FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Semantic Integration and Exploration of Statistical Data

DISSERTATION

zur Erlangung des akademischen Grades

Doktor der Technischen Wissenschaften

eingereicht von

Ba-Lam Do
Matrikelnummer 1229759

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: O.Univ.Prof. Dipl.-Ing. Dr. techn. A Min Tjoa
Zweitbetreuung: Mag.rer.soc.oec. Elmar Kiesling, PhD

Diese Dissertation haben begutachtet:

Prof. Dr. Maurizio Marchese	a.Univ.-Prof. Dr. Josef Küng

Wien, 24. April 2017

Ba-Lam Do

FAKULTÄT
FÜR !NFORMATIK
Faculty of Informatics

Semantic Integration and Exploration of Statistical Data

DISSERTATION

submitted in partial fulfillment of the requirements for the degree of

Doktor der Technischen Wissenschaften

by

Ba-Lam Do
Registration Number 1229759

to the Faculty of Informatics

at the TU Wien

Advisor: O.Univ.Prof. Dipl.-Ing. Dr. techn. A Min Tjoa
Second advisor: Mag.rer.soc.oec. Elmar Kiesling, PhD

The dissertation has been reviewed by:

Prof. Dr. Maurizio Marchese	a.Univ.-Prof. Dr. Josef Küng

Vienna, 24th April, 2017

Ba-Lam Do

Erklärung zur Verfassung der Arbeit

Ba-Lam Do
Radelmayergasse 16/3/2, 1190 Wien

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

Wien, 24. April 2017

Ba-Lam Do

Acknowledgements

I would like to send sincere thanks to my advisor Prof. A Min Tjoa. Over four years ago, I had an interview with Prof. Tjoa and subsequently gained a scholarship to conduct my research in Austria. This opened a new horizon for me in both academic and social fields. During my four-years of study at TU Wien, he has transmitted his passion in academic research to me and provided me with valuable advice to overcome challenges in research. Without his support, I would not have completed my study.

I want to express my thanks to Dr. Amin Anjomshooa, who enlightened me the first steps in research in nearly my first two years at TU Wien. My research proposal took shape in this period after many interesting discussions with him. I am so grateful to Dr. Elmar Kiesling, who supported me in two last years at TU Wien. Working with him is a great pleasure to me because of his knowledge and passion in research. His detailed comments not only supported me in finishing the study but also gave me valuable lessons. I also want to thank my colleagues at Linked Data lab including Tuan Dat Trinh, Peter Wetz, Peb Ruswono Aryan, and Fajar Juang Ekaputra. I will never forget their supports and memories that we had together.

My four years in Austria were not a lonely period thanks to the love and enormous encouragement from my parents, who always respected my choices and decisions. I would like to send my special thanks to my wife, Thi Huong Tran, who shared both the fun and sadness with me and encouraged me at each difficult time.

Finally, I would like to thank the Vietnam-Austria Scholarship Programme, which offered a grant for me to study at TU Wien. The offices of OeAD (Austria) and VIED (Vietnam) supported me wherever I had a trouble or a question. This support allowed me to overcome difficulties and to focus on my study.

Kurzfassung

Die Menge an verfügbaren statistischen Daten im Web ist in den letzten Jahren deutlich gestiegen. Viele Organisationen und öffentliche Institutionen veröffentlichen statistische Daten in einer Vielzahl von Formaten und Kodierungen, verwenden dabei verschiedene Skalen und stellen diese mittels unterschiedlicher Zugangsmechanismen zur Verfügung. Diese Inkonsistenzen erschweren die Analyse der dadurch entstehenden heterogenen und schwer zugänglichen Vielfalt an Daten.

