenced ontology of the same class	100

Table 5.19: Analysis of the structure of Statistical Linked Data (as of December 2013).

in most cases of the Data Hub and the Planet Data data sets (see Table 5.18). The detected code lists have a list-type character, like country lists, age groups, or entries of a scale. Only in a few cases are hierarchies within these code lists, e.g. in a geographical classification with different administrative levels. In the OAEI and IM challenges, only in two cases could references to code lists be detected, i.e. object properties that link to individuals of imported ontologies.

Finally, we examined whether the different ontologies of the detected structure can be distinguished by different namespaces and URIs. The individuals of an ontology are considered to be defined in one namespace. Moreover, the classes and individuals of an imported ontology have to be addressed by the same object property of the input ontology. The results in Table 5.19 show that this is indeed the case. However, some ontologies may be subsumed under one namespace since they can be distinguished by different URI paths. For all studied data sets, observing the URI path was sufficient for identifying and distinguishing the ontologies.

5.4.4 Approach and Similarity Measures

Knowing the structural differences between current benchmarks and statistical data according to our problem statement leads us to the following algorithm to improve the matching of object properties. Revisiting our use case, we complement the matching process of the two data sets by identifying those code lists that contain the referenced individuals like `/dic/geo#BE` and `code:NLD`. Then, the overlap between these code lists is computed, which we conjecture to represent a semantic similarity between the object properties `geo` and `property:LOCATION`. This is used as correspondence for the overall matching between the data sets.

The algorithm is formalized as follows. Given as input are two ontologies O and O' with classes C and C', properties P and P', and individuals I_n and I'_n. The objects RC and RC' of the object property instances are classes or individuals of imported ontologies R and R'. These are ontologies with homogeneous entities of the same type, e.g. code lists. The imported ontologies are either T-Boxes or A-Boxes of their own with different namespaces. Thus, based on the data analysis conducted in Section 5.4.3, we formulate the following definition of *Object Property Instances* and *Property Objects.*

Theorem 5.2. *Object Property Instance and Property Object*

An instance OPI of an object property P is a tuple of the form $\langle I, P, RC \rangle$, where I is an individual of ontology O and P is the particular object property of O. A property object RC of OPI is a class or individual of a referenced ontology R.

Furthermore, we define *Imported Ontologies* as follows.

Theorem 5.3. *Imported Ontology*

An imported ontology R is either a T-Box or an A-Box ontology with classes or individuals RC that are objects in the object property instances OPI of the ontology O. An imported ontology R and its entities RC are held in a namespace different from the namespace of O and all its entities.

The objective of our algorithm is to detect an alignment A as output with correspondences between all entities of O and O'. Additionally, overlaps between all R_n are used in order to generate additional correspondences between object properties P and P'. In the algorithm, we apply any given ontology matching system that generates correspondences between two input ontologies. As mentioned earlier, the matcher is used as a black box in our algorithm. The algorithm goes through five phases for matching two input ontologies O and O'.

1. All RC inside each ontology are grouped in order to identify the imported ontologies R_n and R'_m per each ontology O and O'.
2. The input ontologies O and O' are matched by an ontology matching tool. The resulting correspondences are included in the alignment A.
3. All pairs of R_n and R'_m are matched with each other by the same matcher. The resulting correspondences are the basis for calculating the overlap scores in the next phase.
4. Overlap scores are computed pairwise for each R_n and R'_m. Different similarity measures can be applied. We utilize the Jaccard coefficient [vR79, DMD$^+$03]. However, the Jaccard coefficient is known for its unbalance [IMSW07], especially when two sets are highly different in their size. This may complicate the choice of a suitable threshold. Hence, we introduce three additional similarity measures for addressing this problem in Theorem 5.4. The overlap between two ontologies is computed by assuming that a correspondence between two individuals of the ontologies indicates that they are part of the intersection set of R_n and R'_m. In this way, we can determine $|R_n \bigcap R'_m|$. If the overlap is higher than a specific threshold t, we assume that there is a correspondence between the object properties P and P' that hold R_n and R'_m as objects in OPI and OPI'.

5. We add the detected correspondence with the calculated overlap score between their imported ontologies R_n and R'_m as confidence value to the alignment A. If a correspondence between two object properties already exists in A, the correspondence with the higher confidence value is retained and the other one is discarded.

Theorem 5.4. *Overlap utilizing Jaccard Coefficient and Variations*

The overlap between two imported ontologies R_n and R'_m is computed as

$$JC = \frac{|R_n \cap R'_m|}{|R_n \cup R'_m|}$$

$$JC_{min} = \frac{|R_n \cap R'_m|}{|min(|R_n|, |R'_m|)|}$$

$$JC_{res} = \frac{|R_n \cap R'_m|}{|R_{n-Linked} \cup R'_{m-Linked}|}$$

$$JC_{min+res} = \frac{|R_n \cap R'_m|}{|min(|R_{n-Linked}|, |R'_{m-Linked}|)|}$$

where

$|R_n \bigcap R'_m|$ *is the number of all correspondences between R_n and R'_m;*

$|R_n \bigcup R'_m|$*is the number of all entities in R_n and R'_m;*

$|R_{n-Linked}|$ *and* $|R'_{m-Linked}|$ *is the number of those classes of R_n and R'_m that are linked in the ontologies O and O'.*

The computation of similarities between ontologies is discussed in several works. In [MS02], the ontology similarity is based on the terminological similarity of concepts. Different similarities are combined in [EHHS05], where strings, concepts and usage traces are considered. Stuckenschmidt [Stu09] presents a calculation involving two A-Box ontologies, while also considering structural information from their T-Box ontology. Similar to our method is [DEvZ10], where several measures are introduced for computing ontology similarity by considering available alignments. In [DMD$^+$03], the Jaccard coefficient is introduced as a similarity measure for ontology matching. According to [IMSW07], simple similarity measures like the Jaccard coefficient perform best for instance-based matching, which is why we chose it as our method.

Theorem 5.5. *Correspondence between Two Object Properties*

A correspondence between two object properties P and P' is described by the following five-tuple adapted from [ES07].

$$\langle id, e_1, e_2, r, n \rangle$$

where

id is an identifier for the particular correspondence;

e_1*and* e_2 *are the object properties* P *and* P'*;*

r *determines the type of the relation between* P *and* P'*, in our case an equivalence relationship;*

n *represents the confidence values, which in our case is the overlap*(R_n, R'_m).

This method is simple to implement with any instance-based matcher and enables us to match object properties like `geo` and `property:LOCATION` in our example. The runtime is comparable to matching the whole ontologies. The split between the different ontologies decreases the time needed for matching the particular ontologies, offsetting the need to run additional matching processes.

5.4.5 Evaluation Setup

We evaluate our method on both artificial and real-world data to demonstrate the impact of our method on object property matching. The results show a significant improvement in both scenarios, especially the sought-after improvement of recall.

The evaluation consists of two scenarios. The first scenario *Benchmark* is conducted on an artificially created benchmark for statistical data that is introduced in Section 5.4.6. In the second evaluation scenario *Real-world Data*, we apply our method on the two real-world data sets from Eurostat and OECD from our use case. In each scenario, the matching systems are executed with the input ontologies at first (*State-of-the-Art*). In a second run, our method (*Object Property Matching*) is applied by matching the imported ontologies additionally.

In both scenarios, the resulting correspondences are validated with their particular reference alignments. We compute precision, recall and F-measure for each alignment task, since they are standard evaluation measures for ontology matching evaluation [ES07]. For computing the overlap value, we utilize a threshold of 0.3 in the benchmark scenario. This has turned out to be a suitable value during pretests. Since the Jaccard coefficient can get unbalanced [IMSW07], we compare the different similarity measures defined in Section 5.4.4 in the second scenario.

We chose FALCON-AO [HQ08] and AgreementMaker [CAS09] as the black box matcher from which we assume representative results. FALCON-AO has been chosen because it applies extensional matching techniques like object similarity, while AgreementMaker contains an instance-based matching algorithm. Our instance-based object property matching approach is compared best to those techniques. Both systems have been successful regarding their performances in previous OAEI campaigns [EFM$^+$10, EFvH$^+$11] and are executed without any manipulation in their standard configurations.

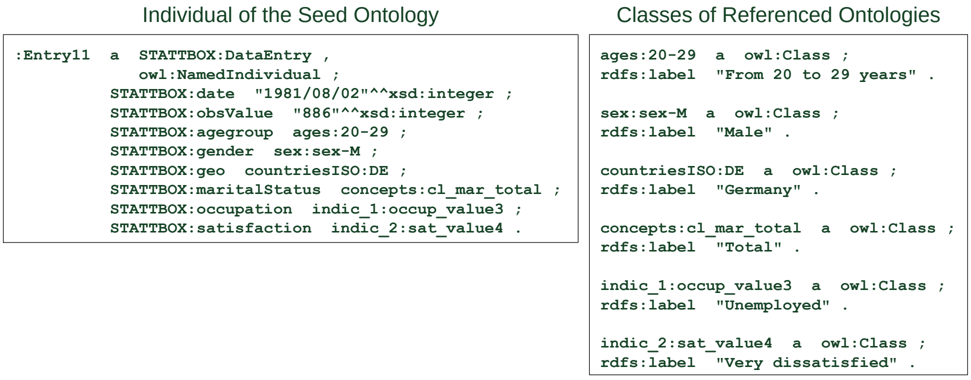
Real-World Data For the *Real-World Data* scenario, we revisit the data sets from our use case. These hold many different properties that semantically overlap and are representative for statistical data. The idea is to examine many different cases in just one pair of data sets, as the preparation is quite labour-intensive. The EUROSTAT data set covers *Labour input in industry.* This data set has 16,783 instances and seven object properties. The OECD data set covers *Outward activity of multinationals – Share in national total (manufacturing).* It has 5,343 instances and eight object properties. In both data sets, the object properties link to classes of particular code lists. Also, both data sets have some object properties that are not linked inside the actual instances. We manually identified five properties that match semantically. In order to use the code lists with the matching systems, they had to be preprocessed. The changes include generic transformations of the referenced code lists from SKOS to T-Box ontologies. Similar preprocessing has been carried out previously in Library Tracks[23] of the OAEI, where SKOS thesauri have been transformed to OWL. The reference alignment, which serves as a gold standard for the evaluation, has been done manually by domain experts. The data is available at `http://code.google.com/p/matching-statistics/`.

5.4.6 Benchmark for Statistical Linked Data

It was not possible to evaluate our method on a gold standard because, unfortunately, no such standard exists yet. This has also been a drawback for the evaluation in the previous section. The benchmark data set of the OAEI is an established source for evaluating ontology matching approaches. Based on an ontology describing bibliographic resources, it covers various kinds of transformations on structural and terminological levels and is used for different alignment tasks. The Islab Instance Matching Benchmark (IIMB) [FLMV08] has been created for evaluating instance matching systems. Both benchmarks are well received in the ontology matching community. However, both benchmarks do not consider the underlying problem of our approach. Similar to the data in our use case is the RDF version [KH13] of the Star Schema Benchmark [OOC09], which comprises five single data sets. However, while the distributed structure is similar to the structure of Statistical Linked Data, it is not processable by most of the current ontology matching systems. Hence, we decided to design a benchmark specific to the problem based on the principles of established benchmarks [OAE, FLMV08].

The benchmark reflects the assumptions made in Section 5.4.3 concerning heterogeneous object properties and their linking to classes of code lists, located in other namespaces and URI paths. The T-Box is a simplified version of a data model for statistical data: one named class representing a data entry and several object properties linking to classes of imported ontologies. These imported ontologies are included with different URI paths. We populate this seed ontology with 50 randomly generated individuals as A-Box. An example is given in Figure 5.3: `:Entry11` represents an observation on the satisfaction

[23] see http://oaei.ontologymatching.org/2007/, http://oaei.ontologymatching.org/2008/library/, http://oaei.ontologymatching.org/2009/library/, and http://web.informatik.uni-mannheim.de/oaei-library/2012/



Individual of the Seed Ontology

```
:Entry11  a  STATTBOX:DataEntry ,
             owl:NamedIndividual ;
          STATTBOX:date  "1981/08/02"^^xsd:integer ;
          STATTBOX:obsValue  "886"^^xsd:integer ;
          STATTBOX:agegroup  ages:20-29 ;
          STATTBOX:gender  sex:sex-M ;
          STATTBOX:geo  countriesISO:DE ;
          STATTBOX:maritalStatus  concepts:cl_mar_total ;
          STATTBOX:occupation  indic_1:occup_value3 ;
          STATTBOX:satisfaction  indic_2:sat_value4 .
```

Classes of Referenced Ontologies

```
ages:20-29  a  owl:Class ;
rdfs:label  "From 20 to 29 years" .

sex:sex-M  a  owl:Class ;
rdfs:label  "Male" .

countriesISO:DE  a  owl:Class ;
rdfs:label  "Germany" .

concepts:cl_mar_total  a  owl:Class ;
rdfs:label  "Total" .

indic_1:occup_value3  a  owl:Class ;
rdfs:label  "Unemployed" .

indic_2:sat_value4  a  owl:Class ;
rdfs:label  "Very dissatisfied" .
```

Figure 5.3: Example individual of the seed ontology.

Tests	Variations
001	Duplicate of seed ontology
010-011	Names of object properties
020-024	Labels, class names, URIs, namespaces of imported classes and ontologies
030-031	No overlap

Table 5.20: Variations within the benchmark.

level of young German adults. The object properties link to classes of code lists from different namespaces[24].

The seed ontology is used to produce variations. The namespaces of all involved code lists, i.e. imported ontologies, were changed. Additionally, specific properties were changed in accordance with our observations regarding statistical data. In the variations 010-011, the names of the object properties are changed on a random basis. In 020-024, the code lists that are referenced are changed with respect to label name, class name, URI path, etc. This notably lowers the overlap. In 030-031, we test the matching without any overlap, to see how our system works on standard ontologies. Each variation forms together with the seed ontology an alignment task. The variations are summarized in Table 5.20. The complete benchmark, the variations and the single tests are available at `http://code.google.com/p/matching-statistics/`.

[24] In the actual benchmark, they are differentiated by the URI path and not by the namespace as this has the greater coverage on statistical data. This example uses namespaces for clarification. For the algorithm, there is no practical difference.

5.4.7 Results

The results in Table 5.21 indicate major improvements on matching object properties in all scenarios[25]. The results of the tests 010-011, which hold differently labelled object properties, reveal the strengths of our method compared to the state-of-the-art. The matchers could not find correspondences between the heterogeneous object properties, even if their referenced individuals are equal or similar, like `concepts:geo#geo_DE` and `vocab:country#DE`. The information given in the labels of these classes is not considered for detecting correspondences between the referring object properties. The recall of our method is much higher for these tests. The results of the tests 020-024 show that the distance between our method and the state-of-the-art is decreasing depending on the matching between the imported ontologies and the different resulting overlaps. However, the results of these tests are always better or at least equal to the state-of-the-art approach when utilizing our method. This is also demonstrated with the counter check 030-031 (no overlap), which shows at least no worsening.

Approach	State of the Art (SotA)						Object Property Matching					
System	AgreementMaker			FALCON-AO			AgreementMaker			FALCON-AO		
	P	R	F	P	R	F	P	R	F	P	R	F
Test 001	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Tests 010-011	1.00	0.45	0.62	1.00	0.34	0.51	**1.00**	**0.89**	**0.94**	1.00	0.78	0.88
Tests 020-024	1.00	0.42	0.59	1.00	0.29	0.45	**1.00**	**0.85**	**0.92**	1.00	0.67	0.80
Tests 030-031	**1.00**	**0.45**	**0.62**	1.00	0.34	0.51	**1.00**	**0.45**	**0.62**	1.00	0.34	0.51
Real-World Data	1.00	0.40	0.57	1.00	0.40	0.57	**0.83**	**1.00**	**0.91**	0.45	1.00	0.62

Table 5.21: Results for both evaluation scenarios.

The results using real-world data are similar to the benchmark tests 020-024, because there are not necessarily any overlaps between the code lists. The object properties in both data sets are named differently, the number of classes in all code lists is unbalanced, and there may not necessarily be correspondences between all object properties. While recall improves, there is some loss of precision (see Table 5.21). False positives occur when the matchers find correspondences between dissimilar code lists, e.g. `geo` (containing country names) of Eurostat with `property:ISIC3` (containing branches of industry) of OECD. Nevertheless, the higher recall shows that our method has detected new correspondences that have not been identified by the state-of-the-art approach.

[25] The best result per row is bold. For the single tasks of the variations, the means have been computed. P = Precision, R = Recall, F = F-measure.

Significance of Results

We have conducted a two-sample t-test as we have carried out in Section 5.3.4. We apply the test on the computed F-measures, since they represent the harmonic mean of precision and recall. The null hypothesis is that there is no significant difference between the performances of the state-of-the-art approach and our approach (represented by the F-measures of both approaches), while the alternate hypothesis states that there is a significant difference. We chose $\alpha = 0.05$ as confidence interval, which means that there is 95% confidence that the conclusion of the test will be valid [Bor93]. In accordance to the sample sizes, the degrees of freedom are $df = 18$. The t-statistic value is computed as in Theorem 5.1 [Bor93] defined. For our evaluation, we have computed a t-statistic value of 40.391, which is much higher than the corresponding critical t value of 1.734 in the t distribution table [Bor93]. That means that the difference between the F-measures of both approaches is significant. This result suggests that our approach performs significantly better than the state-of-the-art approach.

Cutting off False Positives

In order to cut off the detected false positives, we choose a threshold value. Since the unbalance of the simple Jaccard coefficient makes it difficult to set a suitable threshold, we have compared the similarity measures defined in Theorem 5.4 regarding their impact on the real-world data scenario.

Found Correspondences	JC	JC_{min}	JC_{res}	$JC_{min+res}$	SotA
Correct Correspondences					
geo = LOCATION	0.002	1	0.688	1	0
indic_bt = VAR	0.132	0.909	1	1	0
nace_r2 = ISIC3	0.006	0.979	0.959	1	0
obs_status = OBS_STATUS	0.75	1	x	x	0.969
timeformat = TIME_FORMAT	0.571	1	1	1	0.872
False Positives					
geo = ISIC3	0.571	1	1	1	0.872

Table 5.22: Similarity measures for detected correspondences (AgreementMaker).

The different overlap values computed for each detected correspondence are shown in Tables 5.22 and 5.23 and are compared to the confidence values of the state-of-art approach (SotA). An x means that no value could have been computed, because no classes of the referenced code lists have been linked in the data set (this would result in a division by zero). Balanced values make it easier to distinguish false positives because the difference between valid correspondences and non-valid correspondences is increased. For example, the overlap for the correspondence between geo and LOCATION is at 0.002 for JC, but this is much higher for the others. The best approach, in this sample, would

Found Correspondences	JC	JC_{min}	JC_{res}	$JC_{min+res}$	SotA
Correct Correspondences					
geo = LOCATION	0.002	0.909	0.588	0.909	0
indic_bt = VAR	0.012	0.090	0.083	0.333	0
nace_r2 = ISIC3	0.007	0.188	0.103	0.191	0
obs_status = OBS_STATUS	0.647	0.917	x	x	1
timeformat = TIME_FORMAT	0.571	1	1	1	1
False Positives					
geo = ISIC3	0.004	0.354	0.340	1	0
nace_2 = VAR	0.001	0.090	0.017	0.1	0
nace_2 = OBS_STATUS	0.012	0.938	0.306	0.306	0
freq = VAR	0.053	0.111	0.1	0.1	0
freq = OBS_STATUS	0.389	0.778	x	x	0
freq = TIME_FORMAT	0.3	0.75	1	1	0

Table 5.23: Similarity measures for detected correspondences (FALCON-AO).

be to use $JC_{min+res}$ and to use JC_{min} when that fails. However, the actual effect is minimal. Only one false positive is excluded. More thorough testing might bring a clearer distinction. So far, it seems that the choice of the similarity measure to compute an overlap is much less relevant than the choice of the matcher to increase precision.

5.4.8 Discussion and Limitations

We have shown that object properties in statistical data are used differently than in data sets typically used for ontology and schema matching. By leveraging this difference for object property matching, we gain an improvement of recall up to 2.5 times. Loss in precision occurs, but is relatively small in comparison. Since this loss occurs while matching the imported ontologies, adjusting the matching systems towards this problem may be helpful. For these experiments, we have used the standard parameters for both matchers in order to keep it clearer.

While our use case has been motivated by statistical data, a lot of Linked Data sources share this data model structure since many of them are derived from relational databases. We chose statistical data because (1) there is a clear need to integrate the data and (2) although the data sets cover semantically similar topics, standardization usually does not cover the object properties, only the code lists themselves, if at all. This demand may increase with the number of LOD sets.

The technique is simple to implement with any instance-based matcher. The runtime is comparable to matching the whole ontologies.

5.5 Summary

In this chapter, we have presented three methods that improve the process of data matching with Statistical Linked Data **(1a)**. Our methods can be applied during either preprocessing (assessment tests in Section 5.2) or the matching process itself (Sections 5.3 and 5.4). By considering instance values (datatype properties) and code lists (object properties) separately with specialized treatment, we achieve a complete consideration of Statistical Linked Data for the matching process **(1b/1c)**. The assessment tests for Statistical Linked Data allow for determining the suitability of that data for the matching process **(2a)**. Since these tests can be carried out before the actual matching process begins **(2c)**, the researcher may spend less time on inspecting the data manually, which is beneficial when he has a limited technical background and is not familiar with the RDF format and structure **(2b)**.

We have shown that our methods for data matching apply better on Statistical Linked Data than existing approaches and with a significant improvement in recall **(3a/3b)**. The t-tests that were conducted after the evaluations supported these results. However, the presented methods in Sections 5.3 and 5.4 can also be applied to other Linked Data sets and ontologies that hold a similar structure or similar characteristics, e.g. data originated from relational databases often holds a distributed structure, since it is spread over multiple tables **(4b)**. A drawback of our methods is that they focus, in particular, on the special characteristics and structure of statistical or similar data, which means that they may not perform better than state-of-the-art approaches when being applied on completely different data sources **(4a)**.

In the next chapter, we will conclude and summarize our results regarding the investigation of methods for matching Linked Open Social Science Data. We will revisit our research questions and review our contributions in detail. We will discuss whether such methods can serve as door openers for wider applicability of Linked Open Social Science Data. Furthermore, we will present future possibilities for research.

6 Conclusion and Future Work

In Chapter 1, we reflected the current expectations of the Semantic Web community and science politics that LOD and Semantic Web technologies may have a promising impact on various scientific disciplines. This anticipation is based on an increasing volume of available Linked Data sets as well as on standardization, integration and interoperability reasons. This thesis sought to investigate methods for the data matching of statistical data as an example for a possible applicability of LOD in the social science domain, envisioned as Linked Open Social Science Data. Although the available tools and technologies are mature enough, we identified by conducting expert interviews and developing a prototype application in Chapter 3 that there is currently no consumption of Linked Data in the social sciences at all. This insight led us to the identification of five topics of interest in Section 3.3 – namely, data modelling, data access, data retrieval, data matching, and data interaction – that would need further investigation in with regard to our objective. Considering the objectives of this thesis, we continued our work by focusing on the publication of Linked Open Social Science Data in Chapter 4, which is the basis for data matching. In Chapter 5, we developed methods that support and improve data matching with Statistical Linked Data.

In this Chapter, we conclude this thesis and review the contributions of our work together with the objectives and research questions formulated in Sections 1.2 and 1.3. Additionally, in Section 6.2, we provide an outlook on future research that extends our previous work seamlessly and creates further inroads to enable the applicability of Linked Open Social Science Data.

6.1 Summary of Contributions

In this section, we summarize the contributions of this thesis in consideration of our research questions and show that our work on the publication and matching of Linked Open Social Science Data supports its applicability and consumption.

1. General Applicability of LOD for Data Matching

Revisiting the first block of our research questions, we can conclude that Linked Data can be used for data matching **(1a)**. We have contributed in this direction by allowing to assess Statistical Linked Data regarding its usage for scientific research (see Section 5.2) and by introducing two novel approaches (see Sections 5.3 and 5.4) that focus particularly on special data patterns of Statistical Linked Data. The first approach considers datatype properties and patterns of instance values in particular, while the second method focuses on object property matching. The results of all three approaches are promising. Especially for the two matching approaches, significant improvements in recall were measured.

However, the distributed structure of Statistical Linked Data and an extensive use of code lists and occurrence of codes in instance values (see Sections 2.2, 5.3.1 and 5.4.3) **(1b)** can decrease the matching results especially when schema elements, i.e. the datatype and object properties of Statistical Linked Data, are labelled completely differently. These characteristics of Statistical Linked Data are still not fully considered by current matching systems **(1c)**. Numerous tools for schema and ontology matching (see Section 2.3) as well as for link detection and discovery (see Section 2.1.1) enable matching Linked Open Social Science Data on the schema level and between single entities. However, the focus of current matching tools on particular types of data, which are primarily used for evaluation, is a drawback when adapting them for use with Statistical Linked Data.

2. Suitability of LOD for Data Matching

Considering the second block of research questions, we developed assessment tests for Statistical Linked Data in Section 5.2. These tests allow for determining whether one or two Statistical Linked Data sets are valid regarding basic requirements for statistical data and whether they can be matched technically and semantically **(2a)**. Since, data matching can be a time-consuming job, these tests can be executed before the actual matching process begins **(2c)**. Since the results of the tests provide insight regarding whether dimensions of one or two data sets – with respect to their instance values – can be combined or compared with each other, the impact of these tests is that researchers may spend less time on inspecting the data sets manually **(2b)**, which is, in particular, a benefit when the data sets are large and the researcher has a limited technical background and understanding of the RDF format and structure.

3. Influence on the Matching Result

The development of two matching approaches provided insight into the third block of research questions. In both the approaches presented in Sections 5.3 and 5.4, we considered the investigated particular structure of Statistical Linked Data and achieved significant improvements in recall **(3a/3b)**. Although a minor loss in precision was detected when matching real-world Statistical Linked Data sets, the improvement in recall and F-measure outweighed this loss. In addition, the results of the t-tests that were conducted after the evaluations show empirically that our methods perform significantly better than the state-of-the-art approaches. However, since in social research, data sets are often merged or integrated according to one or two particular key variables, i.e. one or two particular schema elements of data sets, a high recall of the matching result is important; this means that the most correct correspondences between two data sets must be detected.

4. Limitations of the Approach and Applicability on Other Data Sources

Regarding the fourth block of research questions, the developed approaches are not limited to the domain of the social sciences, since statistical data is analysed in other domains as well **(4b)**. The developed methods can be applied not only with Statistical Linked Data, but also with data holding a similar structure or characteristics. Moreover, the method on object property matching presented in Section 5.4 can also be applied with any instance-based matching system, since the matching system is used as a black box in this approach. The requirement for data with a similar structure or characteristics is simultaneously the drawback of our methods **(4a)**.

5. Beneficial Use of LOD in the Social Sciences

Considering the fifth block of research questions (see Section 1.3), the results of the first four research questions prove that there can be a beneficial use of LOD in the social sciences. In particular, a value addition can be observed by the improvement on the matching results **(5a)**.

A meaningful application of Linked Open Social Science Data depends on use cases that define the functionality that a potential application should have. While numerous applications for Linked Data exist that enable the browsing and visualization of data (see Section 2.1.3), only a few applications focus on use cases out of a scientific domain [vHEM12, Pau12, KOH12]. Based on our use case of 'Analysing Research Data', we have developed and implemented a framework for a Semantic Data Library, a specialization of a digital library for exposing, retrieving and combining research data using Semantic Web technologies in Section 3.2. This framework focuses on the actual consumption of Linked Data comparable to digital libraries and is independent of a particular domain. By using and adapting technical standards and interfaces like RDF, SPARQL and JSON, among others, our framework shows that Linked Open Social Science Data can easily be reused and processed in other established tools and interfaces **(5a)** like the Google Visualization API[1], Vizgr [HZSM12], the statistics tool R Project[2] or Solr Search[3].

Although neither consuming nor being familiar with LOD, all participants of the expert interviews were convinced that especially data modelling and its publication on the web as well as matching and consuming may benefit from Semantic Web technologies (see Section 3.1.3) **(5a)**. The results of the expert interviews presented in Section 3.1.3 revealed that there is currently no consumption of LOD in the domain of the social sciences. On the one hand, this insight may be interpreted as a surprise considering the amount of relevant

[1] https://developers.google.com/chart/interactive/docs/reference

[2] http://cran.r-project.org/

[3] http://lucene.apache.org/solr/

Linked Data sets and the self-confidence of the Linked Data community regarding the maturity and ready-to-apply potential of Linked Data technologies; on the other hand, this observation could also be presumed because of the novelty of LOD compared to the maturity of the social research itself and of already applied technologies and tools. When a researcher is used to knowing which agency provides data, how it can be accessed, and when processing and analysing this data is an established, though not always satisfying working process, there is – at least at first sight – no convincing reason for trying new technologies and new data sources. Nevertheless, the experts have evinced interest in the idea of LOD and its prospects.

The complete publication of Linked Open Social Science Data is a necessary precondition for its further consumption, e.g. through data matching **(5b)**. A publication of Linked Open Social Science Data is enabled technically in the first place by the semantic technologies empowering LOD (see Sections 2.1.1 and 2.1.2). However, an adequate and complete publication of Social Science Data (see Definition in Section 1.1) as LOD takes place only if vocabularies and ontologies consider and cover particular requirements of the domain, i.e. its processes, data and structures. Beside ongoing work on representing scientific related data as LOD [KdE11, CRT14, VIV, AKT01], there were previously no ontologies for meeting adequately the particular requirements given by the social science domain. The results of Chapter 4 fulfil these requirements by allowing the representation of fine-grained micro data using the DDI-RDF Discovery vocabulary (see Section 4.3), by completing the research process with the inclusion of research data (see Section 4.2) and by enabling consistent content indexing with a domain-specific thesaurus (see Section 4.4).

Applicability to Other Domains

Although encouraged in the social sciences, our results are not restricted to this domain, since the completion of processes, data and structures for a comprehensive publication of data is also relevant for other domains. Person-level micro data, which is particularly addressed by the DDI-RDF Discovery vocabulary, is also part of other scientific domains like economics and behavioural sciences as are complex relationships in thesauri and controlled vocabularies are. In addition, the developed methods for data matching are not restricted to the social sciences, since statistical data is also analysed in other domains.

Additional Findings

While working on the objectives of the thesis, we made the following additional findings. The expert interviews and the work on the prototype application for a Semantic Data Library (presented in Sections 3.1 and 3.2) revealed that there is still a lack in the retrieval of LOD. This is a major obstacle for applying Linked Open Social Science Data when considering that a user cannot be aware of all available and potentially interesting Linked Data sets. Searching for data is a relevant step in research (see Sections 1.2 and

3.1.3), which cannot be satisfied by browsing data catalogues like Data Hub[4], where data is tagged with free keywords chosen by the data provider. However, promising work has started regarding this challenge related to the retrieval of LOD [FCOO12, HL10, LT10].

Research interests are widely spread. This can be seen as an obstacle in terms of data linking. It is difficult to identify to which Linked Data sets a matching might be meaningful and useful. The interests in relevant context information are spread widely as are the interests for combining multiple data sets (see Section 3.1.3). Thus, it is important to provide easy-to-use matching and linking tools that can be executed by users with a limited technical background and understanding of LOD and its underlying techniques.

Until now, consuming Linked Data requires a certain amount of technical expertise from the user. However, from a user's perspective, at least in the case of the experts in our interviews (see Section 3.1.3), it is not important to know how the data is technically published; rather, it is relevant to know which is the data source (provenance, which is a data modelling task), whether it is accessible and how it can be further processed. The last step is preferentially performed in familiar and established applications, e.g. statistics tools, since some operations on data may prove very complex. However, there are still only a few connections to established tools outside of the Semantic Web like the SPARQL plugin[5] for the R Project. On the other hand, another reason for the lacking application of Linked Data lies in the Semantic Web tools themselves. A lot of tools, e.g. in the area of ontology and schema matching, are mature in a technical sense, i.e. they perform more or less well for particular technical aspects, which is shown in evaluations and with used data sets. However, when using Linked Data that has been published out of the context of these tools, e.g. Statistical Linked Data, it is difficult to perform these tools on such data because they differ from the data used in evaluations (see Section 5.3). There is still a gap between Semantic Web tools and the capability of using Linked Data, which has been generated out of other data sources and for other reasons. This issue needs to be addressed because at least Semantic Web tools should be capable of processing Linked Data.

Another reason for the hesitant application for LOD in general and Linked Open Social Science Data in particular may lie in too high or incorrect expectations. According to its initial and original idea [BL06], Linked Data describes methods for publishing and linking data on the web. Since then, numerous architectures, tools and applications have been developed that enable the storage, maintenance, consumption and interaction of Linked Data. However, they can only serve as a general pattern or guideline on how to realize an application of LOD. A meaningful application for a particular community or domain does not appear right from the start. For application in a particular discipline, it is fundamental to adapt the technical concepts of the underlying Semantic Web technologies to concrete infrastructures (technical conditions), use cases (for which users are expected to use the data) and needs of the particular discipline, e.g. representation of specific

[4] http://thedatahub.io/

[5] http://cran.r-project.org/web/packages/SPARQL/index.html

information types, data access and privacy issues. Only when these aspects are brought together can an application of LOD be realized.

6.2 Future Work

Our contributions towards publishing Linked Open Social Science Data and matching Statistical Linked Data encourage efforts in further research. In this section, we present an outlook on future research possibilities geared towards enabling a meaningful and comprehensive consumption of Linked Open Social Science Data.

While the standards and technologies are available and sufficiently mature for publishing Linked Open Social Science Data, there are still obstacles that hinder major publication efforts of research data by data providers. One of the obstacles is certainly the openness of major data holdings and the specific privacy restrictions that have to be followed. Social Science Data is mostly very sensitive, e.g. in survey data that contains opinions from individual people or personal data like salary data, zip codes and others. This is a general challenge that affects not only the social sciences, but rather all domains where sensitive data is collected and used. It is also dependent on the infrastructure, based on which the data or parts of it are published and accessed. In [WBZMZ13], we discuss the promise of semantic technologies for this aim and current challenges with a focus on architecture, retrieval and privacy. For the latter aspect, the manual and time-consuming process of data anonymization can be supported by Semantic Web technologies. This could be done by detecting (semi-)automatically the combinations of attributes in the data that violate the privacy.

Data cannot always be accessed to the fullest extent on the web. However, the reproducibility of research results is an area that has gained in importance. In [BZT13], initial works have begun on a data restore model that enables the reproducibility of research experiments. By referencing particular data values that have been used in a scientific experiment, these values can be included into one's own program code without necessarily requiring the entire data set that may be access-restricted. Since a data template is used to point to the original data values, this model can theoretically be applied on Linked Data that underlies privacy restrictions.

A big step in the direction of consuming and interacting with Linked Data is to consider applications that allow user interaction with such data. In [ZM11b], we present an approach for performing statistical calculations on Linked Data. While current approaches focus on the use of Linked Data in statistical tools [vHEM12, KOH12], our approach follows a more visionary goal and allows for performing statistical calculations directly on the web without any dependencies on particular tools or the need for a background knowledge of statistics. Thus, the approach relies only on Semantic Web standards and technologies like SPARQL query processing. By combining distributed sources with SPARQL, we are able to apply simple statistical calculations, such as linear regression and present the results to the user. The results of testing these calculations using heterogeneous data sources reveals a wide range of typical issues on data integration that one has to be aware of when working with heterogeneous statistical data. With this approach, we demonstrate that it is technically feasible to execute mathematical calculations directly on Linked Data. Such an approach can be independent of existing statistical tools, when relying only on standards (data modelling and vocabularies) and technologies (SPARQL query processing) from the Semantic Web. Issues relating to including the user e.g. for choosing desired observation pairs remain challenging.

Bibliography

The correctness and availability of the URLs in this thesis have been last reviewed on 01 September 2014.

[ADL$^+$09] Sören Auer, Sebastian Dietzold, Jens Lehmann, Sebastian Hellmann, and David Aumueller. Triplify: light-weight linked data publication from relational databases. In Juan Quemada, Gonzalo León, Yoëlle S. Maarek, and Wolfgang Nejdl, editors, *Proceedings of the 18th International Conference on World Wide Web (WWW), Madrid, Spain*, pages 621–630, April 2009.

[ADMR05] David Aumueller, Hong Hai Do, Sabine Massmann, and Erhard Rahm. Schema and ontology matching with COMA++. In Fatma Özcan, editor, *Proceedings of the ACM SIGMOD International Conference on Management of Data, Baltimore, Maryland, USA, June 14-16, 2005*, pages 906–908, 2005.

[AEE$^+$12] José Luis Aguirre, Kai Eckert, Jérôme Euzenat, Willem Robert van Ferrara, Alfio amd Hage, Laura Hollink, Christian Meilicke, Andriy Nikolov, Dominique Ritze, François Scharffe, Pavel Shvaiko, Ondrej Sváb-Zamazal, Cássia Trojahn, Ernesto Jiménez-Ruiz, Bernardo Cuenca Grau, and Benjamin Zapilko. Results of the ontology alignment evaluation initiative 2012. In Pavel Shvaiko, Jérôme Euzenat, Anastasios Kementsietsidis, Ming Mao, Natasha Noy, and Heiner Stuckenschmidt, editors, *Proceedings of the 7th International Workshop on Ontology Matching, Boston, MA, USA, November 11, 2012*, volume 946 of *CEUR Workshop Proceedings*, 2012.

[AHSB13] Ben Adida, Ivan Herman, Manu Sporny, and Mark Birbeck. RDFa 1.1 primer. http://www.w3.org/TR/rdfa-primer/, 2013. Rich Structured Data Markup for Web Documents. W3C Working Group Note 22 August 2013.

[AHSdV11] Samur Araújo, Jan Hidders, Daniel Schwabe, and Arjen P. de Vries. SERIMI - resource description similarity, RDF instance matching and interlinking. In Pavel Shvaiko, Jérôme Euzenat, Tom Heath, Christoph Quix, Ming Mao, and Isabel F. Cruz, editors, *Proceedings of the 6th International Workshop on Ontology Matching, Bonn, Germany, October 24, 2011*, volume 814 of *CEUR Workshop Proceedings*, 2011.

[AKT01] AKT. AKT reference ontology. http://www.aktors.org/publications/ontology/, 2001.

[All] The Data Documentation Initiative Alliance. The Data Documentation Initiative (DDI). http://www.ddialliance.org/.

[All09] The Data Documentation Initiative Alliance. DDI technical specification, part ii, 2009.

[ALL10] ALLBUS/GGSS 1980-2008 (Kumulierte Allgemeine Bevölkerungsumfrage der Sozialwissenschaften/Cumulated German General Social Survey 1980-2008). doi:10.4232/1.10080, 2010.

[AR10] Janneke Adema and Paul Rutten. Digital monographs in the humanities and social sciences: Report on user needs. Technical report, 2010. OAPEN Project Report.

[AS13] Phil Archer and Gofran Shukair. Asset Description Metadata Schema (ADMS). http://www.w3.org/TR/vocab-adms/, 2013.

[ATV08] Bogdan Alexe, Wang-Chiew Tan, and Yannis Velegrakis. STBenchmark: towards a benchmark for mapping systems. *Proc. VLDB Endow.*, 1(1):230–244, August 2008.

[BAB$^+$10] Sean Bechhofer, John Ainsworth, Jitenkumar Bhagat, Iain Buchan, Phillip Couch, Don Cruickshank, Mark Delderfield, Ian Dunlop, Matthew Gamble, Carole Goble, Danius Michaelides, Paolo Missier, Stuart Owen, David Newman, David De Roure, and Shoaib Sufi. Why linked data is not enough for scientists. In *Sixth International Conference on e-Science, e-Science 2010, 7-10 December 2010, Brisbane, QLD, Australia*. IEEE Computer Society, 2010.