Diese Arbeit behandelt drei grundlegende Problemstellungen der Integration und Exploration von ungleichen statistischen Datenquellen: (i) Datenheterogeneität; (ii) Verknüpfung von statistischen Datensets; und (iii) die Bereitstellung eines einheitlichen und integrierten Zugangs zu einzelnen Datensets. Dazu verwenden wir semantische Technologien, Standards, Services und Vokabularien. Wir nutzen das RDF Datenmodell und das Data Cube Vokabular, um die heterogene Repräsentation von Daten zu vereinheitlichen. Mittels der RDF Mapping Language werden die ursprünglichen Daten so in ein dem RDF Data Cube Vokabular entsprechendes RDF Modell konvertiert. Im nächsten Schritt definieren wir URI-Vorlagen, die verwendet werden, um eine Reihe von gemeinsam benutzten URIs zu erstellen. Diese URIs ermöglichen die Verknüpfung der zuvor erstellten RDF Modelle. Zu diesem Zwecke entwickeln wir Algorithmen, mit denen RDF Komponenten (zum Beispiel räumliche oder zeitliche Dimensionen) oder Werte über Datensets hinweg abgestimmt werden können. Um statistische Daten abfragen und integrieren zu können, erstellen wir ein einheitliches Modell der Metadaten-Beschreibungen von den Datensets. Diese Metadaten-Beschreibungen enthalten (i) die Detail-Struktur und die Zugangsmethode, um Abfragen zu erstellen, und (ii) die Verknüpfungen, die Komponenten und Werte der Datensets mit einem Satz von gemeinsam benutzten URIs verbinden. Diese wohldefinierten Metadaten stellen somit eine standardisierte konzeptionelle Ebene für jedes Datenset dar. Eine Mediator-Komponente, die auf einem Metadaten-Speicher basiert, ermöglicht die semantische Integration und einheitlichen Zugang zu verschiedenen heterogenen Datenquellen. Der Mediator kann generische Abfragen so transformieren, dass sie auf individuelle Datensets passen, Werte zwischen unterschiedlichen Skalen konvertieren und Resultate in ein konsolidiertes Ergebnis umschreiben.

Die präsentierte Methode ist in Form des Systems StatSpace umgesetzt. StatSpace ist ein Linked Statistical Data Space, der einheitlichen Zugang zu mehr als 1,800 Datensets bereitstellt. Diese Daten werden von verschiedenen Anbietern, wie zum Beispiel der

Weltbank, der Europäischen Union, oder der Europäischen Umweltagentur zur Verfügung gestellt. Wir evaluieren die Methode hinsichtlich Abdeckung, Validität und Performance

Abstract

In recent years, the amount of statistical data available on the web has grown dramatically. Numerous organizations and governments publish statistical data in a multitude of formats and encodings, using different scales, and providing access through a wide range of mechanisms. Due to such inconsistent data publishing practices, analysis of heterogeneous and dispersed statistical data is challenging.

This thesis addresses three major challenges to integrate and explore disparate statistical data sources, i.e.,: (i) data heterogeneity; (ii) interconnection between statistical data sets; and (iii) providing uniform and integrated access to individual data sets. To this end, we rely on semantic technologies, standards, services, and vocabularies. We use RDF data model and Data Cube vocabulary to consolidate heterogeneous data representations. Based on the RDF mapping language, raw data sets are lifted into RDF following the Data Cube vocabulary. To link URIs used in data sources, we define URI design patterns to coin a set of shared URIs. In addition, we develop algorithms to map components (e.g., spatial dimension, temporal dimension) and align values. In order to query and integrate individual data sets, we model statistical data sets in metadata descriptions in a uniform manner. Each metadata description contains information of (i) the detailed data structure and access method for query building, and (ii) link relationships that connect components and values used in the data set to a set of shared URIs. The well-defined metadata, hence, provides a standardized conceptual layer over each data set. Relying on the metadata repository, a mediator provides a semantic integration of and uniform access to multiple heterogeneous data sources. The mediator can transform generic queries into suitable queries for individual data sets, perform scale transformation, rewrite individual results, and integrate them into a consolidated result.

We implement this approach in StatSpace, a linked statistical data space that provides uniform access to more than 1,800 data sets published by a variety of data providers including the World Bank, the European Union, and the European Environment Agency. We evaluate our approach in terms of coverage, validity, and performance.

Contents

Kurzfassung **ix**

Abstract **xi**

Contents **xiii**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Problem Statement . . . 2
1.3 Research Questions . . . 3
1.4 Research Methodology . . . 3
1.5 Main Contributions . . . 4
1.6 Thesis Structure . . . 5
1.7 Publications . . . 6

2 Background **7**
2.1 Statistical Data . . . 7
2.2 Statistical Data and Metadata Exchange (SDMX) . . . 8
2.3 Resource Description Framework (RDF) . . . 9
2.4 RDF Data Cube Vocabulary . . . 10
2.5 Role of Standards in Statistical Data Integration and Exploration . . . 12
2.6 Existing Architectures for Data Integration and Exploration . . . 12
2.6.1 Warehousing and Virtual Integration Architectures . . . 12
2.6.2 Data Integration Architecture for Enterprise Information Systems 14
2.6.3 Data Integration Architecture for Multimedia Sources . . . 14
2.6.4 Statistical Data Exploration Architecture . . . 15

3 Architecture for Statistical Data Integration and Exploration **17**
3.1 Architecture . . . 17
3.2 RML Mapping Service . . . 19
3.3 URI Design Patterns . . . 19
3.4 Metadata Repository . . . 21
3.5 Metadata Generator . . . 24
3.6 Mediator . . . 25

3.7 Explorer . . . 26

4 Implementation **29**
4.1 Elements . . . 29
4.1.1 RML Mapping Service . . . 29
4.1.2 URI Design Patterns . . . 32
4.1.3 Metadata Generator . . . 36
4.1.4 Metadata Repository . . . 47
4.1.5 Mediator . . . 48
4.1.6 Explorer . . . 49
4.2 Example Use Cases . . . 53
4.2.1 Statistical Data Integration . . . 53
4.2.2 Data Quality Assessment . . . 56
4.2.3 Correlation Mining . . . 57
4.2.4 Spatial Data Visualization . . . 58

5 Evaluation **63**
5.1 RML Mapping Service . . . 63
5.2 URI Design Patterns . . . 65
5.3 Metadata Repository . . . 66
5.4 Metadata Generator . . . 67
5.4.1 Mapping Endorsement . . . 67
5.4.2 Limitation of Endorsement . . . 69

6 Related Work **71**
6.1 Mapping Languages . . . 71
6.2 Coreference Resolution in Linked Data context . . . 73
6.2.1 Link Discovery Frameworks . . . 73
6.2.2 Coreference Resolution Services . . . 74
6.3 Data Integration Research . . . 77
6.4 Data Exploration Research . . . 80

7 Conclusions and Future Work **83**
7.1 Summary . . . 83
7.2 Future Work . . . 84

Appendices **87**
A URI Design Patterns for Code Lists . . . 87
A.1 Reference Area Dimension (cl_area) . . . 87
A.2 Reference Period Dimension (cl_period) . . . 87
A.3 Age Dimension (cl_age) . . . 88
A.4 Education Level Dimension (cl_educationLev) . . . 89
A.5 Occupation Dimension (cl_occupation) . . . 89
A.6 Currency Dimension (cl_currency) . . . 90

A.7 Civil Status Dimension (cl_civilStatus) . . . 90
A.8 Frequency Dimension (cl_frequency) . . . 91
A.9 Sex Dimension (cl_sex) . . . 92
A.10 Economic Activity Dimension (cl_economicActivity) . . . 92
A.11 Expenditure Dimension . . . 93
A.12 Unit of Measure (cl_unitMeasure) . . . 94
A.13 Subject (cl_subject) . . . 94
B RDF mapping for World Bank Data . . . 95
C RML Mapping for Office for National Statistics (ONS) Data . . . 98
C.1 RML Mapping for the first Spreadsheet Data Set Collected . . . 98
C.2 RML Mapping for the second Spreadsheet Data Set Collected . . . 101
D RML Queries used in Evaluation . . . 106
D.1 RML Queries used to transform data of one country into RDF . . . 106
D.2 RML Queries used to transform data of all countries into RDF . . . 110
E Runtime of RML mapping service in Evaluation . . . 113
E.1 Time Consumption for Data Transformation of one Country . . . 113
E.2 Time Consumption for Data Transformation of all Countries . . . 115

List of Figures **117**

List of Tables **118**

List of Listings **120**

List of Algorithms **121**

List of Abbreviations **124**

Bibliography **125**

CHAPTER 1

Introduction

1.1 Motivation

In recent years, open data publishing practices have been widely adopted by numerous governmental and non-governmental organizations [1, 2, 3, 4, 5]. As a consequence, a large number of data sets have become available on the web. This proliferation of open data has created opportunities for research and has the potential to facilitate more informed citizen participation.