[BB10] Uldis Bojārs and John G. Breslin. SIOC Core Ontology Specification. http://rdfs.org/sioc/spec/, 2010.

[BBDV11] Zohra Bellahsene, Angela Bonifati, Fabien Duchateau, and Yannis Velegrakis. On evaluating schema matching and mapping. In Zohra Bellahsene, Angela Bonifati, and Erhard Rahm, editors, *Schema Matching and Mapping*, pages 253–291. Springer, 2011.

[BBR11] Zohra Bellahsene, Angela Bonifati, and Erhard Rahm, editors. *Schema Matching and Mapping*. Springer, 2011.

[BC06] Christian Bizer and Richard Cyganiak. D2R Server - publishing relational databases on the semantic web. In *Poster at the 5th International Semantic Web Conference, Athens, GA, USA*, November 5-9 2006.

[BCH07] Christian Bizer, Richard Cyganiak, and Tom Heath. How to publish linked data on the web. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/, 2007.

[BCWZ12] Thomas Bosch, Richard Cyganiak, Joachim Wackerow, and Benjamin Zapilko. Leveraging the DDI model for linked statistical data in the social, behavioural, and economic sciences. In *Proceedings of the 2012*

International Conference on Dublin Core and Metadata Applications. Dublin Core Metadata Initiative, 2012.

[BFF08] Diego Berrueta, Sergio Fernández, and Iván Frade. Cooking HTTP content negotiation with Vapour. In *Proceedings of 4th workshop on Scripting for the Semantic Web 2008 (SFSW2008), Tenerife, Canary Islands, Spain*, 2008.

[BG04] Dan Brickley and R.V. Guha. RDF Vocabulary Description Language 1.0: RDF Schema. http://www.w3.org/TR/rdf-schema/, 2004. W3C Recommendation 10 February 2004.

[BHBL09] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, 2009.

[BKK11] Alkyoni Baglatzi, Tomi Kauppinen, and Carsten Keßler. Linked science core vocabulary specification. http://linkedscience.org/lsc/ns/, 2011.

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

[BLCC+06] Tim Berners-Lee, Yuhsin Chen, Lydia Chilton, Dan Connolly, Ruth Dhanaraj, James Hollenbach, Adam Lerer, and David Sheets. Tabulator: Exploring and analyzing linked data on the semantic web. In *Proceedings of the 3rd International Semantic Web User Interaction Workshop*, 2006.

[BLHL01] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific American*, 284(5):34–43, May 2001.

[BLK+09] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - a crystallization point for the web of data. *J. Web Sem.*, 7(3):154–165, 2009.

[BM02] Jacob Berlin and Amihai Motro. Database schema matching using machine learning with feature selection. In Anne Banks Pidduck, John Mylopoulos, Carson C. Woo, and M. Tamer Özsu, editors, *Advanced Information Systems Engineering, 14th International Conference, CAiSE 2002, Toronto, Canada, May 27-31, 2002, Proceedings*, volume 2348 of *Lecture Notes in Computer Science*, pages 452–466. Springer, 2002.

[BM11] Thomas Bosch and Brigitte Mathiak. Generic multilevel approach designing domain ontologies based on XML schemas. In *Proceedings of the Workshop Ontologies Come of Age in the Semantic Web*, CEUR Workshop Proceedings, pages 1–12, 2011.

[BM12] Thomas Bosch and Brigitte Mathiak. XSLT transformation generating owl ontologies automatically based on xml schemas. In *International Conference for Internet Technology and Secured Transactions (ICITST)*, 2012.

[BMR11] Philip A. Bernstein, Jayant Madhavan, and Erhard Rahm. Generic schema matching, ten years later. In *PVLDB, 2011 (VLDB 10 Year Best Paper Award Paper)*, 2011.

[BN05] Alexander Bilke and Felix Naumann. Schema matching using duplicates. In Karl Aberer, Michael J. Franklin, and Shojiro Nishio, editors, *Proceedings of the 21st International Conference on Data Engineering, ICDE 2005, 5-8 April 2005, Tokyo, Japan*, pages 69–80. IEEE Computer Society, 2005.

[Bor93] Jürgen Bortz. *Lehrbuch der Statistik für Sozialwissenschaftler*. Springer, 1993.

[Bri10] Dan Brickley. FOAF Vocabulary Specification 0.98. http://xmlns.com/foaf/spec/, 2010. Namespace Document 9 August 2010.

[BZT13] Daniel Bahls, Benjamin Zapilko, and Klaus Tochermann. A data restore model for reproducibility in computational statistics. In *i-Know '13: Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies*, Graz, Austria, 2013.

[BZTM11] Arnim Bleier, Benjamin Zapilko, Mark Thamm, and Peter Mutschke. Using SKOS to integrate social networking sites with scholarly information portals. In Alexandre Passant, Sergio Fernandez, John Bresli, and Uldis Bojārs, editors, *SDoW2011 - Social Data on the Web: workshop at the 10th International Semantic Web Conference, Bonn*, October 23 2011.

[BZWG13] Thomas Bosch, Benjamin Zapilko, Joachim Wackerow, and Arofan Gregory. Towards the discovery of person-level data: reuse of vocabularies and related use cases. In *Proceedings of the International Workshop on Semantic Statistics (SemStats 2013) collocated with the 12th International Semantic Web Conference (ISWC-2013)*, 2013.

[CAS09] Isabel F. Cruz, Flavio Palandri Antonelli, and Cosmin Stroe. AgreementMaker: Efficient matching for large real-world schemas and ontologies. *Proc. VLDB Endow.*, 2(2):1586–1589, August 2009.

[CB07] Richard Cyganiak and Christian Bizer. Pubby: A linked data frontend for SPARQL endpoints. http://www4.wiwiss.fu-berlin.de/pubby/, 2007.

[CBG$^+$11] Marcus Cobden, Jennifer Black, Nicholas Gibbins, Les Carr, and Nigel R. Shadbolt. A research agenda for linked closed dataset. In Olaf Hartig, Andreas Harth, and Juan Sequeda, editors, *Proceedings of the Second International Workshop on Consuming Linked Data (COLD2011), Bonn, Germany, October 23, 2011*, volume 782 of *CEUR Workshop Proceedings*, 2011.

[CCGC09] Karen Calhoun, Joanne Cantrell, Peggy Gallagher, and Diane Cellentani. Online catalogs what users and librarians want: an OCLC report. Technical report, Dublin, Ohio, 2009.

[CDR10] Richard Cyganiak, Chris Dollin, and Dave Reynolds. Expressing statistical data in RDF with SDMX-RDF. http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/index.html, 2010.

[CFG$^+$10] Richard Cyganiak, Simon Field, Arofan Gregory, Wolfgang Halb, and Jeni Tennison. Semantic statistics: Bringing together SDMX and SCOVO. In Christian Bizer, Tom Heath, Tim Berners-Lee, and Michael Hausenblas, editors, *Proceedings of the WWW2010 Workshop on Linked Data on the Web, LDOW 2010, Raleigh, USA, April 27, 2010*, volume 628 of *CEUR Workshop Proceedings*, 2010.

[CFM06] Silvana Castano, Alfio Ferrara, and Stefano Montanelli. Matching ontologies in open networked systems: Techniques and applications. In Stefano Spaccapietra, Paolo Atzeni, WesleyW. Chu, Tiziana Catarci, and KatiaP. Sycara, editors, *J. Data Semantics V*, volume 3870 of *Lecture Notes in Computer Science*, pages 25–63. Springer Berlin Heidelberg, 2006.

[Cot14] Franck Cotton. XKOS - An SKOS extension for representing statistical classifications. http://rdf-vocabulary.ddialliance.org/xkos.html, 2014.

[CQ09] Gong Cheng and Yuzhong Qu. Searching linked objects with falcons: Approach, implementation and evaluation. *Int. J. Semantic Web Inf. Syst.*, 5(3):49–70, 2009.

[CRT14] Richard Cyganiak, Dave Reynolds, and Jeni Tennison. The RDF Data Cube vocabulary. http://www.w3.org/TR/vocab-data-cube/, 2014.

[CS11] Thierry Coquand and Vincent Siles. A decision procedure for regular expression equivalence in type theory. In *Proceedings of the First international conference on Certified Programs and Proofs*, CPP'11, pages 119–134, Berlin, Heidelberg, 2011. Springer-Verlag.

[CZAH11] Richard Cyganiak, Jun Zhao, Keith Alexander, and Michael Hausenblas. Vocabulary of Interlinked Datasets (VoID). http://vocab.deri.ie/void/, 2011.

[DD12] Leigh Dodds and Ian Davis. Linked data patterns. http://patterns.dataincubator.org/book/, 2012.

[DDH01] AnHai Doan, Pedro Domingos, and Alon Y. Halevy. Reconciling schemas of disparate data sources: A machine-learning approach. In Walid G. Aref, editor, *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data, Santa Barbara, California, USA*, pages 509–520, 2001.

[De 09] Johan De Smedt. SKOS extensions for the EUROVOC thesaurus. In *European Semantic Technology Conference*, 2009.

[DEvZ10] Jérôme David, Jérôme Euzenat, and Ondřej Šváb Zamazal. Ontology similarity in the alignment space. In Peter F. Patel-Schneider, Yue Pan, Pascal Hitzler, Peter Mika, Lei Zhang, Jeff Z. Pan, Ian Horrocks, and Birte Glimm, editors, *The Semantic Web - ISWC 2010 - 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I*, volume 6496 of *Lecture Notes in Computer Science*, pages 129–144. Springer, 2010.

[DG08] Bruce D'Arcus and Frédéric Giasson. The Bibliographic Ontology (BIBO). http://bibliontology.com/, 2008.

[DMD+03] AnHai Doan, Jayant Madhavan, Robin Dhamankar, Pedro Domingos, and Alon Halevy. Learning to match ontologies on the semantic web. *The VLDB Journal*, 12(4):303–319, November 2003.

[DMDH03] Anhai Doan, Jayant Madhavan, Pedro Domingos, and Alon Halevy. Ontology matching: A machine learning approach. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies in Information Systems*, International Handbooks on Information Systems, pages 397–416. Springer, 2003.

[DMQ05] Dejing Dou, Drew V. McDermott, and Peishen Qi. Ontology translation on the semantic web. *J. Data Semantics*, 2:35–57, 2005.

[DMR03] Hong Hai Do, Sergey Melnik, and Erhard Rahm. Comparison of schema matching evaluations. In *Revised Papers from the NODe 2002 Web and Database-Related Workshops on Web, Web-Services, and Database Systems*, pages 221–237, London, UK, 2003. Springer-Verlag.

[DP11] Thorsten Dresing and Thorsten Pehl. *Praxisbuch Transkription. Regelsysteme, Software und praktische Anleitungen für qualitative ForscherInnen.* 2011. ISBN 978-3-8185-0489-2.

[DR02] Hong Hai Do and Erhard Rahm. COMA - a system for flexible combination of schema matching approaches. In *VLDB 2002, Proceedings of 28th International Conference on Very Large Data Bases, August 20-23, 2002, Hong Kong, China*, pages 610–621. Morgan Kaufmann, 2002.

[DSC12] Souripriya Das, Seema Sundara, and Richard Cyganiak. R2RML: RDB to RDF Mapping Language. http://www.w3.org/TR/r2rml/, 2012. W3C Proposed Recommendation 14 August 2012.

[Dum] Edd Dumbill. Description of a project. https://github.com/edumbill/doap/wiki.

[Dun46] Halbert L. Dunn. Record linkage. *American Journal of Public Health and the Nations Health*, 36(12):1412–1416, 1946.

[EFM+10] Jérôme Euzenat, Alfio Ferrara, Christian Meilicke, Juan Pane, François Scharffe, Pavel Shvaiko, Heiner Stuckenschmidt, Ondrej Sváb-Zamazal, Vojtech Svátek, and Cássia Trojahn dos Santos. Results of the ontology alignment evaluation initiative 2010. In Pavel Shvaiko, Jérôme Euzenat, Fausto Giunchiglia, Heiner Stuckenschmidt, Ming Mao, and Isabel F. Cruz, editors, *Proceedings of the 5th International Workshop on Ontology Matching (OM-2010), Shanghai, China, November 7, 2010*, volume 689 of *CEUR Workshop Proceedings*, 2010.

[EFvH^{+}11] Jérôme Euzenat, Alfio Ferrara, Willem Robert van Hage, Laura Hollink, Christian Meilicke, Andriy Nikolov, Dominique Ritze, François Scharffe, Pavel Shvaiko, Heiner Stuckenschmidt, Ondrej Sváb-Zamazal, and Cássia Trojahn dos Santos. Results of the ontology alignment evaluation initiative 2011. In Pavel Shvaiko, Jérôme Euzenat, Tom Heath, Christoph Quix, Ming Mao, and Isabel F. Cruz, editors, *Proceedings of the 6th International Workshop on Ontology Matching, Bonn, Germany, October 24, 2011*, volume 814 of *CEUR Workshop Proceedings*, 2011.

[EHHS05] Marc Ehrig, Peter Haase, Mark Hefke, and Nenad Stojanovic. Similarity for ontologies - a comprehensive framework. In Dieter Bartmann, Federico Rajola, Jannis Kallinikos, David E. Avison, Robert Winter, Phillip Ein-Dor, Jörg Becker, Freimut Bodendorf, and Christof Weinhardt, editors, *Proceedings of the 13th European Conference on Information Systems, Information Systems in a Rapidly Changing Economy, ECIS 2005, Regensburg, Germany, May 26-28, 2005*, pages 1509–1518, 2005.

[EIV07] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. Duplicate record detection: A survey. *IEEE Trans. Knowl. Data Eng.*, 19(1):1–16, 2007.

[EM07] Daniel Engmann and Sabine Maßmann. Instance matching with COMA++. In Matthias Jarke, Thomas Seidl, Christoph Quix, David Kensche, Stefan Conrad, Erhard Rahm, Ralf Klamma, Harald Kosch, Michael Granitzer, Sven Apel, Marko Rosenmüller, Gunter Saake, and Olaf Spinczyk, editors, *Datenbanksysteme in Business, Technologie und Web (BTW 2007), Workshop Proceedings, 5.-6. März 2007, Aachen, Germany*, pages 28–37. Verlagshaus Mainz, Aachen, 2007.

[EMS09] Kai Eckert, Christian Meilicke, and Heiner Stuckenschmidt. Improving ontology matching using meta-level learning. In Lora Aroyo, Paolo Traverso, Fabio Ciravegna, Philipp Cimiano, Tom Heath, Eero Hyvönen, Riichiro Mizoguchi, Eyal Oren, Marta Sabou, and Elena Paslaru Bontas Simperl, editors, *The Semantic Web: Research and Applications, 6th European Semantic Web Conference, ESWC 2009, Heraklion, Crete, Greece, May 31-June 4, 2009, Proceedings*, volume 5554 of *Lecture Notes in Computer Science*, pages 158–172. Springer, 2009.

[ES04a] Marc Ehrig and Steffen Staab. QOM - quick ontology mapping. In Sheila A. McIlraith, Dimitris Plexousakis, and Frank van Harmelen, editors, *The Semantic Web - ISWC 2004: Third International Semantic Web Conference,Hiroshima, Japan, November 7-11, 2004. Proceedings*, volume 3298 of *Lecture Notes in Computer Science*, pages 683–697. Springer, 2004.

[ES04b] Marc Ehrig and York Sure. Ontology mapping - an integrated approach. In Christoph Bussler, John Davies, Dieter Fensel, and Rudi Studer, editors, *The Semantic Web: Research and Applications, First European Semantic Web Symposium, ESWS 2004, Heraklion, Crete, Greece, May 10-12, 2004, Proceedings*, volume 3053 of *Lecture Notes in Computer Science*, pages 76–91. Springer, 2004.

[ES07] Jérôme Euzenat and Pavel Shvaiko. *Ontology matching.* Springer-Verlag, Heidelberg (DE), 2007.

[EV04] Jérôme Euzenat and Petko Valtchev. Similarity-based ontology alignment in OWL-Lite. In Ramon López de Mántaras and Lorenza Saitta, editors, *Proceedings of the 16th Eureopean Conference on Artificial Intelligence, ECAI'2004, including Prestigious Applicants of Intelligent Systems, PAIS 2004, Valencia, Spain, August 22-27, 2004*, pages 333–337. IOS Press, 2004.

[FCOO12] Andre Freitas, Edward Curry, Joao Gabriel Oliveira, and Sean O'Riain. Querying heterogeneous datasets on the linked data web: Challenges, approaches, and trends. *IEEE Internet Computing*, 16(1):24–33, January 2012.

[fECoD02] Organisation for Economic Co-operation and Development, editors. *Measuring the non-observed economy: a handbook.* Statistics. Paris, 2002.

[FH11] Christian Fürber and Martin Hepp. Towards a vocabulary for data quality management in semantic web architectures. In *Proceedings of the 1st International Workshop on Linked Web Data Management*, LWDM '11, pages 1–8, New York, NY, USA, 2011. ACM.

[Fie99] Roy Fielding. Hypertext transfer protocol – HTTP/1.1. http://www.w3.org/Protocols/rfc2616/rfc2616.html, 1999.

[FLMV08] Alfio Ferrara, Davide Lorusso, Stefano Montanelli, and Gaia Varese. Towards a benchmark for instance matching. In Pavel Shvaiko, Jérôme Euzenat, Fausto Giunchiglia, and Heiner Stuckenschmidt, editors, *Proceedings of the 3rd International Workshop on Ontology Matching (OM-2008) Collocated with the 7th International Semantic Web Conference (ISWC-2008), Karlsruhe, Germany, October 26, 2008*, volume 431 of *CEUR Workshop Proceedings.* CEUR-WS.org, 2008.

[FvKK$^+$95] Uwe Flick, Ernst v. Kardorff, Heiner Keupp, Lutz v. Rosenstiel, and Stephan Wolff. *Handbuch Qualitative Sozialforschung: Grundlagen, Konzepte, Methoden und Anwendungen.* Beltz, Psychologie-Verlag-Union, 1995.

[GCAR06] Carole A. Goble, Óscar Corcho, Pinar Alper, and David De Roure. e-science and the semantic web: A symbiotic relationship. In *Discovery Science*, pages 1–12, 2006.

[GGSL12] Christophe Guéret, Paul Groth, Claus Stadler, and Jens Lehmann. Assessing linked data mappings using network measures. In Elena Simperl, Philipp Cimiano, Axel Polleres, Óscar Corcho, and Valentina Presutti, editors, *The Semantic Web: Research and Applications - 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings*, volume 7295 of *Lecture Notes in Computer Science*, pages 87–102. Springer, 2012.

[GHHZ11] Thomas Gottron, Christian Hachenberg, Andreas Harth, and Benjamin Zapilko. Towards a semantic data library for the social sciences. In Livia Predoiu, Steffen Hennicke, Andreas Nürnberger, Annett Mitschick, and Seamus Ross, editors, *SDA 2011 - Semantic Digital Archives: Proceedings of the International Workshop on Semantic Digital Archives, Berlin, Germany, September 29, 2011*, volume 801 of *CEUR Workshop Proceedings*, 2011.

[Gol07] Anna Gold. Cyberinfrastructure, data and libraries. part 1 & 2. *D-Lib Magazine*, 13(9/10), 2007.

[GSY04] Fausto Giunchiglia, Pavel Shvaiko, and Mikalai Yatskevich. S-match: an algorithm and an implementation of semantic matching. In Christoph Bussler, John Davies, Dieter Fensel, and Rudi Studer, editors, *The Semantic Web: Research and Applications, First European Semantic Web Symposium, ESWS 2004, Heraklion, Crete, Greece, May 10-12, 2004, Proceedings*, volume 3053 of *Lecture Notes in Computer Science*, pages 61–75. Springer, 2004.

[GV10] Arofan Gregory and Mary Vardigan. The web of linked data: Realizing the potential for the social sciences. Nsf sbe 2020 paper 186, 2010.

[Hal05] Alon Halevy. Why your data won't mix. *Queue*, 3(8):50–58, October 2005.

[Har12] Andreas Harth. VisiNav: A system for visual search and navigation on web data. *J. Web Sem.*, 8(4), 2012.

[Hau09] Michael Hausenblas. Exploiting linked data to build web applications. *IEEE Internet Computing*, 13(4):68–73, July 2009.

[HB11] Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool, 2011.

[HBF09] Olaf Hartig, Christian Bizer, and Johann Christoph Freytag. Executing SPARQL queries over the web of linked data. In Abraham Bernstein, David R. Karger, Tom Heath, Lee Feigenbaum, Diana Maynard, Enrico Motta, and Krishnaprasad Thirunarayan, editors, *The Semantic Web - ISWC 2009, 8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, October 25-29, 2009. Proceedings*, volume 5823 of *Lecture Notes in Computer Science*, pages 293–309. Springer, 2009.

[HCCQ10] Wei Hu, Jianfeng Chen, Gong Cheng, and Yuzhong Qu. ObjectCoref & Falcon-AO: results for oaei 2010. In Pavel Shvaiko, Jérôme Euzenat, Fausto Giunchiglia, Heiner Stuckenschmidt, Ming Mao, and Isabel F. Cruz, editors, *Proceedings of the 5th International Workshop on Ontology Matching (OM-2010), Shanghai, China, November 7, 2010*, volume 689 of *CEUR Workshop Proceedings*, 2010.

[HE09] Alison J. Head and Michael B. Eisenberg. Lessons learned: How college students seek information in the digital age. Technical report, Information School, University of Washington, December 2009.

[Hel05] Cornelia Helfferich. *Die Qualität qualitativer Daten.* VS, Verlag für Sozialwiss., 2005.

[Hep08] Martin Hepp. GoodRelations: An ontology for describing products and services offers on the web. In Aldo Gangemi and Jérôme Euzenat, editors, *Knowledge Engineering: Practice and Patterns, 16th International Conference, EKAW 2008, Acitrezza, Italy, September 29 - October 2, 2008. Proceedings*, volume 5268 of *Lecture Notes in Computer Science*, pages 332–347. Springer, 2008.

[HFH+09] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.

[HHD+07] Andreas Harth, Aidan Hogan, Renaud Delbru, Jürgen Umbrich, Seán O'Riain, and Stefan Decker. SWSE: Answers before links! In Jennifer Golbeck and Peter Mika, editors, *Proceedings of the Semantic Web Challenge 2007 co-located with ISWC 2007 + ASWC 2007, Busan, Korea, November 13th, 2007*, volume 295 of *CEUR Workshop Proceedings*, 2007.

[HHR+09] Michael Hausenblas, Wolfgang Halb, Yves Raimond, Lee Feigenbaum, and Danny Ayers. SCOVO: Using statistics on the web of data. In Lora Aroyo, Paolo Traverso, Fabio Ciravegna, Philipp Cimiano, Tom Heath, Eero Hyvönen, Riichiro Mizoguchi, Eyal Oren, Marta Sabou, and Elena Paslaru Bontas Simperl, editors, *The Semantic Web: Research and Applications, 6th European Semantic Web Conference, ESWC 2009, Heraklion, Crete, Greece, May 31-June 4, 2009, Proceedings*, volume 5554 of *Lecture Notes in Computer Science*, pages 708–722. Springer, 2009.

[HISW10] Harry Halpin, Renato Iannella, Brian Suda, and Norman Walsh. Representing vcard objects in RDF. http://www.w3.org/Submission/vcard-rdf, 2010. W3C Member Submission 20 January 2010.

[HK10] Jan Hannemann and Jürgen Kett. Linked data for libraries. In *World Library and Information Congress: 76th IFLA General Conference and Assembly 10-15 August 2010, Gothenburg, Sweden*, 2010.

[HL10] Olaf Hartig and Andreas Langegger. A database perspective on consuming linked data on the web. *Datenbankspektrum, Semantic Web Special Issue*, July 2010.

[HP09] James Hollenbach and Joe Presbrey. Using RDF metadata to enable access control on the social semantic web. In *Proceedings of the Workshop on Collaborative Construction, Management and Linking of Structured Knowledge, CK'09*, 2009.

[HPSB+04] Ian Horrocks, Peter F. Patel-Schneider, Harold Boley, Said Tabet, Benjamin Grosof, and Mike Dean. SWRL: A semantic web rule language combining OWL and RuleML. http://www.w3.org/Submission/SWRL/, 2004. W3C Member Submission 21 May 2004.

[HPV+02] Mauricio A. Hernández, Lucian Popa, Yannis Velegrakis, Renée J. Miller, Felix Naumann, and Ching-Tien Ho. Mapping XML and relational schemas with clio. In Rakesh Agrawal and Klaus R. Dittrich, editors, *Proceedings of the 18th International Conference on Data Engineering, San Jose, CA, USA, February 26 - March 1, 2002*, pages 498–499. IEEE Computer Society, 2002.

[HQ08] Wei Hu and Yuzhong Qu. Falcon-AO: A practical ontology matching system. *J. Web Sem.*, 6(3):237–239, September 2008.

[HSNM11] Jakob Huber, Timo Sztyler, Jan Nößner, and Christian Meilicke. CODI: Combinatorial optimization for data integration: results for oaei 2011. In Pavel Shvaiko, Jérôme Euzenat, Tom Heath, Christoph Quix, Ming Mao, and Isabel F. Cruz, editors, *Proceedings of the 6th International Workshop on Ontology Matching, Bonn, Germany, October 24, 2011*, volume 814 of *CEUR Workshop Proceedings*, 2011.

[HT12] Daniel M. Herzig and Thanh Tran. Heterogeneous web data search using relevance-based on the fly data integration. In Alain Mille, Fabien L. Gandon, Jacques Misselis, Michael Rabinovich, and Steffen Staab, editors, *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012*, WWW '12, pages 141–150, New York, NY, USA, 2012. ACM.

[HZ09] Olaf Hartig and Jun Zhao. Using web data provenance for quality assessment. In Juliana Freire, Paolo Missier, and Satya Sanket Sahoo, editors, *Proceedings of the First International Workshop on the role of Semantic Web in Provenance Management (SWPM 2009), collocated with the 8th International Semantic Web Conference (ISWC-2009), Washington DC, USA, October 25, 2009*, volume 526 of *CEUR Workshop Proceedings*, 2009.

[HZ12] Olaf Hartig and Jun Zhao. Provenance Vocabulary Core Ontology Specification. http://trdf.sourceforge.net/provenance/ns.html, 2012.

[HZSM11a] Daniel Hienert, Benjamin Zapilko, Philipp Schaer, and Brigitte Mathiak. Vizgr - combining data on a visual level. In JosÃ© Cordeiro and Joaquim Filipe, editors, *WEBIST 2011, Proceedings of the 7th International Conference on Web Information Systems and Technologies, Noordwijker-*

hout, Netherlands, 6-9 May, 2011, pages 202–211. SciTePress, 2011. ISBN: 978-989-8425-51-5.

[HZSM11b] Daniel Hienert, Benjamin Zapilko, Philipp Schaer, and Brigitte Mathiak. Web-based multi-view visualizations for aggregated statistics. In *2nd International Workshop on Data Visualization and Integration on the Web (DATAVIEW); Proceedings of the 5th International Workshop on Web APIs and Services Mashups Proceedings (Mashups '11)*. ACM, 2011. ISBN: 978-1-4503-0823-6.

[HZSM12] Daniel Hienert, Benjamin Zapilko, Philipp Schaer, and Brigitte Mathiak. Vizgr: linking data in visualizations. In Joaquim Filipe and José Cordeiro, editors, *Web information systems and technologies : 7th international conference ; revised selected papers*, number 101 in Lecture notes in business information processing, pages 177–191. Springer, 2012. ISBN: 978-3-642-28081-8.

[IMSW07] Antoine Isaac, Lourens Van Der Meij, Stefan Schlobach, and Shenghui Wang. An empirical study of instance-based ontology matching. In Karl Aberer, Key-Sun Choi, Natasha Fridman Noy, Dean Allemang, Kyung-Il Lee, Lyndon J. B. Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, volume 4825 of *Lecture Notes in Computer Science*, pages 253–266. Springer, 2007.

[IMvdM$^+$08] Antoine Isaac, Henk Matthezing, Lourens van der Meij, Stefan Schlobach, Shenghui Wang, and Claus Zinn. Putting ontology alignment in context: Usage scenarios, deployment and evaluation in a library case. In Sean Bechhofer, Manfred Hauswirth, Jörg Hoffmann, and Manolis Koubarakis, editors, *The Semantic Web: Research and Applications, 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008, Proceedings*, volume 5021 of *Lecture Notes in Computer Science*, pages 402–417. Springer, 2008.

[Ini] Dublin Core Metatdata Initiative. DCMI Metadata Terms. http://dublincore.org/.

[Inm05] William H. Inmon. *Building the Data Warehouse*. John Wiley & Sons, Inc., New York, NY, USA, 4rth edition, 2005.

[IS09a] Antoine Isaac and Ed Summers. Correspondences between ISO-2788/5964 and SKOS constructs, 2009.

[IS09b] Antoine Isaac and Ed Summers. SKOS Simple Knowledge Organization System Primer. http://www.w3.org/TR/skos-primer/, 2009. W3C Working Group Note 18 August 2009.

[ISO85] ISO. ISO 5964:1985: Documentation - guidelines for the establishment and development of multilingual thesauri, 1985.

[ISO86] ISO. ISO 2788:1986: Documentation - guidelines for the establishment and development of monolingual thesauri, 1986.

[ISO00] ISO. ISO 1087-1:2000: Terminology work - vocabulary - part 1: Theory and application, 2000.

[ISO03] ISO. ISO 19115:2003/Cor 1:2006: Geographic information - metadata - part 1: Fundamentals, 2003. will be revised by ISO/DIS 19115-1.

[ISO04] ISO. ISO/IEC 11179-1:2004: Information technology - metadata registries (mdr) - part 1: Framework, 2004.

[IUBH10] Robert Isele, Jürgen Umbrich, Chris Bizer, and Andreas Harth. LDSpider: An open-source crawling framework for the web of linked data. In Axel Polleres and Huajun Chen, editors, *Proceedings of the ISWC 2010 Posters & Demonstrations Track: Collected Abstracts, Shanghai, China, November 9, 2010*, volume 658 of *CEUR Workshop Proceedings*, 2010.

[JFNP12] Frederik Janssen, Faraz Fallahi, Jan Noessner, and Heiko Paulheim. Towards rule learning approaches to instance-based ontology matching. In *1st International Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data (Know@LOD)*, 2012.

[JH04] James A. Jacobs and Charles Humphrey. Preserving research data. *Commun. ACM*, 47(9):27–29, September 2004.

[JMSK09] Yves R. Jean-Mary, E. Patrick Shironoshita, and Mansur R. Kabuka. Ontology matching with semantic verification. *J. Web Sem.*, 7(3):235–251, 2009.

[JYV^{+}11] Prateek Jain, Peter Z. Yeh, Kunal Verma, Reymonrod G. Vasquez, Mariana Damova, Pascal Hitzler, and Amit P. Sheth. Contextual ontology alignment of LOD with an upper ontology: A case study with proton. In Grigoris Antoniou, Marko Grobelnik, Elena Paslaru Bontas Simperl, Bijan Parsia, Dimitris Plexousakis, Pieter De Leenheer, and Jeff Z. Pan, editors, *The Semantic Web: Research and Applications - 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29-June 2, 2011, Proceedings, Part I*, volume 6643 of *Lecture Notes in Computer Science*, pages 80–92. Springer, 2011.

[KBK12] Tomi Kauppinen, Alkyoni Baglatzi, and Carsten Keßler. Linked Science: Interconnecting Scientific Assets. In Terence Critchlow and Kerstin Kleese-Van Dam, editors, *Data Intensive Science*. CRC Press, USA, 2012.

[KC04] Graham Klyne and Jeremy J. Carroll. Resource description framework (RDF): Concepts and abstract syntax. http://www.w3.org/TR/rdf-concepts/, 2004. W3C Recommendation 10 February 2004.

[KdE11] Tomi Kauppinen and Giovana Mira de Espindola. Linked open science-communicating, sharing and evaluating data, methods and results for executable papers. *Procedia CS*, 4:726–731, 2011.

[KH13] Benedikt Kämpgen and Andreas Harth. No size fits all - running the star schema benchmark with SPARQL and RDF aggregate views. In Philipp Cimiano, Oscar Corcho, Valentina Presutti, Laura Hollink, and Sebastian Rudolph, editors, *The Semantic Web: Semantics and Big Data, 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings*, volume 7882 of *Lecture Notes in Computer Science*, pages 290–304. Springer Berlin Heidelberg, 2013.

[KII11] Kommission Zukunft der Informationsinfrastruktur KII. Gesamtkonzept für die Informationsinfrastruktur in Deutschland. Technical report, 2011.

[KKV94] Gary King, Robert O. Keohane, and Sidney Verba. *Designing Social Inquiry: Scientific Inference in Qualitative Research.* Princeton University Press, Princeton, 1994.

[KM09a] Sebastian Ryszard Kruk and Bill McDaniel. Goals of semantic digital libraries. In Sebastian Ryszard Kruk and Bill McDaniel, editors, *Semantic Digital Libraries*, pages 71–76. 2009.

[KM09b] Sebastian Ryszard Kruk and Bill McDaniel, editors. *Semantic Digital Libraries.* Springer, 2009.

[KNS03] Jürgen Krause, Elisabeth Niggemann, and Roland Schwänzl. Normierung und Standardisierung in sich verändernden Kontexten: Beispiel: Virtuelle Fachbibliotheken. *ZfBB: Zeitschrift für Bibliothekswesen und Bibliographie*, 50(1):19–28, 2003.

[KOH12] Benedikt Kämpgen, Sean O'Riain, and Andreas Harth. Interacting with statistical linked data via OLAP operations. In *Proceedings of the International Workshop on Interacting with Linked Data (ILD 2012), Extended Semantic Web Conference (ESWC).* CEUR-WS.org, Mai 2012.

[Kos00] Donald Kossmann. The state of the art in distributed query processing. *ACM Comput. Surv.*, 32(4):422–469, 2000.

[KPGC11] Michael A. Keller, Jerry Persons, Hugh Glaser, and Mimi Calter. Be part of the web - not just on it. report of the stanford linked data workshop. Technical report, 2011.

[Kra03a] Jürgen Krause. Standardisierung von der Heterogenität her denken: Zum Entwicklungsstand Bilateraler Transferkomponenten für digitale Fachbibliotheken. Technical Report 28, IZ-Sozialwissenschaften, Bonn, 2003.

[Kra03b] Jürgen Krause. Standardization, heterogeneity and the quality of content analysis: a key conflict of digital libraries and its solution. In *69th IFLA World Library and Information Congress*, 2003.

[Kra08] Jürgen Krause. Semantic heterogeneity: comparing new semantic web approaches with those of digital libraries. *Library Review*, 57(3):235–248, 2008.

[KREZ14] Andreas Oskar Kempf, Dominique Ritze, Kai Eckert, and Benjamin Zapilko. New ways of mapping knowledge organization systems. using a semi-automatic matching procedure for building up vocabulary crosswalks. *Knowledge Organization*, 41(1), 2014.

[Kru11] Jan Kruse. *Einführung in die Qualitative Interviewforschung*. 2011.

[KZ13] Andreas Oskar Kempf and Benjamin Zapilko. Normdatenpflege in Zeiten der Automatisierung. Zur Evaluation automatisch aufgebauter Thesaurus-Crosskonkordanzen. *Information - Wissenschaft & Praxis*, 64(4):199–208, 2013.

[Lan12] Christoph Lange. Ontologies and languages for representing mathematical knowledge on the semantic web. *Semantic Web*, 4(2):119–158, 2012.

[LC00] Wen-Syan Li and Chris Clifton. Semint: A tool for identifying attribute correspondences in heterogeneous databases using neural networks. *Data Knowl. Eng.*, 33(1):49–84, 2000.

[LCBF09] Luiz André P.Paes Leme, Marco A. Casanova, Karin K. Breitman, and Antonio L. Furtado. Instance-based OWL schema matching. In Joaquim Filipe and José Cordeiro, editors, *Enterprise Information Systems*, volume 24 of *Lecture Notes in Business Information Processing*, pages 14–26. Springer Berlin Heidelberg, 2009.

[LdS08] Carl Lagoze and Herbert Van de Sompel. ORE User Guide - Resource Map Implementation in RDF/XML. http://www.openarchives.org/ore/rdfxml, 2008.

[Lev66] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. Technical Report 8, 1966.

[Lit08] Beate Littig. Interviews mit eliten - interviews mit expertinnen: Gibt es unterschiede? *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, 9(3):17, 2008.

[LJC+08] Boris Lauser, Gudrun Johannsen, Caterina Caracciolo, Willem Robert van Hage, Johannes Keizer, and Philipp Mayr. Comparing human and automatic thesaurus mapping approaches in the agricultural domain. In *Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications*, pages 43–53. Dublin Core Metadata Initiative, 2008.

[LOV12] LOV. Linked Open Vocabularies (LOV). http://labs.mondeca.com/dataset/lov/, 2012.

[LPA+09] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo L. Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Social science. computational social science. *Science (New York, N.Y.)*, 323(5915):721–723, 2009.

[LSM13] Timothy Lebo, Satya Sahoo, and Deborah L. McGuinness. PROV-O: The PROV Ontology. http://www.w3.org/TR/prov-o/, 2013.

[LT10] Günter Ladwig and Thanh Tran. Linked data query processing strategies. In Peter F. Patel-Schneider, Yue Pan, Pascal Hitzler, Peter Mika, Lei Zhang, Jeff Z. Pan, Ian Horrocks, and Birte Glimm, editors, *The Semantic Web - ISWC 2010 - 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I*, volume 6496 of *Lecture Notes in Computer Science*, pages 453–469. Springer, 2010.

[LTLL09] Juanzi Li, Jie Tang, Yi Li, and Qiong Luo. RiMOM: A dynamic multistrategy ontology alignment framework. *IEEE Transactions on Knowledge and Data Engineering*, 21:1218–1232, 2009.

[LWB08] Andreas Langegger, Wolfram Wöß, and Martin Blöchl. A semantic web middleware for virtual data integration on the web. In Sean Bechhofer, Manfred Hauswirth, Jörg Hoffmann, and Manolis Koubarakis, editors, *The Semantic Web: Research and Applications, 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008, Proceedings*, volume 5021 of *Lecture Notes in Computer Science*, pages 493–507. Springer, 2008.

[Mal07] Vèronique Malaisè. SKOS: a model for metadata representation and interoperability dutch cultural heritage institution thesaurus conversion use case. Technical report, 2007. DELOS Multimatch workshop.

[Mal08] Martin Malmsten. Making a library catalogue part of the semantic web. In *Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications*. Dublin Core Metadata Initiative, 2008.

[May02] Philipp Mayring. *Einführung in die qualitative Sozialforschung.* Beltz, Weinheim, 2002.

[MB09a] A. Miles and S. Bechhofer. SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL). http://www.w3.org/TR/skos-reference/skos-xl.html, 2009.

[MB09b] A. Miles and S. Bechhofer. SKOS Simple Knowledge Organization System Reference. http://www.w3.org/TR/skos-reference/, 2009. W3C Recommendation 18 August 2009.

[MBD+07] Cecily Marcus, Stephanie Ball, Leslie Delserone, Amy Hribar, Wayne Loftus, Linda Watson, and Karen Williams. Understanding research behaviors, information resources, and service needs of scientists and graduate students: A study by the university of minnesota libraries. Technical report, June 2007.

[MBR01] Jayant Madhavan, Philip A. Bernstein, and Erhard Rahm. Generic schema matching with cupid. In Peter M. G. Apers, Paolo Atzeni, Stefano Ceri, Stefano Paraboschi, Kotagiri Ramamohanarao, and Richard T. Snodgrass, editors, *VLDB 2001, Proceedings of 27th International Conference on Very Large Data Bases, September 11-14, 2001, Roma, Italy*, pages 49–58. Morgan Kaufmann, 2001.