A considerable share of published data is statistical data [6, 7, 8, 9, 10]. This type of data has attracted great interest because of its diversity and value. It covers a wide range of domains such as finance, demographics, transportation, and employment and plays an increasingly important role in public policy formation and as a facilitator for informed decision-making in the private sector. Therefore, efficient data exploration is necessary to allow knowledge workers to make use of the fast growing amounts of statistical data that are available on the web through different access mechanisms and formats.

Currently, the majority of statistical data is published in raw formats like Comma Separated Values (CSV), JavaScript Object Notation (JSON) or in proprietary formats such as images and Microsoft Excel Spreadsheet (XLS)[1] [11, 12, 4, 13, 9, 14]. This choice is typically driven by existing workflows and motivated by its simplicity and low implementation cost [4]. However, to make use of the data, users need to download it entirely, extract subsets of interest, and manually reconcile data from scattered sources. Furthermore, they need to be proficient in the use of appropriate tools to work with various data formats. This is exacerbated by the heterogeneity in formats that make it difficult to analyze statistical data from multiple sources.

[1] e.g., `https://www.ons.gov.uk`, accessed December 30, 2016

Many data providers expose their data to developers and applications via APIs[2]. Statistical data is typically represented in Extensible Markup Language (XML) format using Statistical Data and Metadata eXchange (SDMX)[3], a standard sponsored by a consortium of seven major international institutions including the World Bank, the European Central Bank, and the United Nations. Although this approach provides a flexible access, the data exposed through APIs typically cannot be integrated automatically and therefore remains isolated and dispersed.

Some initial steps towards uniform publication of machine-readable statistical data have already been made. Many organizations, including the European Union[4] [15], the European Environment Agency[5], and Ireland's Central Statistics Office[6], have adopted RDF Data Cube Vocabulary (QB) [16] for statistical data publishing. Thereby, users can query data sets published by these organizations via their respective SPARQL Protocol and RDF Query Language (SPARQL) [17, 18] endpoints. However, each data provider typically coins its own Uniform Resource Identifiers (URIs) to represent entities, which results in the same entities being represented by different URIs in different data sources [8, 19]. Exploring related data sets and integrating data from multiple sources are therefore still difficult.

Due to diverse and inconsistent data publishing practices, integrating and exploring data across multiple data sets are still a challenging task. Our motivation in this thesis is to provide a uniform access to multiple statistical data sources and facilitate automated data integration.

1.2 Problem Statement

We identify three major challenges with respect to statistical data integration and exploration.

Data heterogeneity. The first challenge is to gather statistical data sets published in a wide range of formats such as CSV, JSON, XML, and RDF, and manage them uniformly. In recent years, RDF has become a widely-accepted standard to represent and exchange data [20, 21]. Therefore, a popular approach to address this issue is to use the QB vocabulary to transform raw statistical data into RDF format and then store data in SPARQL endpoints. Although this approach provides a useful way to gather and manage data sets, it is associated with two main issues. First, transformation into RDF increases the data volume, which can reduce the scalability and effectiveness of the approach, and secondly, we may not have the most current data when the original data sources are updated.

[2] e.g., `http://data.worldbank.org`, accessed December 30, 2016
[3] `http://sdmx.org`, accessed December 30, 2016
[4] `http://data.europa.eu/euodp/en/linked-data`, accessed December 30, 2016
[5] `http://semantic.eea.europa.eu/sparql`, accessed December 30, 2016
[6] `http://data.cso.ie/query.html`, accessed December 30, 2016

Interconnection between RDF data sets. To integrate data sets published by different providers, we need to identify equivalent entities used in these data sets. Approaches to deal with this issue by crawling *owl:sameAs* relationships in data sources exist [22, 23], but this approach cannot detect equivalent entities if the data sources lack *owl:sameAs* relationships, which is often the case. In addition, the generated coreference resolution services [22, 23] cover only a small number of sources. Another approach is to manually provide *owl:sameAs* relationships to integrate data from heterogeneous data sources [24, 25]. This approach, while helpful, is cumbersome and time consuming and therefore only applicable to a moderate amount of data sources.