[MCA$^+$11] Stella Mitchell, Shanshan Chen, Mansoor Ahmed, Brian Lowe, Paula Markes, Nick Rejack, Jon Corson-Rikert, Bing He, and Ying Ding. The VIVO ontology: Enabling networking of scientists. In David De Roure and Marshall Scott Poole, editors, *Web Science 2011, WebSci '11, Koblenz, Germany - June 15 - 17, 2011*, 2011.

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

[MGMR02] Sergey Melnik, Hector Garcia-Molina, and Erhard Rahm. Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In Rakesh Agrawal and Klaus R. Dittrich, editors, *Proceedings of the 18th International Conference on Data Engineering, San Jose, CA, USA, February 26 - March 1, 2002*, pages 117–128. IEEE Computer Society, 2002.

[MMB12] Pablo N. Mendes, Hannes Mühleisen, and Christian Bizer. Sieve: linked data quality assessment and fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, EDBT-ICDT '12, pages 116–123, New York, NY, USA, 2012. ACM.

[Moc11] Ekkehard Mochmann. E-infrastructure for the social sciences. In *Building on Progress: Expanding the Research Infrastructure for the Social, Economic and Behavioral Sciences*. German Data Forum, 2011.

[MP08a] Philipp Mayr and Vivien Petras. Building a terminology network for search: the komohe project. In *Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications*, pages 177–182. Dublin Core Metadata Initiative, 2008.

[MP08b] Philipp Mayr and Vivien Petras. Cross-concordances: Terminology mapping and its effectiveness for information retrieval. In *74th IFLA World Library and Information Congress*, 2008.

[MS02] Alexander Maedche and Steffen Staab. Measuring similarity between ontologies. In Asunción Gómez-Pérez and V. Richard Benjamins, editors, *Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, 13th International Conference, EKAW 2002, Siguenza,* *Spain, October 1-4, 2002, Proceedings*, volume 2473 of *Lecture Notes in Computer Science*, pages 251–263. Springer, 2002.

[MvH04] Deborah L. McGuinness and Frank van Harmelen. Owl web ontology language. http://www.w3.org/TR/owl-features/, 2004. W3C Recommendation 10 February 2004.

[MZJ+11] Ahsan Morshed, Benjamin Zapilko, Gudrun Johannsen, Philipp Mayr, and Johannes Keizer. Evaluating approaches to automatically match thesauri from different domains for linked open data. In *10th European NKOS Workshop*, 2011.

[MZS10a] Philipp Mayr, Benjamin Zapilko, and York Sure. Ein Mehr-Thesauri-Szenario basierend auf SKOS und Crosskonkordanzen. In Marlies Ockenfeld, editor, *Recherche im Google-Zeitalter - vollständig und präzise?! : die Notwendigkeit von Informationskompetenz ; Tagungsband / 25. Oberhofer Kolloquium zur Praxis der Informationsvermittlung, Barleben/Magdeburg, 22. bis 24. April 2010*, number 13 in Deutsche Gesellschaft für Informationswissenschaft und Informationspraxis: Tagungen der Deutschen Gesellschaft für Informationswissenschaft und Informationspraxis, pages 163–172. DGI, 2010. ISBN: 978-3-925474-68-2.

[MZS10b] Philipp Mayr, Benjamin Zapilko, and York Sure. Establishing a multi-thesauri-scenario based on SKOS and cross-concordances. In *Proceedings of the 2010 International Conference on Dublin Core and Metadata Applications*. Dublin Core Metadata Initiative, 2010.

[NA11] Axel-Cyrille Ngonga Ngomo and Sören Auer. Limes a time-efficient approach for large-scale link discovery on the web of data. In Toby Walsh, editor, *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 2312–2317. IJCAI/AAAI, 2011.

[Neu09] Joachim Neubert. Bringing the thesaurus for economics on to the web of linked data. In Christian Bizer, Tom Heath, Tim Berners-Lee, and Kingsley Idehen, editors, *Proceedings of the WWW2009 Workshop on Linked Data on the Web, LDOW 2009, Madrid, Spain, April 20, 2009*, volume 538 of *CEUR Workshop Proceedings*, 2009.

[NIS11] NISO. Project ISO 25964: Thesauri and interoperability with other vocabularies. http://www.niso.org/schemas/iso25964/, 2011.

[NKIV11] Venkata Narasimha, Pavan Kappara, Ryutaro Ichise, and O. P. Vyas. LiDDM: A data mining system for linked data. In Christian Bizer, Tom Heath, Tim Berners-Lee, and Michael Hausenblas, editors, *WWW2011 Workshop on Linked Data on the Web, Hyderabad, India, March 29, 2011*, volume 813 of *CEUR Workshop Proceedings*, 2011.

[NLAH11] Axel-Cyrille Ngonga Ngomo, Jens Lehmann, Sören Auer, and Konrad Höffner. Raven - active learning of link specifications. In Pavel Shvaiko, Jérôme Euzenat, Tom Heath, Christoph Quix, Ming Mao, and Isabel F. Cruz, editors, *Proceedings of the 6th International Workshop on Ontology Matching, Bonn, Germany, October 24, 2011*, volume 814 of *CEUR Workshop Proceedings*, 2011.

[OAE] Ontology Alignment Evaluation Initiative OAEI. http://oaei.ontologymatching.org/.

[oC] The Library of Congress. PREMIS Preservation Metadata Maintenance Activity. http://www.loc.gov/standards/premis/.

[OOC09] P. ONeil, E. ONeil, and X. Chen. Star schema benchmark - revision 3. Technical report, UMass, Boston, 2009.

[Pau12] Heiko Paulheim. Generating possible interpretations for statistics from linked open data. In Elena Simperl, Philipp Cimiano, Axel Polleres, Óscar Corcho, and Valentina Presutti, editors, *The Semantic Web: Research and Applications - 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings*, volume 7295 of *Lecture Notes in Computer Science*, pages 560–574. Springer, 2012.

[PH10] Niko P. Popitsch and Bernhard Haslhofer. DSNotify: handling broken links in the web of data. In *Proceedings of the 19th World Wide Web Conference 2010, WWW 2010, Raleigh, North Carolina, USA, 2010*, WWW '10, pages 761–770, New York, NY, USA, 2010. ACM.

[PNMC$^+$13] Bernardo Pereira Nunes, Alexander Mera, MarcoAntonio Casanova, Besnik Fetahu, Luiz Andre P. Paes Leme, and Stefan Dietze. Complex matching of RDF datatype properties. In Hendrik Decker, Lenka Lhotska, Sebastian Link, Josef Basl, and Amin Tjoa, editors, *Database and Expert Systems Applications*, volume 8055 of *Lecture Notes in Computer Science*, pages 195–208. Springer Berlin Heidelberg, 2013.

[Pol04] Roswitha Poll. Informationsverhalten und informationsbedarf der wissenschaft. teil 1 der nutzungsanalyse des systems der überregionalen literatur- und informationsversorgung. *Zeitschrift für Bibliothekswesen und Bibliographie*, (2):59–65, 2004.

[Pow92] Patrick Powell. Resim - an algorithm for finding the similarity of regular expression based patterns and strings. In *Conference Record of The Twenty-Sixth Asilomar Conference on Signals, Systems and Computers*, 1992.

[PZ09] Michael Panzer and Marcia Lei Zeng. Modeling classification systems in SKOS: some challenges and best-practice recommendations. In *Proceedings of the 2009 International Conference on Dublin Core and Metadata Applications*, pages 3–14. Dublin Core Metadata Initiative, 2009.

[PZC06] Cristian Pérez de Laborda, Matthäus Zloch, and Stefan Conrad. RDQuery - Querying relational databases on-the-fly with RDF-QL. In *Poster and Demo Proceedings of the 15th International Conference on Knowledge Engineering* *and Knowledge Management, EKAW 2006, Podebrady, Czech Republic*, pages 19–20, 2006.

[RA07] Yves Raimond and Samer Abdallah. The event ontology. http://motools.sf.net/event/event.html, 2007.

[Rah11] Erhard Rahm. Towards large-scale schema and ontology matching. In Zohra Bellahsene, Angela Bonifati, and Erhard Rahm, editors, *Schema Matching and Mapping*, pages 3–27. 2011.

[RB01] Erhard Rahm and Philip A. Bernstein. A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4):334–350, December 2001.

[Rey14] Dave Reynolds. The organization ontology. http://www.w3.org/TR/vocab-org/, 2014.

[RNX$^+$12] Shu Rong, Xing Niu, EvanWei Xiang, Haofen Wang, Qiang Yang, and Yong Yu. A machine learning approach for instance matching based on similarity metrics. In Philippe Cudré-Mauroux, Jeff Heflin, Evren Sirin, Tania Tudorache, Jérôme Euzenat, Manfred Hauswirth, Josiane Xavier Parreira, Jim Hendler, Guus Schreiber, Abraham Bernstein, and Eva Blomqvist, editors, *The Semantic Web - ISWC 2012 - 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I*, volume 7649 of *Lecture Notes in Computer Science*, pages 460–475. Springer, 2012.

[RTM08] Hajo Rijgersberg, Jan L. Top, and Marcel Meinders. Use of a quantitative research ontology in e-science. In *AAAI Spring Symposium: Semantic Scientific Knowledge Integration*, pages 87–92, 2008.

[SAS11] Fabian M. Suchanek, Serge Abiteboul, and Pierre Senellart. PARIS: probabilistic alignment of relations, instances, and schema. *Proc. VLDB Endow.*, 5(3):157–168, November 2011.

[SBH$^+$05] York Sure, Stephan Bloehdorn, Peter Haase, Jens Hartmann, and Daniel Oberle. The SWRC ontology - semantic web for research communities. In *Proceedings of the 12th Portuguese Conference on Artificial Intelligence*, pages 218–231. Springer, 2005.

[SBJC14] Max Schmachtenberg, Christian Bizer, Anja Jentzsch, and Richard Cyganiak. Linking open data cloud diagram. http://lod-cloud.net/, 2014.

[SC08] Leo Sauermann and Richard Cyganiak. Cool URIs for the semantic web. http://www.w3.org/TR/cooluris/, 2008. W3C Interest Group Note 03 December 2008.

[SDM02] SDMX. Statistical data and metadata exchange. http://sdmx.org/, 2002.

[SDM09] SDMX. SDMX Content-Oriented Guidelines, 2009.

[SE05] Pavel Shvaiko and Jérôme Euzenat. A survey of schema-based matching approaches. In Stefano Spaccapietra, editor, *Journal on Data Semantics*

IV, volume 3730 of *Lecture Notes in Computer Science*, chapter 5, pages 146–171. Springer, 2005.

[SE11] Pavel Shvaiko and Jérôme Euzenat. Ontology matching: State of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, 99(PrePrints), 2011.

[SHE05] Rainer Schnell, Paul B. Hill, and Elke Esser. *Methoden der empirischen Sozialforschung*. Oldenburg, München, 9 edition, 2005.

[SHH^{+}09] Satya S. Sahoo, Wolfgang Halb, Sebastian Hellmann, Kingsley Idehen, Ted Thibodeau, Sören Auer, Juan Sequeda, and Ahmed Ezzat. A survey of current approaches for mapping of relational databases to RDF. Technical report, W3C, 2009.

[SIRK08] Ed Summers, Antoine Isaac, Clay Redding, and Dan Krech. LCSH, SKOS and linked data. In *Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications*, pages 25–33. Dublin Core Metadata Initiative, 2008.

[SM01] Gerd Stumme and Alexander Maedche. FCA-MERGE: Bottom-up merging of ontologies. In Bernhard Nebel, editor, *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI 2001, Seattle, Washington, USA, August 4-10, 2001*, pages 225–230. Morgan Kaufmann, 2001.

[SP11] Owen Sacco and Alexandre Passant. A privacy preference ontology (PPO) for linked data. In Christian Bizer, Tom Heath, Tim Berners-Lee, and Michael Hausenblas, editors, *WWW2011 Workshop on Linked Data on the Web, Hyderabad, India, March 29, 2011*, volume 813 of *CEUR Workshop Proceedings*, 2011.

[SS05] York Sure and Rudi Studer. Semantic web technologies for digital libraries. *Library Management*, 26(4/5):190–195, April 2005. Special Issue: Semantic Web.

[Stu09] Heiner Stuckenschmidt. A semantic similarity measure for ontology-based information. In *Proceedings of the 8th International Conference on Flexible Query Answering Systems*, FQAS '09, pages 406–417, Berlin, Heidelberg, 2009. Springer-Verlag.

[Sve07] Lars G. Svensson. National libraries and the semantic web: Requirements and applications. In *ICSD - International Conference for Digital Libraries and the Semantic Web - Proceedings*, 2007.

[SZ09a] Maximilian Stempfhuber and Benjamin Zapilko. Ein Ebenenmodell für die semantische Integration von Primärdaten und Publikationen in Digitalen Bibliotheken. In *Proceedings der 12. Tagung der Deutschen ISKO, International Society for Knowledge Organization*, 2009.

[SZ09b] Maximilian Stempfhuber and Benjamin Zapilko. Einbindung von Primärdaten in Digitale Bibliotheken. In Rainer Kuhlen, editor, *Information: Droge, Ware oder Commons? Proceedings des 11. Internationalen Symposiums für Informationswissenschaft (ISI 2009)*, pages 539–546. vwh Verlag Werner Hülsbusch, 2009. ISBN: 978-3-940317-43-8.

[SZ09c] Maximilian Stempfhuber and Benjamin Zapilko. Integrated retrieval of research data and publications in digital libraries. In Susanna Mornati and Turid Hedlund, editors, *Rethinking Electronic Publishing: Innovation in Communication Paradigms and Technologies. Proceedings of the 13th International Conference on Electronic Publishing*, pages 613–620, 2009.

[SZ09d] Maximilian Stempfhuber and Benjamin Zapilko. Modelling text-fact-integration in digital libraries. In *Electronic Proceedings of the 5th International Conference on e-Social Science*, 2009.

[TCC$^+$10] Giovanni Tummarello, Richard Cyganiak, Michele Catasta, Szymon Danielczyk, Renaud Delbru, and Stefan Decker. Sig.ma: Live views on the web of data. *J. Web Sem.*, 8(4):355–364, 2010.

[TDO07] Giovanni Tummarello, Renaud Delbru, and Eyal Oren. Sindice.com: weaving the open linked data. In Karl Aberer, Key-Sun Choi, Natasha Fridman Noy, Dean Allemang, Kyung-Il Lee, Lyndon J. B. Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007*, volume 4825 of *Lecture Notes in Computer Science*, pages 552–565. Springer, 2007.

[TP11] Anne E. Thessen and David J. Patterson. Data issues in the life sciences. *ZooKeys*, 150(150):15–51, November 2011.

[TQJ11] Yao Tie, Xu Qiang, and He Jin. A grouping algorithm based on regular expression similarity for dfa construction. In *13th International Conference on Communication Technology (ICCT)*, 2011.

[Tra10] Duc Thanh Tran. *Semantic Web Search - A Process-Oriented Perspective on Data Retrieval on the Semantic Web.* Phd thesis, KIT, Fakultät für Wirtschaftswissenschaften, Karlsruhe, 2010.

[UG04] Michael Uschold and Michael Gruninger. Ontologies and semantics for seamless connectivity. *SIGMOD Rec.*, 33(4):58–64, December 2004.

[vAMMS06] Mark van Assem, Véronique Malaisé, Alistair Miles, and Guus Schreiber. A method to convert thesauri to SKOS. In York Sure and John Domingue, editors, *The Semantic Web: Research and Applications, 3rd European Semantic Web Conference, ESWC 2006, Budva, Montenegro, June 11-14, 2006, Proceedings*, volume 4011 of *Lecture Notes in Computer Science*, pages 95–109. Springer, 2006.

[Vat10] Bernard Vatant. Porting library vocabularies to the semantic web, and back. a win-win round trip. In *76th IFLA General Conference And Assembly*, 2010.

[VBGK09] Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. Discovering and maintaining links on the web of data. In Abraham Bernstein, David R. Karger, Tom Heath, Lee Feigenbaum, Diana Maynard, Enrico Motta, and Krishnaprasad Thirunarayan, editors, *The Semantic Web - ISWC 2009, 8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, October 25-29, 2009. Proceedings*, volume 5823 of *Lecture Notes in Computer Science*, pages 650–665. Springer, 2009.

[VHC10] Robert Vesse, Wendy Hall, and Les Carr. Preserving linked data on the semantic web by the application of link integrity techniques from hypermedia. In Christian Bizer, Tom Heath, Tim Berners-Lee, and Michael Hausenblas, editors, *Proceedings of the WWW2010 Workshop on Linked Data on the Web, LDOW 2010, Raleigh, USA, April 27, 2010*, volume 628 of *CEUR Workshop Proceedings*, 2010.

[vHEM12] Willem Robert van Hage, Marieke Erp, and Véronique Malaisé. Linked open piracy: A story about e-science, linked data, and statistics. *J. Data Semantics*, 1(3):187–201, 2012.

[VIV] VIVO. VIVO Ontology. http://sourceforge.net/apps/mediawiki/vivo/index.php?title=Ontolog

[vK90] Ernst v. Kardorff. Qualitative Sozialforschung: Versuch einer Standortbestimmung. In Uwe Flick, Ernst v. Kardorff, Heiner Keupp, Lutz v. Rosenstiel, and Stephan Wolff, editors, *Handbuch qualitativer Sozialforschung*, pages 3–8. Psychologie Verlagsunion, 1990.

[VLH$^+$10] Denny Vrandecic, Christoph Lange, Michael Hausenblas, Jie Bao, and Li Ding. Semantics of governmental statistics data. In *Proceedings of the WebSci10: extending the frontiers of society on-line; April 26-27th 2010, Raleigh, NC, USA*, 2010.

[vOHdB11] Jacco van Ossenbruggen, Michiel Hildebrand, and Viktor de Boer. Interactive vocabulary alignment. In *Theory and Practice of Digital Libraries - First International Conference, TPDL 2011. Proceedings*, pages 296–307, 2011.

[vR79] Cornelis J. van Rijsbergen. *Information Retrieval.* Butterworth, 1979.

[WAZM10] Andias Wira-Alam, Benjamin Zapilko, and Philipp Mayr. An experimental approach for collecting snippets describing the relations between wikipedia articles. In *Proceedings of the WebSci10: extending the frontiers of society on-line; April 26-27th 2010, Raleigh, NC, USA*, 2010.

[WBZMZ13] Dennis Wegener, Erdal Baran, Wolfgang Zenk-Möltgen, and Benjamin Zapilko. Towards integrating the research data life cycle of the social sciences based on semantic technology. In *Workshop Applications of Semantic*

Technologies (AST2013), 43. Jahrestagung der Gesellschaft für Informatik e.V. (GI), Koblenz, Germany, 2013.

[Win06] William E. Winkler. Overview of record linkage and current research directions. Technical report, 2006.

[Wis11] Wissenschaftsrat. *Empfehlungen zu Forschungsinfrastrukturen.* Geschäftsstelle des WR, 2011.

[WLT11] Andreas Wagner, Günter Ladwig, and Thanh Tran. Browsing-oriented semantic faceted search. In Abdelkader Hameurlain, Stephen W. Liddle, Klaus-Dieter Schewe, and Xiaofang Zhou, editors, *Database and Expert Systems Applications - 22nd International Conference, DEXA 2011, Toulouse, France, August 29 - September 2, 2011. Proceedings, Part I*, volume 6860 of *Lecture Notes in Computer Science*, pages 303–319. Springer, 2011.

[WSB+09] William Wong, Hanna Stelmaszewska, Nazlin Bhimani, Sukhbinder Barn, and Balbir Barn. JISC user behaviour observational study: User behaviour in resource discovery. final report. Technical report, November 2009.

[WSRH10] Andreas Wagner, Sebastian Speiser, Oliver Raabe, and Andreas Harth. Linked data for a privacy-aware smart grid. In *GI Jahrestagung*, pages 449–454, 2010.

[WWH+10] Anja Wilde, Agnieszka Wenninger, Oliver Hopt, Philipp Schaer, and Benjamin Zapilko. Aktivitäten von GESIS im Kontext von Open Data und Zugang zu sozialwissenschaftlichen Forschungsergebnissen. In M. Ockenfeld, editor, *Semantic Web & Linked Data: Elemente zukünftiger Informationsinfrastrukturen; 1. DGI-Konferenz; 62. Jahrestagung der DGI, Frankfurt am Main, 7.-9. Oktober 2010; Proceedings*, pages 183–192. Dt. Ges. für Informationswiss. u. Informationspraxis, 2010.

[ZC09] Katrin Zaiss and Stefan Conrad. Partial ontology matching using instance features. In Robert Meersman, Tharam S. Dillon, and Pilar Herrero, editors, *On the Move to Meaningful Internet Systems: OTM 2009, Confederated International Conferences, CoopIS, DOA, IS, and ODBASE 2009, Vilamoura, Portugal, November 1-6, 2009, Proceedings, Part II*, volume 5871 of *Lecture Notes in Computer Science*, pages 1201–1208. Springer, 2009.

[ZHM11] Benjamin Zapilko, Andreas Harth, and Brigitte Mathiak. Enriching and analysing statistics with linked open data. In *NTTS - Conference on New Techniques and Technologies for Statistics.* Eurostat, 2011.

[ZM11a] Benjamin Zapilko and Brigitte Mathiak. Defining and executing assessment tests on linked data for statistical analysis. In Olaf Hartig, Andreas Harth, and Juan Sequeda, editors, *Proceedings of the Second International Workshop on Consuming Linked Data (COLD2011), Bonn, Germany, October 23, 2011*, volume 782 of *CEUR Workshop Proceedings*, 2011.

[ZM11b] Benjamin Zapilko and Brigitte Mathiak. Performing statistical methods on linked data. In *Proceedings of the 2011 International Conference on Dublin Core and Metadata Applications*. Dublin Core Metadata Initiative, 2011.

[ZM14] Benjamin Zapilko and Brigitte Mathiak. Object property matching utilizing the overlap between imported ontologies. In Valentina Presutti, Claudia d'Amato, Fabien Gandon, Mathieu d'Aquin, Steffen Staab, and Anna Tordai, editors, *The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, volume 8465 of *Lecture Notes in Computer Science*. Springer, 2014.

[ZS09a] Benjamin Zapilko and York Sure. Converting the TheSoz to SKOS. GESIS Technical Report 07/2009, GESIS - Leibniz Institute for the Social Sciences, 2009. ISSN: 1868-9051.

[ZS09b] Benjamin Zapilko and York Sure. Neue Möglichkeiten für die Wissensorganisation durch die Kombination von Digital Library Verfahren mit Standards des Semantic Web. In Heinz-Peter Ohly, editor, *Proceedings der 12. Tagung der Deutschen ISKO, International Society for Knowledge Organization*, 2009.

[ZS09c] Benjamin Zapilko and York Sure. Transferring the shell model to the semantic web and the impact on text-fact-integration. In *ICSD - International Conference for Digital Libraries and the Semantic Web - Proceedings*, pages 1–7. University of Trento, 2009. ISBN: 978-88-8443-302-2.

[ZSC09] Katrin Zaiss, Tim Schlueter, and Stefan Conrad. Instance-based ontology matching using different kinds of formalisms. In *Proceedings of the International Conference on Semantic Web Engineering, Oslo, Norway, July*, pages 29–31, 2009.

[ZSMM13] Benjamin Zapilko, Johann Schaible, Philipp Mayr, and Brigitte Mathiak. TheSoz: A SKOS representation of the Thesaurus for the Social Sciences. *Semantic Web*, 4(3):257–263, 2013.

[ZZS12] Benjamin Zapilko, Matthäus Zloch, and Johann Schaible. Utilizing regular expressions for instance-based schema matching. In Pavel Shvaiko, Jérôme Euzenat, Anastasios Kementsietsidis, Ming Mao, Natasha Noy, and Heiner Stuckenschmidt, editors, *Proceedings of the 7th International Workshop on Ontology Matching, Boston, MA, USA, November 11, 2012*, volume 946 of *CEUR Workshop Proceedings*, 2012.

Appendix

A Guideline for the Qualitative Interviews

The following section contains the complete guideline according to which the qualitative interviews have been conducted. The guideline has been written in German language, because all participants of the interviews have been German native speakers or have been at least fluent in German.

Allgemeine Einleitung:

- [Kurze Vorstellung der eigenen Person]
- Mein Forschungsinteresse bezieht sich auf die Verknüpfung und Integration von heterogenen Daten und Informationen, die ursprünglich voneinander separiert vorliegen, aber für verschiedene Informationsbedürfnisse gemeinsam genutzt werden.
- [Motivation zu den Experteninterviews: Hier sollte glaubhaft vermittelt werden, warum gerade mit der Person X ein Interview geführt wird. Dieser Teil wird individuell auf die jeweilige Person angepasst.]
- Der Ablauf des Interviews wird sich wie folgt gestalten. Zunächst stelle ich Dir/Ihnen ein paar allgemeine Fragen zur Datenverknüpfung und -integration. Danach möchte ich gerne anhand eines konkreten Beispiels aus Deinem/Ihren Arbeitsumfeld Probleme der Datenverknüpfung eruieren und Lösungsansätze innerhalb eines alternativen Szenarios diskutieren.

In meiner eigenen Forschungsarbeit beschäftige ich mich mit der Verknüpfung von Daten und deren Integration basierend auf standardisierten Webtechnologien. Dabei geht es insbesondere darum, Daten und Informationen, die im Web veröffentlicht werden (z.B. auf Webseiten, aber auch über Datenbanken) mit mehr aussagekräftiger Semantik auszuzeichnen und darüber mit anderen Daten in Beziehung zu setzen. Das können beispielsweise Literaturinformationen, Studien auf verschiedenen Erschließungsebenen oder Statistikdaten sein, die alle wiederum u.U. von verschiedenen Datenanbietern an unterschiedlichen Orten im Web veröffentlicht werden. Um allerdings noch genauer zu verstehen, welche Arten und Teile von Daten miteinander kombinierbar und integrierbar sind – vor allem aus inhaltlicher Sicht – und zu welchen Problemen und Herausforderungen es bei einer Verknüpfung kommen kann, möchte ich Dich/Sie bitten, mir zu diesem Thema aus Deinem/Ihrem Arbeitsalltag zu erzählen.

- 1. Beschreibe mir bitte, bei welchen Arbeiten Du/Sie mit der Verknüpfung oder Integration von Daten oder Informationen zu tun hast oder konfrontiert wirst.
- 2. Welches Beispiel lässt sich dabei aus Deiner/Ihrer Arbeit nennen, bei dem die Verknüpfung oder Integration von unterschiedlichen Daten von großer Bedeutung oder besonders typisch sind?

In den nachfolgenden Fragen möchte ich gerne mehr über besondere Herausforderungen bei dieser Arbeit erfahren. Ich möchte Sie/Dich daher zunächst bitten, zu beschreiben, welche Arbeitsschritte für Sie/Dich dabei anfallen und welche davon besonders zeitintensiv sind.

- 3. Welche Arbeitsschritte sind zu bewältigen und welche davon würden Sie/würdest Du dabei als besonders zeitintensiv beschreiben? Was genau ist dabei aufwändig?
- 4. Bei welchen Arbeitsschritten kann es dabei zu Hürden oder Schwierigkeiten kommen? Was ist dabei besonders schwierig?
- 5. Wie häufig wird man mit den zeitintensiven Aufgaben / Problemstellungen konfrontiert?
- 6. Welche Verknüpfungen oder Integration von Daten oder Informationen, mit denen Du/Sie arbeitest, vermisst Du/Sie? Welche Verbindungen wären aus Deiner/Ihrer Sicht sinnvoll?

Bei den nachfolgenden Fragen möchte ich Dich/Sie bitten, dass Du/Sie sich in konkrete Situationen hineindenken bzw. sich konkrete Situationen vorstellen und im Folgenden beschreiben, wie Du/Sie mit den Situationen umgehst/umgehen.

Beispielszenario vorstellen: Angenommen Du suchst/Sie suchen im Web nach zwei verschiedenen Informationen oder Studien, die in einem bestimmten Kriterium übereinstimmen oder vergleichbar sein sollen [Beispiele individuell auf die Person zugeschnitten, z.B., die Suche nach vergleichbaren Variablen und Indikatoren bei Forschungsdaten, die Suche nach Publikationen zu einer bestimmten Studie (und umgekehrt), eine themenorientierte Recherche unabhängig vom Typ des Ergebnisses, . . .].

- 7a. Wie gehst Du/Sie vor, wenn die Informationen, die Du/Sie suchen, sich nicht gemeinsam finden lassen?
- 7b. Wie oder wo kannst Du/Sie herausfinden, ob es die Information gibt?
- 7c. Worin besteht hierbei die Schwierigkeit?

Alternativszenario vorstellen: Angenommen die Information über die fehlende Verknüpfung würde existieren und wäre Nutzern verfügbar, bspw. durch die Metadaten eines Ergebnisses einsehbar, direkt integriert im Ergebnis dargestellt oder anklickbar [z.B. interne und externe Verlinkung zu anderen Daten, z.B. Literatur und Studien, informationstypen-übergreifende Verschlagwortung, Variablen und Indikatoren von Forschungsdaten, Codelisten, Autorenvernetzung, ...].

- 7d. Wo würdest Du/Sie Anwendungen oder Anknüpfungspunkte sehen?
- 7e. In welcher Art und Weise würde sich das Problem aus dem Beispielszenario durch diese Gegebenheiten verändern?
- 7f. Wie würde solch eine Veränderung genau aussehen?
- 7g. Welche weiteren Veränderungen – abseits des konkreten Problems – könntest Du/Sie sich unter solchen Rahmenbedingungen vorstellen?
- 8. Gibt es von Ihrer/Deiner Seite weitere Themen oder Aspekte, die Sie/Du in dem Kontext der Verknüpfung und Integration heterogener Daten ansprechen möchten?

B Extension of the SWRC Ontology

The following Listing B.1 depicts the extended class and properties of the SWRC ontology presented in Section 4.2.

Listing B.1: Extension of the SWRC Ontology.

```
swrc:Dataset
a owl:Class ;
rdfs:label "Datensatz"@de ;
owl:equivalentClass qb:DataSet .

swrc:acronym a owl:DatatypeProperty ;
rdfs:label "akronym"@de ;
rdfs:comment "added 2011-11-01"^^xsd:string ;
rdfs:subPropertyOf swrc:title .

swrc:datasetInfo a owl:ObjectProperty ;
rdfs:label "datensatzInfo"@de ;
rdfs:comment "added 2011-11-01"^^xsd:string ;
rdfs:domain swrc:Dataset ;
rdfs:range swrc:Document ;
owl:inverseOf swrc:describesDataset .

swrc:describesDataset a owl:ObjectProperty ;
rdfs:label "beschreibtDatensatz"@de ;
rdfs:comment "added 2011-11-01"^^xsd:string ;
rdfs:domain swrc:Document ;
rdfs:range swrc:Dataset ;
owl:inverseOf swrc:datasetInfo .

swrc:doi a owl:DatatypeProperty ;
rdfs:label "doi"@de ;
rdfs:comment "added 2011-11-01"^^xsd:string ;
rdfs:subPropertyOf dcterms:identifier .

swrc:endHour a owl:DatatypeProperty ;
rdfs:label "EndeZeit"@de ;
rdfs:comment "added 2011-11-01"^^xsd:string ;
rdfs:subPropertyOf swrc:date .

swrc:fee a owl:DatatypeProperty ;
rdfs:label "gebühren"@de ;
rdfs:comment "added 2011-11-01"^^xsd:string ;
rdfs:domain swrc:Event .

swrc:fundedBy a owl:ObjectProperty ;
```

```
rdfs:label "gefördertVon"@de ;
owl:inverseOf swrc:funds .

swrc:funds a owl:ObjectProperty ;
rdfs:label "fördert"@de ;
owl:inverseOf swrc:fundedBy .

swrc:graduatedIn a owl:ObjectProperty ;
rdfs:label "abschlussIn"^^xsd:string ;
rdfs:comment "added 2011-11-01"^^xsd:string ;
rdfs:domain swrc:Person ;
rdfs:range swrc:Topic .

swrc:handle a owl:DatatypeProperty ;
rdfs:label "handle"^^xsd:string ;
rdfs:comment "added 2011-11-01"^^xsd:string ;
rdfs:subPropertyOf dcterms:identifier .

swrc:hasDataCollector a owl:ObjectProperty ;
rdfs:label "hatDatenerheber"@de ;
rdfs:domain swrc:Dataset ;
rdfs:range [ a owl:Class ;
owl:unionOf ( swrc:Person swrc:Organization ) ] .

swrc:hasPrimaryResearcher a owl:ObjectProperty ;
rdfs:label "hatPrimaerforscher"@de ;
rdfs:domain [ a owl:Class ;
owl:unionOf ( swrc:Dataset swrc:Project ) ] ;
rdfs:range [ a owl:Class ;
owl:unionOf ( swrc:Person swrc:Organization ) ] .

swrc:hostedBy a owl:ObjectProperty ;
rdfs:label "veranstaltetVon"@de ;
rdfs:domain swrc:Event ;
owl:inverseOf swrc:hosts .

swrc:hosts a owl:ObjectProperty ;
rdfs:label "veranstaltet"@de ;
rdfs:range swrc:Event ;
owl:inverseOf swrc:hostedBy .

swrc:isbn a owl:DatatypeProperty ;
rdfs:label "isbn"@de ;
rdfs:comment "added 2011-11-01"^^xsd:string ;
rdfs:subPropertyOf dcterms:identifier .

swrc:issn a owl:DatatypeProperty ;
rdfs:label "issn"@de ;
rdfs:comment "added 2011-11-01"^^xsd:string ;
rdfs:subPropertyOf dcterms:identifier .

swrc:organizedBy a owl:ObjectProperty ;
rdfs:label "organisiertVon"@de ;
owl:inverseOf swrc:organizes .
```

```
swrc:organizes a owl:ObjectProperty ;
rdfs:label "organisiert"@de ;
owl:inverseOf swrc:organizedBy .

swrc:outcomeDataset a owl:ObjectProperty ;
rdfs:label "ergebnisDatensatz"@de ;
rdfs:comment "added 2011-11-01"^^xsd:string ;
rdfs:domain swrc:Project ;
rdfs:range swrc:Dataset .

swrc:producedBy a owl:ObjectProperty ;
rdfs:label "produziertVon"@de ;
rdfs:domain [ a owl:Class ;
owl:unionOf ( swrc:Product swrc:Dataset ) ] ;
rdfs:range [ a owl:Class ;
owl:unionOf ( swrc:Person swrc:Organization ) ] .

swrc:producer a owl:ObjectProperty ;
rdfs:label "produzent"@de ;
rdfs:domain [ a owl:Class ;
owl:unionOf ( swrc:Organization swrc:Person ) ] ;
rdfs:range [ a owl:Class ;
owl:unionOf ( swrc:Product swrc:Dataset ) ] .

swrc:startHour a owl:DatatypeProperty ;
rdfs:label "startZeit"@de ;
rdfs:comment "added 2011-11-01"^^xsd:string ;
rdfs:subPropertyOf swrc:date .

swrc:subtitle a owl:DatatypeProperty ;
rdfs:label "untertitel"@de ;
rdfs:comment "addad 2011-11-01"^^xsd:string ;
rdfs:subPropertyOf swrc:title .

swrc:urn a owl:DatatypeProperty ;
rdfs:label "urn"@de ;
rdfs:comment "added 2011-11-01"^^xsd:string ;
rdfs:subPropertyOf dcterms:identifier .
```

C DDI-RDF Discovery Vocabulary

Listing C.1 depicts the full model of the DDI-RDF Discovery Vocabulary presented in Section 4.3.

Listing C.1: The DDI-RDF Discovery Vocabulary.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix dc: <http://purl.org/dc/elements/1.1/>.
@prefix dcterms: <http://purl.org/dc/terms/>.
@prefix dcat: <http://www.w3.org/ns/dcat#>.
@prefix skos: <http://www.w3.org/2004/02/skos/core#>.
@prefix qb: <http://purl.org/linked-data/cube#>.
@prefix owl: <http://www.w3.org/2002/07/owl#>.
@prefix disco: <http://rdf-vocabulary.ddialliance.org/discovery#>.
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix adms: <http://www.w3.org/ns/adms#>.
@prefix org: <http://www.w3.org/ns/org#>.
@prefix prov: <http://www.w3.org/ns/prov#>.
@prefix xkos: <http://purl.org/linked-data/xkos#>.

#################################################################
# Ontology #
#################################################################

<http://rdf-vocabulary.ddialliance.org/discovery>
a owl:Ontology;
dc:title "DDI-RDF Discovery Vocabulary"@en;
rdfs:comment "This specification defines the DDI Discovery Vocabulary, an
    RDF Schema vocabulary that enables discovery of research and survey
    data on the Web. It is based on DDI (Data Documentation Initiative)
    XML formats."@en;
dc:contributor "Thomas Bosch", "Richard Cyganiak", "Joachim Wackerow",
    "Benjamin Zapilko";
dc:creator "Thomas Bosch", "Sarven Capadisli", "Franck Cotton", "Richard
    Cyganiak", "Arofan Gregory", "Benedikt Kämpgen", "Olof Olsson", "Heiko
    Paulheim", "Joachim Wackerow", "Benjamin Zapilko";
owl:versionInfo "Version 0.6 - 2014-09-25".

#################################################################
# Classes #
#################################################################

# AnalysisUnit class
# DDI3.1 r:AnalysisUnit
```

```
disco:AnalysisUnit
a rdfs:Class, owl:Class;
rdfs:label "Analysis Unit"@en, "Analyseeinheit"@de;
rdfs:comment "The process collecting data is focusing on the analysis of a
    particular type of subject. If, for example, the adult population of
    Finland is being studied, the AnalysisUnit would be individuals or
    persons."@en;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>;
rdfs:subClassOf skos:Concept.