Uniform and integrated access to multiple heterogeneous data sources. A large number of applications of statistical data integration and exploration [26, 27, 28, 29, 30, 31] have been introduced in recent years. These tools support users in exploring statistical data stored in SPARQL endpoints in various ways. To integrate data, these applications need to generate separate SPARQL queries for interlinked data sources. Therefore, integrating new data sources requires a considerable effort to identify equivalent relationships, create SPARQL queries, and integrate a set of results obtained from individual data sources. These applications are therefore limited not only to a single input format (i.e., RDF format), but also in the number of involved sources for data integration, which is frequently limited to two data sources. As a result, so far we have not had a single point of access where users can explore and integrate statistical data sets published on web sites, portals, and SPARQL endpoints, in a uniform manner.

1.3 Research Questions

The central problem addressed in this thesis is:
How can users be enabled to integrate and explore heterogeneous statistical data sources?

To address this problem, we need to propose suitable approaches to overcome the existing challenges of data heterogeneity, data linking, and data access. Therefore we break the research question into the following three sub-questions.

Research Question 1. *How can we address data heterogeneity in terms of formats?*

Research Question 2. *How can we establish interconnections between statistical data sets?*

Research Question 3. *How can we provide uniform and integrated access to individual data sets?*

1.4 Research Methodology

We make use of the design-science research methodology proposed by Peffers et al. [32]. Our approach consists of four main phases: (i) *Analysis*, (ii) *Design*, (iii) *Evaluation*, and (iv) *Communication*. In the *Analysis* phase, we undertake a review of literature on data

transformation, linking, integration, and exploration. From the knowledge obtained, we identify requirements and challenges, which will be addressed in the thesis. In the *Design* phase, we find answers to all the research questions and a prototype implementation will be introduced to users. In the next phase, the *Evaluation* phase, we will evaluate the implementation with respect to two aspects: (i) example use cases that our approach provides for users; and (ii) performance, data coverage, and the validity of the approach. Finally, the *Communication* phase runs simultaneously to the other phases. We will focus on publishing papers and writing the thesis. This will encourage participation in the research community, eventually resulting in a high quality thesis.

1.5 Main Contributions

In this thesis, we show that based on a coherent combination of available semantic technologies, standards, services, and vocabularies, we can successfully address research questions to benefit users. In the following, we describe our contributions with respect to the existing challenges.

A metadata repository for gathering and managing data sets. For each statistical data set, we use metadata to describe the information needed for querying and integrating the data set, i.e.,: (i) the data structure and access method for query building, and (ii) link relationships that connect components and values used in the data set to a set of shared URIs. The well-defined metadata, hence, provides a standardized conceptual layer for each data set. For statistical data sets published in raw formats, these respective metadata descriptions contain RDF Mapping Language (RML) [33] mappings to transform these data sets into RDF at query time. As a result, users can explore up-to-date data at runtime. Additionally, the mappings needed to execute the on-demand conversions require only very limited space, which makes the approach more scalable and easier to manage. Two supporting services to construct the *metadata repository* include a *metadata generator* and an *RML mapping service*.

URIs design patterns for data linking. We first reuse existing URIs defined by organizations and researchers. Next, we define a number of new patterns based on recommended values of SDMX and the Work Bank (WB), which enable and enhance their re-usability and reliability. In addition, we present a pattern to generate URIs for geographical areas because existing sources cover only a limited number of areas. We also develop relevant algorithms to match URIs used in data sets to their respective shared URIs. Finally, we introduce a coreference resolution service for statistical data sources and allow users to endorse the correctness of mappings at `statspace.linkedwidgets.org/sameas/`.

A mediator for uniform data querying. We develop a *mediator* that provides uniform and integrated access to all data sets described in the *metadata repository*. The *mediator* receives SPARQL queries that use shared URIs to represent components and filter values as the input. Based on metadata descriptions of data sets, the *mediator* can

rewrite the input query into specific queries for relevant data sets, rewrite the results obtained from individual sources, and then integrate them into a consolidated result.