# RepresentedVariable class
disco:RepresentedVariable
a rdfs:Class, owl:Class;
rdfs:label "Data element"@en, "Élément de donnée"@fr;
rdfs:comment "RepresentedVariables encompasse study-independent, re-usable
    parts of variables like occupation classification."@en;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# DataFile class
disco:DataFile
a rdfs:Class, owl:Class;
rdfs:label "Data file"@en, "Fichier de données"@fr;
rdfs:comment "The class DataFile, which is also a dcterms:Dataset,
    represents all the data files containing the microdata datasets."@en ;
rdfs:subClassOf dcat:Distribution;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# DescriptiveStatistics class
disco:DescriptiveStatistics
a rdfs:Class, owl:Class;
rdfs:label "Descriptive statistics"@en, "Statistique descriptive"@fr;
rdfs:comment "SummaryStatistics pointing to variables and
    CategoryStatistics pointing to categories and codes are both
    DescriptiveStatistics."@en;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# SummaryStatistics class
disco:SummaryStatistics
a rdfs:Class, owl:Class;
rdfs:label "Summary statistics"@en;
rdfs:comment "For SummaryStatistics, maximum values, minimum values, and
    standard deviations can be defined."@en;
rdfs:subClassOf disco:DescriptiveStatistics;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# CategoryStatistics class
# DDI3.1 p:CategoryStatistics
disco:CategoryStatistics
a rdfs:Class, owl:Class;
rdfs:label "Category statistics"@en;
rdfs:comment "For CategoryStatistics, frequencies, percentages, and
    weighted percentages can be defined."@en;
rdfs:subClassOf disco:DescriptiveStatistics;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.
```

```

# Instrument class (e.g., questionnaire, sensors, registers)
# XXX: Additional subclasses to be discussed.
# DDI3.1 d:Instrument
disco:Instrument
a rdfs:Class, owl:Class;
rdfs:label "Instrument"@en; rdfs:label "Instrument de collecte"@fr;
rdfs:comment "The data for the study are collected by an Instrument. The
    purpose of an Instrument, i.e. an interview, a questionnaire or
    another entity used as a means of data collection, is in the case of a
    survey to record the flow of a questionnaire, its use of questions,
    and additional component parts. A questionnaire contains a flow of
    questions."@en;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# LogicalDataSet class
disco:LogicalDataSet
a rdfs:Class, owl:Class;
rdfs:label "LogicalDataSet"@en, "Ensemble de données"@fr;
rdfs:comment "Each study has a set of logical metadata associated with the
    processing of data, at the time of collection or later during
    cleaning, and re-coding. LogicalDataSet represents the microdata
    dataset."@en;
rdfs:subClassOf dcat:Dataset;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# Question class
# skos:prefLabel represents question name
# DDI3.1 d:QuestionItem|d:MultipleQuestionItem
disco:Question
a rdfs:Class, owl:Class;
rdfs:label "Question"@en, "Question"@fr;
rdfs:comment "A Question is designed to get information upon a subject, or
    sequence of subjects, from a respondent."@en;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# disco:responseDomain
# cardinality at disco:Question: 0..n
# cardinality at disco:Representation: 1..n
disco:responseDomain
a rdf:Property, owl:ObjectProperty;
rdfs:label "responseDomain"@en;
rdfs:comment "The response domain of questions."@en;
rdfs:domain disco:Question;
rdfs:range disco:Representation;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# Questionnaire class
disco:Questionnaire
a rdfs:Class, owl:Class;
rdfs:label "Questionnaire"@en, "Fragebogen"@de;
rdfs:comment "A questionnaire contains a flow of questions. "@en;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>;
rdfs:subClassOf disco:Instrument.
```

```

# Study class
# DDI3.1 s:StudyUnit
disco:Study
a rdfs:Class, owl:Class;
rdfs:label "Study"@en, "Étude"@fr;
rdfs:comment "A Study represents the process by which a data set was
    generated or collected."@en;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# Study group class
disco:StudyGroup
a rdfs:Class, owl:Class;
rdfs:label "Study Group"@en, "Studiengruppe"@de;
rdfs:comment "In some cases, where data collection is cyclic or on-going,
    data sets may be released as a StudyGroup, where each cycle or wave of
    the data collection activity produces one or more data sets. This is
    typical for longitudinal studies, panel studies, and other types of
    series (to use the DDI term). In this case, a number of Study objects
    would be collected into a single StudyGroup."@en;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# Variable class
# DDI3.1 l:Variable
disco:Variable
a rdfs:Class, owl:Class;
rdfs:label "Variable"@en, "Variable"@fr;
rdfs:comment "Variables provide a definition of the column in a
    rectangular data file. Variable is a characteristic of a unit being
    observed. A variable might be the answer of a question, have an
    administrative source, or be derived from other variables."@en;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# Universe class
# skos:Concept/skos:notation represents universe name
# skos:Concept/skos:prefLabel represents universe label
# DDI3.1 c:Universe
disco:Universe
a rdfs:Class, owl:Class;
rdfs:label "Universe"@en, "Univers"@fr;
rdfs:comment "A Universe is the total membership or population of a
    defined class of people, objects or events."@en;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>;
rdfs:subClassOf skos:Concept.

disco:Mapping
a rdfs:Class, owl:Class;
rdfs:label "Mapping"@en;
rdfs:comment "Mappings betwenn DDI-RDF and DDI-XML"@en;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

#################################################################
# Datatype properties #
################################################################
```

```

# caseQuantity property
# DDI3.1 p:CaseQuantity
disco:caseQuantity
a rdf:Property, owl:DatatypeProperty;
rdfs:label "number of cases"@en;
rdfs:comment "case quantity of a DataFile."@en;
rdfs:domain disco:DataFile;
rdfs:range xsd:nonNegativeInteger;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# frequency property
disco:frequency
a rdf:Property, owl:DatatypeProperty;
rdfs:label "frequency"@en, "fréquence"@fr;
rdfs:comment "frequency"@en;
rdfs:domain disco:CategoryStatistics;
rdfs:range xsd:nonNegativeInteger;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# isPublic property
disco:isPublic
a rdf:Property, owl:DatatypeProperty;
rdfs:label "is public"@en, "ist öffentlich"@de;
rdfs:domain disco:LogicalDataSet;
rdfs:comment "The value true indicates that the dataset can be accessed (usually downloaded) by anyone."@en;     rdfs:range xsd:boolean;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# isValid property
disco:isValid
a rdf:Property, owl:DatatypeProperty;
rdfs:label "is valid"@en;
rdfs:domain skos:Concept;
rdfs:comment "Indicates if the code (represented by skos:Concept) is valid or missing."@en;
rdfs:range xsd:boolean;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# questionText property
# DDI3.1 d:QuestionText
disco:questionText
a rdf:Property, owl:DatatypeProperty;
rdfs:label "question text"@en, "Fragetext"@de;
rdfs:comment "question text"@en;
rdfs:domain disco:Question;
rdfs:range rdf:langString;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# percentage property
disco:percentage
a rdf:Property, owl:DatatypeProperty;
rdfs:label "percentage"@en, "pourcentage"@fr;
rdfs:comment "percentage"@en;
```

```
rdfs:domain disco:CategoryStatistics;
rdfs:range xsd:double;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# computationBase property
disco:computationBase
a rdf:Property, owl:DatatypeProperty;
rdfs:label "computation base"@en, "pourcentage"@fr;
rdfs:comment "computation base"@en;
rdfs:domain disco:CategoryStatistics;
rdfs:range rdf:langString;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# cumulativePercentage property
disco:cumulativePercentage
a rdf:Property, owl:DatatypeProperty;
rdfs:label "cumulative percentage"@en;
rdfs:comment "cumulative percentage"@en;
rdfs:domain disco:CategoryStatistics;
rdfs:range xsd:double;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# purpose property
# DDI3.1 s:Purpose
disco:purpose
a rdf:Property, owl:DatatypeProperty;
rdfs:label "purpose"@en, "Grund"@de;
rdfs:comment "The purpose of a Study of a StudyGroup."@en;
rdfs:domain [ a owl:Class; owl:unionOf (disco:Study disco:StudyGroup)];
rdfs:range rdf:langString;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# subtitle property
# DDI3.1 r:SubTitle
disco:subtitle
a rdf:Property, owl:DatatypeProperty;
rdfs:label "subtitle"@en, "Untertitel"@de;
rdfs:comment "The sub-title of a Study of a StudyGroup."@en;
rdfs:domain [ a owl:Class; owl:unionOf (disco:Study disco:StudyGroup)];
rdfs:range rdf:langString;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

disco:startDate
a rdf:Property, owl:DatatypeProperty;
rdfs:label "start date"@en;
rdfs:comment "start date"@en;
rdfs:domain dcterms:PeriodOfTime;
rdfs:range xsd:date;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

disco:endDate
a rdf:Property, owl:DatatypeProperty;
rdfs:label "end date"@en;
rdfs:comment "end date"@en;
```

```
rdfs:domain dcterms:PeriodOfTime;
rdfs:range xsd:date;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

disco:mappingDDI-L
a rdf:Property, owl:DatatypeProperty;
rdfs:label "Mapping from and to DDI-L"@en;
rdfs:comment "Mapping from and to DDI-L"@en;
rdfs:domain disco:Mapping;
rdfs:range rdf:langString;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

disco:mappingDDI-C
a rdf:Property, owl:DatatypeProperty;
rdfs:label "Mapping from and to DDI-C"@en;
rdfs:comment "Mapping from and to DDI-C"@en;
rdfs:domain disco:Mapping;
rdfs:range rdf:langString;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

disco:context a rdf:Property, owl:DatatypeProperty;
rdfs:label "context specifies conditions which have to be fulfilled for specific mappings"@en;
rdfs:comment "context specifies conditions which have to be fulfilled for specific mappings"@en;
rdfs:domain disco:Mapping;
rdfs:range rdf:langString;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

disco:variableQuantity
a rdf:Property, owl:DatatypeProperty;
rdfs:label "variable quantity"@en;
rdfs:comment "Variable quantity"@en;
rdfs:domain [ a owl:Class; owl:unionOf (disco:LogicalDataSet disco:DataFile)];
rdfs:range xsd:nonNegativeInteger;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

################################################################
# Object properties #
# Cardinalities are noted:
# Notation: Study -> Universe 1,...,n / 0,...,n
# Meaning: Study has 1,...,n universes; Universe has 0,...,n studies. #
################################################################

# analysisUnit property (different meaning than concept property)
# Variable -> AnalysisUnit 0,1 / 0,...,n
# Study -> AnalysisUnit 0,1 / 0,...,n
disco:analysisUnit
a rdf:Property, owl:ObjectProperty;
rdfs:label "analysis unit"@en, "Analyseeinheit"@de;
rdfs:comment "analysis unit of a Study, a StudyGroup, or a Variable."@en;
rdfs:domain [a owl:Class; owl:unionOf ( disco:Study disco:StudyGroup disco:Variable)];
```

```
rdfs:range disco:AnalysisUnit;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# basedOn property
# * Variable -> RepresentedVariable 0,1 / 0,...,n
disco:basedOn
a rdf:Property, owl:ObjectProperty;
rdfs:label "based on"@en;
rdfs:comment "points to the RepresentedVariable the Variable is based on."@en;
rdfs:domain disco:Variable;
rdfs:range disco:RepresentedVariable;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# collectionMode property
disco:collectionMode
a rdf:Property, owl:ObjectProperty;
rdfs:label "collection mode"@en, "Datenerfassungsmodus"@de;
rdfs:comment "mode of collection of a Questionnaire"@en;
rdfs:domain disco:Questionnaire;
rdfs:range skos:Concept;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# concept property
# * RepresentedVariable -> Concept 1 / 0,...,n
# * Question -> Concept 1,...,n / 0,...,n
# * Variable -> Concept 1 / 0,...,n
disco:concept
a rdf:Property, owl:ObjectProperty;
rdfs:label "concept"@en, "a pour concept"@fr;
rdfs:comment "points to the DDI concept of a RepresentedVariable, a Variable, or a Question"@en;
rdfs:domain [a owl:Class; owl:unionOf (disco:RepresentedVariable disco:Question disco:Variable)];
rdfs:range skos:Concept;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# aggregation property
# * LogicalDataSet -> qb:DataSet 0,...,n / 0,...,n (Use Case: Look whether a LogicalDataSet exists for a qb:DataSet)
disco:aggregation
a rdf:Property, owl:ObjectProperty;
rdfs:label "aggregation"@en;
rdfs:comment "points to the aggregated data set of a microdata data set."@en;
rdfs:domain disco:LogicalDataSet;
rdfs:range qb:DataSet;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# dataFile property
# * LogicalDataSet -> DataFile 0,...,n / 0,...,n
disco:dataFile
a rdf:Property, owl:ObjectProperty;
rdfs:label "data file"@en, "a pour fichier de données"@fr;
```

```
rdfs:comment "points to the DataFile of a Study or a LogicalDataSet."@en;
rdfs:domain [a owl:Class; owl:unionOf (disco:Study disco:LogicalDataSet)];
rdfs:range disco:DataFile;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# ddifile property
# (disco:Study disco:StudyGroup) -> foaf:Document 0,* / 0,*
disco:ddifile
a rdf:Property, owl:ObjectProperty;
rdfs:label "DDI file"@en, "DDI-Datei"@de;
rdfs:comment "points from a Study or a StudyGroup to the original DDI file which is a foaf:Document."@en;
rdfs:domain [a owl:Class; owl:unionOf (disco:Study disco:StudyGroup)];
rdfs:range foaf:Document;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# externalDocumentation property
# XXX: check whether skos:Concept is ok and cardinality
disco:externalDocumentation
a rdf:Property, owl:ObjectProperty;
rdfs:label "external documentation"@en, "externe Dokumentation"@de;
rdfs:comment "points from an Instrument to a foaf:Document which is the external documentation of the Instrument."@en;
rdfs:domain disco:Instrument;
rdfs:range foaf:Document;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# fundedBy property
disco:fundedBy
a rdf:Property, owl:ObjectProperty;
rdfs:label "funded by"@en;
rdfs:comment "points from a Study or a StudyGroup to the funding foaf:Agent which is either a foaf:Person or a org:Organization."@en;
rdfs:domain [a owl:Class; owl:unionOf (disco:Study disco:StudyGroup)];
rdfs:range foaf:Agent;
rdfs:subPropertyOf dcterms:contributor;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# inGroup property
# Study -> StudyGroup 0,1 / 0...*
disco:inGroup
a rdf:Property, owl:ObjectProperty;
rdfs:label "in group"@en;
rdfs:comment "points from a Study to the StudyGroup which contains the Study."@en;
rdfs:domain disco:Study;
rdfs:range disco:StudyGroup;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# inputVariable property (links DataSets to DDI variables)
# * qb:DataSet -> Variable 0,...,n / 0,...,n
disco:inputVariable
a rdf:Property, owl:ObjectProperty;
rdfs:label "input variable"@en, "variable en entrée"@fr;
```

```
rdfs:comment "Indicates the original Variable of an aggregated qb:DataSet."@en;
rdfs:domain qb:DataSet;
rdfs:range disco:Variable;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# instrument property
# * Study -> Instrument 1 / 0,...,n (cardinality might have to be changed if we want to have reusable instruments in the future)
disco:instrument
a rdf:Property, owl:ObjectProperty;
rdfs:label "instrument"@en, "a comme instrument"@fr;
rdfs:comment "Indicates the Instrument of a Study or a LogicalDataSet."@en;
rdfs:domain [a owl:Class; owl:unionOf (disco:Study disco:LogicalDataSet)];
rdfs:range disco:Instrument;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# kindOfData property
# (disco:Study disco:StudyGroup) -> skos:Concept 0,* / 0,1
disco:kindOfData
a rdf:Property, owl:ObjectProperty;
rdfs:label "kind of data"@en;
rdfs:domain [a owl:Class; owl:unionOf (disco:Study disco:StudyGroup)];
rdfs:range skos:Concept;
rdfs:comment "The general kind of data (e.g. geospatial, register, survey) collected in this study, given either as a skos:Concept, or as a blank node with attached free-text rdfs:label.";
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# product property
# * Study -> LogicalDataSet 0,...,n / 1,...,n
disco:product
a rdf:Property, owl:ObjectProperty;
rdfs:label "product"@en, "Produkt"@de;
rdfs:comment "Indicates the LogicalDataSets of a Studies."@en;
rdfs:domain disco:Study;
rdfs:range qb:LogicalDataSet;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# question property
# * Variable -> Question 0,...,n / 0,...,n
# * Questionnaire -> Question 1,...,n / 0,...,n
disco:question
a rdf:Property, owl:ObjectProperty;
rdfs:label "question"@en, "a comme question"@fr;
rdfs:comment "Indicates the Questions associated to Variables or contained in Questionnaires."@en;
rdfs:domain [a owl:Class; owl:unionOf (disco:Variable disco:Questionnaire)];
rdfs:range disco:Question;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# representation property
# * Variable -> Representation 1 / 0,...,n
```

```
disco:representation
a rdf:Property, owl:ObjectProperty;
rdfs:label "representation"@en, "a pour représentation"@fr;
rdfs:comment "RepresentedVariables and Variables can have a Representation
    whose individuals are either of the class rdfs:Datatype (to represent
    values) or skos:ConceptScheme (to represent code lists)."@en;
rdfs:domain [a owl:Class; owl:unionOf (disco:RepresentedVariable
    disco:Variable disco:Question)];
rdfs:range [a owl:Class; owl:unionOf (skos:ConceptScheme rdfs:Datatype)];
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# statisticsCategory property
# * DescriptiveStatistics -> Concept 0,...,n / 0,...,n
disco:statisticsCategory
a rdf:Property, owl:ObjectProperty;
rdfs:label "statistics category"@en, "a pour concept statistique"@fr;
rdfs:comment "Indicates the skos:Concept (representing codes and
    categories) of a specific CategoryStatistics individual."@en;
rdfs:domain disco:CategoryStatistics;
rdfs:range skos:Concept;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# statisticsDataFile property
# * DescriptiveStatistics -> DataFile 0,...,n / 0,...,n
disco:statisticsDataFile
a rdf:Property, owl:ObjectProperty;
rdfs:label "statistics data file"@en, "a pour fichier statistique"@fr;
rdfs:comment "Indicates the DataFile of a specific DesciptiveStatistics
    individual."@en;
rdfs:domain disco:DescriptiveStatistics;
rdfs:range disco:DataFile;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# statisticsVariable property
# * DescriptiveStatistics -> Variable 0,...,n / 0,...,n
disco:statisticsVariable
a rdf:Property, owl:ObjectProperty;
rdfs:label "statistics variable"@en, "a pour variable statistique"@fr;
rdfs:comment "Indicates the Variable of a specific SummaryStatistics
    individual."@en;
rdfs:domain disco:SummaryStatistics;
rdfs:range disco:Variable;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# weightedBy property
# * SummaryStatistics, CategoryStatistics -> Variable 0,...,n / 0 .. 1
disco:weightedBy
a rdf:Property, owl:ObjectProperty;
rdfs:label "weighted by"@en, ""@fr;
rdfs:comment "SummaryStatistics or CategoryStatistics resources may be
    weighted by a specific Variable."@en;
rdfs:domain [a owl:Class; owl:unionOf (disco:SummaryStatistics
    disco:CategoryStatistics)];
rdfs:range disco:Variable;
```

```
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# universe property
# * Study/Study -> Universe 1,...,n / 0,...,n
# * RepresentedVariable -> Universe 0,...,n / 0,...,n (Note:
    RepresentedVariable and Variable are the same thing in different
    states)
# * Variable -> Universe 1 / 0,...,n
# * Question -> Universe 1 / 0,...,n
# * LogicalDataSet -> Universe 1 / 0,...,n (Property: dataSetUniverse)
disco:universe
a rdf:Property, owl:ObjectProperty;
rdfs:label "universe"@en, "a comme univers"@fr;
rdfs:comment "Indicates the Universe(s) of Studies, StudyGrous,
    RepresentedVariables, Variables, Questions, and LogicalDataSets."@en;
rdfs:domain [a owl:Class; owl:unionOf (disco:Study disco:StudyGroup
    disco:RepresentedVariable disco:Variable disco:Question
    disco:LogicalDataSet)];
rdfs:range disco:Universe;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

# variable property
# * Study -> Variable 0,...,n / 1,...,n
# * LogicalDataSet -> Variable 0,...,n / 1,...,n
disco:variable
a rdf:Property, owl:ObjectProperty;
rdfs:label "variable"@en, "Variable"@de;
rdfs:comment "Indicates the Variable of a Study and points to Variable
    contained in the LogicalDataSet."@en;
rdfs:domain [a owl:Class; owl:unionOf (disco:Study disco:LogicalDataSet)];
rdfs:range disco:Variable;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.

disco:summaryStatisticsType
a rdf:Property, owl:ObjectProperty;
rdfs:label "summary statistics type"@en;
rdfs:comment "summary statistics type"@en;
rdfs:domain disco:SummaryStatistics;
rdfs:range skos:Concept;
rdfs:isDefinedBy <http://rdf-vocabulary.ddialliance.org/discovery>.
```

D Benchmark for Statistical Data

The following Listing D.1 depicts the seed ontology of the Benchmark for Statistical Data, which was introduced in Section 5.4.6. Additionally, the code lists of the seed ontology are summarized in this ontology. For the evaluation, they have been split in separated files and used as separated data sets.

Listing D.1: Seed Ontology of the Benchmark

```
@prefix : <http://lod.gesis.org/matchingstatistics/000#> .
@prefix STATTBOX: <http://lod.gesis.org/matchingstatistics/STATTBOX/> .
@prefix concepts: <http://lod.gesis.org/matchingstatistics/concepts/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://lod.gesis.org/matchingstatistics/000> a owl:Ontology .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry1>
a STATTBOX:DataEntry , owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#0-9> ;
STATTBOX:date "1958/05/08"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-F> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#LU> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#married>
    ;
STATTBOX:obsValue 738 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value1>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value1>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry10>
a STATTBOX:DataEntry , owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#40-49> ;
STATTBOX:date "1927/10/02"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-M> ;
```

```
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#UA> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#single>
    ;
STATTBOX:obsValue 586 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value3>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value3>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry11>
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#10-19> ;
STATTBOX:date "1981/08/02"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-F> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#FI> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#total>
    ;
STATTBOX:obsValue 886 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value3>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value8>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry12>
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#80-89> ;
STATTBOX:date "1963/17/09"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-N> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#PT> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#total>
    ;
STATTBOX:obsValue 318 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value8>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value8>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry13>
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#20-29> ;
```

```
STATTBOX:date "1993/08/03"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-T> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#FR> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#separated>
    ;
STATTBOX:obsValue 854 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value8>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value8>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry14>
 a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#10-19> ;
STATTBOX:date "1996/11/09"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-T> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#BY> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#married>
    ;
STATTBOX:obsValue 614 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value1>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value3>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry15>
 a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#10-19> ;
STATTBOX:date "1972/24/04"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-M> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#HR> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#separated>
    ;
STATTBOX:obsValue 684 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value1>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value8>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry16>
```

```
a STATTBOX:DataEntry , owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#40−49> ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex−N> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#PL> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#separated>
    ;
STATTBOX:obsValue 710 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value9>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value9>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry17>
a STATTBOX:DataEntry , owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#50−59> ;
STATTBOX:date "1971/31/03"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex−F> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#CZ> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#married>
    ;
STATTBOX:obsValue 71 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value2>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value3>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry18>
a STATTBOX:DataEntry , owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#0−9> ;
STATTBOX:date "1981/07/08"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex−N> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#LU> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#married>
    ;
STATTBOX:obsValue 710 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value1>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value3>
    .
```

```

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry19>
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#20-29> ;
STATTBOX:date "1934/17/03"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-U> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#BG> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#divorced>
    ;
STATTBOX:obsValue 506 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value8>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value1>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry2>
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#60-69> ;
STATTBOX:date "1994/17/04"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-U> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#LI> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#divorced>
    ;
STATTBOX:obsValue 11 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value2>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value2>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry20>
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#60-69> ;
STATTBOX:date "1983/01/08"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-T> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#NL> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#married>
    ;
STATTBOX:obsValue 573 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value8>
    ;
```

```
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value8>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry21>
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#0-9> ;
STATTBOX:date "1918/21/07"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-U> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#BA> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#separated>
    ;
STATTBOX:obsValue 202 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value1>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value3>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry22>
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#30-39> ;
STATTBOX:date "1930/25/01"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-U> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#MD> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#married>
    ;
STATTBOX:obsValue 248 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value5>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value1>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry23>
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#30-39> ;
STATTBOX:date "1973/27/09"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-M> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#BA> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#single>
    ;
STATTBOX:obsValue 232 ;
```

```
STATTBOX:occupation <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value5> ;
STATTBOX:satisfaction <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value9> .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry24>
 a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup <http://lod.gesis.org/matchingstatistics/concepts/agegroup#80-89> ;
STATTBOX:date "1963/27/03"^^xsd:string ;
STATTBOX:gender <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-N> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#NO> ;
STATTBOX:maritalStatus <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#divorced> ;
STATTBOX:obsValue 270 ;
STATTBOX:occupation <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value1> ;
STATTBOX:satisfaction <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value8> .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry25>
 a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup <http://lod.gesis.org/matchingstatistics/concepts/agegroup#60-69> ;
STATTBOX:date "1945/08/04"^^xsd:string ;
STATTBOX:gender <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-F> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#IT> ;
STATTBOX:maritalStatus <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#widowed> ;
STATTBOX:obsValue 77 ;
STATTBOX:occupation <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value5> ;
STATTBOX:satisfaction <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value8> .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry26>
 a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup <http://lod.gesis.org/matchingstatistics/concepts/agegroup#40-49> ;
STATTBOX:date "1922/20/06"^^xsd:string ;
STATTBOX:gender <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-T> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#PL> ;
STATTBOX:maritalStatus
```

```
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#separated>
    ;
STATTBOX:obsValue 755 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value5>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value2>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry27>
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#80-89> ;
STATTBOX:date "1991/02/02"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-T> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#RS> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#widowed>
    ;
STATTBOX:obsValue 515 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value9>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value9>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry28>
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#20-29> ;
STATTBOX:date "1967/05/05"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-N> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#UA> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#married>
    ;
STATTBOX:obsValue 84 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value2>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value2>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry29>
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#80-89> ;
STATTBOX:date "1942/07/05"^^xsd:string ;
STATTBOX:gender
```

```
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-F> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#FI> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#divorced>
    ;
STATTBOX:obsValue 188 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value3>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value9>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry3>
 a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#90> ;
STATTBOX:date "1954/07/02"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-M> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#ES> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#married>
    ;
STATTBOX:obsValue 80 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value3>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value1>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry30>
 a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#30-39> ;
STATTBOX:date "1997/03/04"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-N> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#UA> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#divorced>
    ;
STATTBOX:obsValue 125 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value5>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value1>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry31>
 a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
```

```
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#40-49> ;
STATTBOX:date "1980/23/02"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-T> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#GR> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#divorced>
    ;
STATTBOX:obsValue 861 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value5>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value3>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry32>
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#40-49> ;
STATTBOX:date "1979/27/07"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-U> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#SI> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#total>
    ;
STATTBOX:obsValue 705 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value3>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value2>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry33>
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#50-59> ;
STATTBOX:date "1979/28/09"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-T> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#BY> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#widowed>
    ;
STATTBOX:obsValue 156 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value5>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value1>
    .

```

```
<http://lod.gesis.org/matchingstatistics/STATDATA/Entry34>
 a STATTBOX:DataEntry , owl:NamedIndividual ;
STATTBOX:agegroup
     <http://lod.gesis.org/matchingstatistics/concepts/agegroup#20-29> ;
STATTBOX:date "1941/18/03"^^xsd:string ;
STATTBOX:gender
     <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-U> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#FR> ;
STATTBOX:maritalStatus
     <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#divorced>
     ;
STATTBOX:obsValue 284 ;
STATTBOX:occupation
     <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value3>
     ;
STATTBOX:satisfaction
     <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value1>
     .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry35>
 a STATTBOX:DataEntry , owl:NamedIndividual ;
STATTBOX:agegroup
     <http://lod.gesis.org/matchingstatistics/concepts/agegroup#80-89> ;
STATTBOX:date "1919/04/08"^^xsd:string ;
STATTBOX:gender
     <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-N> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#BA> ;
STATTBOX:maritalStatus
     <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#widowed>
     ;
STATTBOX:obsValue 455 ;
STATTBOX:occupation
     <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value2>
     ;
STATTBOX:satisfaction
     <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value3>
     .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry36>
 a STATTBOX:DataEntry , owl:NamedIndividual ;
STATTBOX:agegroup
     <http://lod.gesis.org/matchingstatistics/concepts/agegroup#70-79> ;
STATTBOX:date "1912/22/08"^^xsd:string ;
STATTBOX:gender
     <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-N> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#IT> ;
STATTBOX:maritalStatus
     <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#single>
     ;
STATTBOX:obsValue 280 ;
STATTBOX:occupation
     <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value4>
     ;
STATTBOX:satisfaction
```

```
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value2>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry37>
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#20-29> ;
STATTBOX:date "1940/11/09"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-M> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#IT> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#separated>
    ;
STATTBOX:obsValue 672 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value4>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value2>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry38>
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#80-89> ;
STATTBOX:date "1927/17/07"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-F> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#CH> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#widowed>
    ;
STATTBOX:obsValue 77 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value5>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value3>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry39>
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#70-79> ;
STATTBOX:date "1972/16/09"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-T> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#RU> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#separated>
    ;
STATTBOX:obsValue 648 ;
STATTBOX:occupation
```

```
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value2>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value9>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry4>
 a STATTBOX:DataEntry , owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#50-59> ;
STATTBOX:date "1959/27/07"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-T> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#MK> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#separated>
    ;
STATTBOX:obsValue 275 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value2>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value3>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry40>
 a STATTBOX:DataEntry , owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#20-29> ;
STATTBOX:date "1996/23/01"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-U> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#EE> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#single>
    ;
STATTBOX:obsValue 428 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value1>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value3>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry41>
 a STATTBOX:DataEntry , owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#50-59> ;
STATTBOX:date "1979/18/09"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-N> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#GB> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#total>
    ;
STATTBOX:obsValue 306 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value2>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value8>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry42>
 a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#10-19> ;
STATTBOX:date "1981/30/04"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-F> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#NO> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#married>
    ;
STATTBOX:obsValue 134 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value8>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value9>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry43>
 a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#20-29> ;
STATTBOX:date "1966/07/02"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-N> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#NL> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#married>
    ;
STATTBOX:obsValue 201 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value4>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value8>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry44>
 a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#80-89> ;
STATTBOX:date "1912/13/07"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-T> ;
```

```
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#BY> ;
STATTBOX:maritalStatus <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#widowed> ;
STATTBOX:obsValue 25 ;
STATTBOX:occupation <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value3> ;
STATTBOX:satisfaction <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value2> .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry45>
 a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup <http://lod.gesis.org/matchingstatistics/concepts/agegroup#70-79> ;
STATTBOX:date "1998/11/08"^^xsd:string ;
STATTBOX:gender <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-U> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#IS> ;
STATTBOX:maritalStatus <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#divorced> ;
STATTBOX:obsValue 676 ;
STATTBOX:occupation <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value4> ;
STATTBOX:satisfaction <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value3> .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry46>
 a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup <http://lod.gesis.org/matchingstatistics/concepts/agegroup#80-89> ;
STATTBOX:date "1967/15/09"^^xsd:string ;
STATTBOX:gender <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-F> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#SE> ;
STATTBOX:maritalStatus <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#single> ;
STATTBOX:obsValue 181 ;
STATTBOX:occupation <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value8> ;
STATTBOX:satisfaction <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value3> .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry47>
 a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup <http://lod.gesis.org/matchingstatistics/concepts/agegroup#30-39> ;
```

```
STATTBOX:date "1971/18/06"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-F> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#IT> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#single>
    ;
STATTBOX:obsValue 857 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value5>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value1>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry48>
 a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#40-49> ;
STATTBOX:date "1911/07/08"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-F> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#LV> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#total>
    ;
STATTBOX:obsValue 484 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value4>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value3>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry49>
 a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#80-89> ;
STATTBOX:date "1976/21/04"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-F> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#NL> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#separated>
    ;
STATTBOX:obsValue 862 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value9>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value3>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry5>
```

```
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#70-79> ;
STATTBOX:date "1936/17/03"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-M> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#CZ> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#total>
    ;
STATTBOX:obsValue 874 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value9>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value2>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry50>
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#40-49> ;
STATTBOX:date "1964/04/06"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-U> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#IE> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#widowed>
    ;
STATTBOX:obsValue 368 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value2>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value2>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry6>
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#30-39> ;
STATTBOX:date "1925/23/08"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-U> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#RU> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#married>
    ;
STATTBOX:obsValue 27 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value3>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value2>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry7>
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#50-59> ;
STATTBOX:date "1994/01/02"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-F> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#MD> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#separated>
    ;
STATTBOX:obsValue 485 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value1>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value1>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry8>
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#40-49> ;
STATTBOX:date "1932/13/08"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-T> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#AZ> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#single>
    ;
STATTBOX:obsValue 308 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value4>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value1>
    .

<http://lod.gesis.org/matchingstatistics/STATDATA/Entry9>
a STATTBOX:DataEntry, owl:NamedIndividual ;
STATTBOX:agegroup
    <http://lod.gesis.org/matchingstatistics/concepts/agegroup#40-49> ;
STATTBOX:date "1989/28/03"^^xsd:string ;
STATTBOX:gender
    <http://lod.gesis.org/matchingstatistics/concepts/gender#sex-N> ;
STATTBOX:geo <http://lod.gesis.org/matchingstatistics/concepts/geo#LT> ;
STATTBOX:maritalStatus
    <http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#widowed>
    ;
STATTBOX:obsValue 440 ;
STATTBOX:occupation
    <http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value1>
    ;
STATTBOX:satisfaction
    <http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value3>
    .

STATTBOX:agegroup
a owl:ObjectProperty .

STATTBOX:date
a owl:DatatypeProperty .

STATTBOX:gender
a owl:ObjectProperty .

STATTBOX:geo
a owl:ObjectProperty .

STATTBOX:maritalStatus
a owl:ObjectProperty .

STATTBOX:obsValue
a owl:DatatypeProperty .

STATTBOX:occupation
a owl:ObjectProperty .

STATTBOX:satisfaction
a owl:ObjectProperty .

<http://lod.gesis.org/matchingstatistics/concepts/geo#AD>
a owl:Class ;
rdfs:label "Andorra"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#AL>
a owl:Class ;
rdfs:label "Albania"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#AM>
a owl:Class ;
rdfs:label "Armenia"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#AT>
a owl:Class ;
rdfs:label "Austria"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#BE>
a owl:Class ;
rdfs:label "Belgium"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#CY>
a owl:Class ;
rdfs:label "Cyprus"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#DE>
```

```
a owl:Class ;
rdfs:label "Germany"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#DK>
a owl:Class ;
rdfs:label "Denmark"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#FO>
a owl:Class ;
rdfs:label "Faeroe Islands"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#GE>
a owl:Class ;
rdfs:label "Georgia"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#GI>
a owl:Class ;
rdfs:label "Gibraltar"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#HU>
a owl:Class ;
rdfs:label "Hungary"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#MC>
a owl:Class ;
rdfs:label "Monaco"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#MT>
a owl:Class ;
rdfs:label "Malta"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#RO>
a owl:Class ;
rdfs:label "Romania"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#SK>
a owl:Class ;
rdfs:label "Slovakia"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#SM>
a owl:Class ;
rdfs:label "San Marino"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#TR>
a owl:Class ;
rdfs:label "Turkey"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#VA>
a owl:Class ;
rdfs:label "Vatican City State"@en .

<http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value4>
a owl:Class ;
rdfs:label "Very dissatisfied"@en .
```

```

rdfs:label a owl:AnnotationProperty .

<http://lod.gesis.org/matchingstatistics/concepts/agegroup#90>
a owl:Class ;
rdfs:label "From 90 years and older"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#BG>
a owl:Class ;
rdfs:label "Bulgaria"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#CH>
a owl:Class ;
rdfs:label "Switzerland"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#EE>
a owl:Class ;
rdfs:label "Estonia"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#ES>
a owl:Class ;
rdfs:label "Spain"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#GB>
a owl:Class ;
rdfs:label "United Kingdom"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#GR>
a owl:Class ;
rdfs:label "Greece"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#HR>
a owl:Class ;
rdfs:label "Croatia"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#IE>
a owl:Class ;
rdfs:label "Ireland"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#IS>
a owl:Class ;
rdfs:label "Iceland"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#LT>
a owl:Class ;
rdfs:label "Lithuania"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#LV>
a owl:Class ;
rdfs:label "Latvia"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#MK>
a owl:Class ;
rdfs:label "Macedonia"@en .
```

```

<http://lod.gesis.org/matchingstatistics/concepts/geo#SE>
a owl:Class ;
rdfs:label "Sweden"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#SI>
a owl:Class ;
rdfs:label "Slovenia"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#CZ>
a owl:Class ;
rdfs:label "Czech Republic"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#FI>
a owl:Class ;
rdfs:label "Finland"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#FR>
a owl:Class ;
rdfs:label "France"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#LU>
a owl:Class ;
rdfs:label "Luxembourg"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#NO>
a owl:Class ;
rdfs:label "Norway"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#PL>
a owl:Class ;
rdfs:label "Poland"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#RU>
a owl:Class ;
rdfs:label "Russian Federation"@en .

<http://lod.gesis.org/matchingstatistics/concepts/agegroup#0-9>
a owl:Class ;
rdfs:label "From 0 to 9 years"@en .

<http://lod.gesis.org/matchingstatistics/concepts/agegroup#60-69>
a owl:Class ;
rdfs:label "From 60 to 69 years"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#BA>
a owl:Class ;
rdfs:label "Bosnia and Herzegovina"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#BY>
a owl:Class ;
rdfs:label "Belarus"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#NL>
```

```
a owl:Class ;
rdfs:label "Netherlands"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#UA>
a owl:Class ;
rdfs:label "Ukraine"@en .

<http://lod.gesis.org/matchingstatistics/concepts/agegroup#10-19>
a owl:Class ;
rdfs:label "From 10 to 19 years"@en .

<http://lod.gesis.org/matchingstatistics/concepts/agegroup#70-79>
a owl:Class ;
rdfs:label "From 70 to 79 years"@en .

<http://lod.gesis.org/matchingstatistics/concepts/geo#IT>
a owl:Class ;
rdfs:label "Italy"@en .

<http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value9>
a owl:Class ;
rdfs:label "Not applicable"@en .

<http://lod.gesis.org/matchingstatistics/concepts/agegroup#30-39>
a owl:Class ;
rdfs:label "From 30 to 39 years"@en .

<http://lod.gesis.org/matchingstatistics/concepts/agegroup#50-59>
a owl:Class ;
rdfs:label "From 50 to 59 years"@en .

<http://lod.gesis.org/matchingstatistics/concepts/gender#sex-M>
a owl:Class ;
rdfs:label "Male"@en .

<http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#total>
a owl:Class ;
rdfs:label "Total"@en .

<http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value4>
a owl:Class ;
rdfs:label "Not working"@en .

<http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value8>
a owl:Class ;
rdfs:label "Not specified or unknown"@en .

<http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value9>
a owl:Class ;
rdfs:label "Not applicable"@en .

<http://lod.gesis.org/matchingstatistics/concepts/agegroup#20-29>
a owl:Class ;
rdfs:label "From 20 to 29 years"@en .
```

```

<http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#single>
a owl:Class ;
rdfs:label "Single"@en .

<http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#divorced>
a owl:Class ;
rdfs:label "Divorced"@en .

<http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#widowed>
a owl:Class ;
rdfs:label "Widowed"@en .

<http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value2>
a owl:Class ;
rdfs:label "Part time employment"@en .

<http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value3>
a owl:Class ;
rdfs:label "Unemployed"@en .

<http://lod.gesis.org/matchingstatistics/concepts/agegroup#40-49>
a owl:Class ;
rdfs:label "From 40 to 49 years"@en .

<http://lod.gesis.org/matchingstatistics/concepts/agegroup#80-89>
a owl:Class ;
rdfs:label "From 80 to 89 years"@en .

<http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value1>
a owl:Class ;
rdfs:label "Full-time employment"@en .

<http://lod.gesis.org/matchingstatistics/concepts/occupation#occup_value5>
a owl:Class ;
rdfs:label "Retired"@en .

<http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value8>
a owl:Class ;
rdfs:label "Dont know"@en .

<http://lod.gesis.org/matchingstatistics/concepts/gender#sex-U>
a owl:Class ;
rdfs:label "Not specified or unknown"@en .

<http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#separated>
a owl:Class ;
rdfs:label "Separated"@en .

<http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value1>
a owl:Class ;
rdfs:label "Very satisfied"@en .

<http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value2>
```

```
a owl:Class ;
rdfs:label "Somewhat satisfied"@en .

<http://lod.gesis.org/matchingstatistics/concepts/gender#sex-N>
a owl:Class ;
rdfs:label "Not applicable"@en .

<http://lod.gesis.org/matchingstatistics/concepts/gender#sex-T>
a owl:Class ;
rdfs:label "Total"@en .

<http://lod.gesis.org/matchingstatistics/concepts/maritalStatus#married>
a owl:Class ;
rdfs:label "Married"@en .

<http://lod.gesis.org/matchingstatistics/concepts/gender#sex-F>
a owl:Class ;
rdfs:label "Females"@en .

<http://lod.gesis.org/matchingstatistics/concepts/satisfaction#sat_value3>
a owl:Class ;
rdfs:label "Somewhat dissatisfied"@en .

STATTBOX:DataEntry a owl:Class .
```