An explorer for data integration and exploration. As a final contribution, we implement a web application named *explorer* to support non-expert users who are not familiar with semantic web technologies in exploring and integrating statistical data sets. Instead of formulating SPARQL queries, users can discover data through search, data visualization, and data integration. Similar to the *mediator*, this application needs to communicate with two other components, i.e., the *RML mapping service* and *metadata repository* to satisfy users' needs.

We have implemented our approach in this thesis in a linked statistical data space named StatSpace that provides uniform access to statistical data sets and facilitates automated data integration. StatSpace is available on the web at `http://statspace.linkedwidgets.org`. It currently provides access to more than 1,800 data sets published by a variety of data providers including the World Bank, the European Union, and the European Environment Agency.

1.6 Thesis Structure

The remainder of this thesis is organized as follows:

Chapter 2 presents background information on statistical data, standards regarding this kind of data, and existing architectures for data integration and exploration. We show that a combination of available standards can provide a sound foundation for statistical data integration and exploration.

Chapter 3 presents our approach for statistical data integration and exploration. We first introduce the StatSpace architecture. Next, we discuss the role of each its constituent element, including the *metada repository*, *URIs design patterns*, *mapping service*, *metadata generator*, *mediator*, and *explorer*.

Chapter 4 introduces the implementation of StatSpace architecture. Next, we illustrate the usefulness of StatSpace through four example use cases of data integration, data quality assessment, correlation mining, and spatial data visualization.

Chapter 5 evaluates StatSpace by means of performance, coverage of data sources, and mapping validity.

Chapter 6 discusses related work. We categorize these efforts into four groups, i.e., mapping definition languages, coreference resolution, data integration, and data exploration. We also provide comparisons between our approach and related work.

Chapter 7 summarizes the contributions of this thesis and provides an outlook on future research.

1.7 Publications

Contributions in this thesis have been published in the following peer reviewed papers.

- Ba-Lam Do, Peter Wetz, Elmar Kiesling, Peb Ruswono Aryan, Tuan-Dat Trinh, and A Min Tjoa. StatSpace: A unified platform for statistical data exploration. In On the Move to Meaningful Internet Systems. OTM2016, Confederated International Conferences: CoopIS, ODBASE, and C&TC. Springer, 2016. [34]
- Ba-Lam Do, Peb Ruswono Aryan, Tuan-Dat Trinh, Peter Wetz, Elmar Kiesling, and A. Min Tjoa. Toward a framework for statistical data integration. In Proceedings of the International Workshop on Semantic Statistics. CEUR-WS.org, 2015. [35]
- Ba-Lam Do, Tuan-Dat Trinh, Peb Ruswono Aryan, Peter Wetz, Elmar Kiesling, and A Min Tjoa. Toward a statistical data integration environment: the role of semantic metadata. In Proceedings of the International Conference on Semantic Systems, pages 25–32. ACM, 2015. [36]
- Ba-Lam Do, Tuan-Dat Trinh, Peter Wetz, Elmar Kiesling, Amin Anjomshoaa, and A Min Tjoa. Multiscale exploration of spatial statistical datasets: a linked data mashup approach. In Proceedings of the International Workshop on Semantic Statistics. CEUR-WS.org, 2014. [37]
- Ba-Lam Do, Peb Ruswono Aryan, Tuan-Dat Trinh, Peter Wetz, Elmar Kiesling, and A. Min Tjoa. Widget-based exploration of linked statistical data spaces. In Proceedings of International Conference on Data Management Technologies and Applications, pages 282-290. ACM, 2014. [38]

CHAPTER 2

Background

In this chapter, we introduce background knowledge related to our research. We first provide a brief introduction of statistical data and its benefits in Section 2.1. Next, in Sections 2.2, 2.3, and 2.4 we introduce three existing standards that have been widely used in organizations and governments to represent and publish their statistical data. We then discuss the roles of these standards for statistical data integration and exploration in Section 2.5. Finally, we present existing data integration and exploration architectures and highlight the lack of an architecture for statistical data integration and exploration in Section 2.6.

2.1 Statistical Data

Statistical data refers to data produced through investigation, observation, and experimentation to report overall trends, to identify risks and opportunities, and to conduct planning [39, 26]. Cyganiak et al. [16] characterize a statistical data set by: (i) a set of dimensions that qualify observations (e.g., time interval of the observation or geographical area that the observation covers), (ii) a set of measures that describe the objects of the observation (e.g., population or annual percentage change), and (iii) attributes that facilitate interpretation of the observed values (e.g., units of measure or scaling factors). Statistical data exploration and integration can potentially provide many benefits to users:

(i) Users can complement incomplete data and obtain a more comprehensive view. For instance, using data published by World Bank, users can obtain economic indicators – such as *GDP per capita* of a country in the period from 1890 to 2015. This data source, however, does not include forecasts for the following years, which can, however, be found in other data sources such as the International Monetary Fund (IMF). Furthermore, the relationship between this indicator and other indicators

such as *Inflation* from the World Bank or the *Corruption Perceptions Index* from Transparency International may be of interest. Therefore, an appropriate and meaningful combination of related data sets from a single (e.g., the World Bank) or multiple sources (e.g., the World Bank, IMF, Transparency International) would allow end users to explore data in a broader and more detailed manner, which a single data source cannot provide.

(ii) Users can make comparisons between related data sets and hence they can obtain sound foundations for decision-making. For example, the Vienna Open Data portal[1] provides various public transport data sets such as the annual number of passengers and offered seats for bus, tram, and metro vehicles. By comparing these datasets, policy-makers can identify appropriate investment for each type of vehicle in upcoming years.

2.2 Statistical Data and Metadata Exchange (SDMX)

Exchange and sharing of statistical data are frequent tasks of major organizations, particularly between international statistical organizations and their member countries. This leads to a pressing need to establish a standardized common agreement how to describe data in order to enable more efficient data exchange and exploration. To this end, SDMX was launched in 2001 by seven institutions including the World Bank, the Statistical Office of the European Union, the United Nations Statistical Division, the European Central Bank (ECB), the International Monetary Fund, the Organization for Economic Cooperation and Development, and the Bank for International Settlements[2]. SDMX has become an international standard[3] approved by the International Organization for Standardization and is used as a preferred standard in the global statistical community [40].

SDMX consists of three main components, i.e., technical standards (such as validation and transformation languages, tutorials), statistical guidelines (such as cross-domain concepts, code lists), and related architectures and tools. In SDMX, statistical data can be represented in either GESMES/TS proprietary format (named SDMX-EDI) or XML open format (named SDMX-XML). Although a transformation between these formats can be easily done through the use of publicly available tools, the main SDMX format is XML [41].

To publish a statistical data set following SDMX-XML, data publishers first have to create a *Data Structure Definition* file to specify dimensions, measures, and attributes used in the data set. In addition, this description file also needs to identify possible values (i.e., code lists) for dimensions and attributes listed in the data set. This component therefore facilitates a better understanding by users and allows the generation of automatic data

[1] `https://open.wien.gv.at/site/open-data/`, accessed December 30, 2016
[2] `https://sdmx.org/?page_id=3425`, accessed December 30, 2016
[3] `http://sdmx.org/?p=1215`, accessed December 30, 2016

exploration applications. In the next step, the providers need to create a data file to describe real values. In this file, they can group data at different levels, such as observation, series, group, and data set levels [42]. Table 2.1 and Listing 2.1 provide an overview of the *Data Structure Definition* and data files for the publishing of the *euro foreign exchange reference rate* [42] of the ECB in XML format.

Dimensions			
Type	**Concept**	**Representation**	**Description**
Dimension 1	FREQ	CL_FREQ	Frequency of observations
Dimension 2	CURRENCY	CL_CURRENCY	The currency whose value is being measured against the base currency
Dimension 3	CURRENCY_DENOM	CL_CURRENCY	The base currency
Dimension 4	EXR_TYPE	CL_EXR_TYPE	The exchange rate type (e.g., spot).
Dimension 5	EXR_SUFFIX	CL_EXR_SUFFIX	Exchange rate series variation
Dimension 6	TIME_PERIOD	Time Point Set	Frequency of observations
Measure			
OBS_VALUE	The measured value		
Attributes			
Concept	**Assignment Level**	**Representation**	**Description**
OBS_STATUS (mandatory)	Observation	CL_OBS_STATUS	The observation status e.g., normal, estimated
OBS_CONF (optional)	Observation	CL_OBS_CONF	The observation confidentiality
TIME_FORMAT (mandatory)	Series	Time Duration Set	ISO 8601compliant way to describe duration
COLLECTION (mandatory)	Series	CL_COLLECTION	When the information was collected
UNIT (mandatory)	Group	CL_UNIT	The unit used
UNIT_MULT (mandatory)	Group	CL_UNIT_MULT	Whether the data is in millions, billions, etc.
DECIMALS (mandatory)	Group	CL_DECIMALS	The number of decimal places
TITLE_COMPL (mandatory)	Group	Up to 1050 characters	A human-readable title describing a certain group

Table 2.1: An example of data structure definition file (excerpt) [42]

2.3 Resource Description Framework (RDF)

The Resource Description Framework (RDF) is data model for describing resources on the web. It is designed to allow computers to understand data representation. The RDF

```
<DataSet xmlns="http://www.ecb.int/vocabulary/stats/exr/1" xsi:schemaLocation="http://
    www.ecb.int/vocabulary/stats/exr/1␣ecb_exr1_compact.xsd" datasetID="ECB_EXR1">
<Group CURRENCY="AUD" CURRENCY_DENOM="EUR" EXR_TYPE="SP00" EXR_SUFFIX="A" DECIMALS="4"
    UNIT="AUD" UNIT_MULT="0" TITLE_COMPL="ECB␣reference␣exchange␣rate,␣Australian␣
    dollar/Euro,␣2:15␣pm␣(C.E.T.)" />
<Series FREQ="D" CURRENCY="AUD" CURRENCY_DENOM="EUR" EXR_TYPE="SP00" EXR_SUFFIX="A"
    TIME_FORMAT="P1D" COLLECTION="A">
<Obs TIME_PERIOD="1999-01-04" OBS_VALUE="1.9100" OBS_STATUS="A" OBS_CONF="F" />
<Obs TIME_PERIOD="1999-01-05" OBS_VALUE="1.8944" OBS_STATUS="A" OBS_CONF="F" />
<Obs TIME_PERIOD="1999-01-06" OBS_VALUE="1.8820" OBS_STATUS="A" OBS_CONF="F" />
<Obs TIME_PERIOD="1999-01-07" OBS_VALUE="1.8474" OBS_STATUS="A" OBS_CONF="F" />
<Obs TIME_PERIOD="1999-01-08" OBS_VALUE="1.8406" OBS_STATUS="A" OBS_CONF="F" />
</Series>
</DataSet>
```

Listing 2.1: Reference exchange rate of the ECB (excerpt) [42]

data model makes use of URIs to name resources and describes data through a set of triples. Each triple consists of three components, i.e., a subject, predicate, and object in the form of *<subject, predicate, object>* expression. A triple, for instance, *<ex:Bob, rdf:type, ex:Employee>* describes that Bob, who is identified by the URI *ex:Bob*, is an instance of the Employee class.

2.4 RDF Data Cube Vocabulary

The RDF Data Cube vocabulary (QB) [16] is a W3C recommendation for publishing statistical data on the web. This vocabulary builds upon the SDMX version 2.0, hence, it can be considered a further development of SDMX to facilitate more efficient data exploration and integration. Figure 2.1 provides an overview of this vocabulary.[4] The QB focuses on describing relationships between elements in a statistical data set including data set (class qb:DataSet), observation (class qb:Observation), data structure definition (class qb:DataStructureDefinition), dimension (class qb:DimensionProperty), measure (class qb:MeasureProperty), attribute (class qb:AttributeProperty), group of observations (class qb:ObservationGroup), slice (qb:Slice), and values of each component (class qb:CodedProperty).

To illustrate this vocabulary, by means of an example data set which shows the life expectancy from regions in Wales, we represent it in RDF format. Table 2.2 shows that this data set can be represented by three dimensions (i.e., *time period*, *region*, and *sex*), together with one measure (i.e., life expectancy), and one attribute (i.e., number of years in this case). Figure 2.2 shows an excerpt of representation of the data set according to the QB vocabulary.

[4] All prefixes used in this paper can be looked up at `http://prefix.cc`

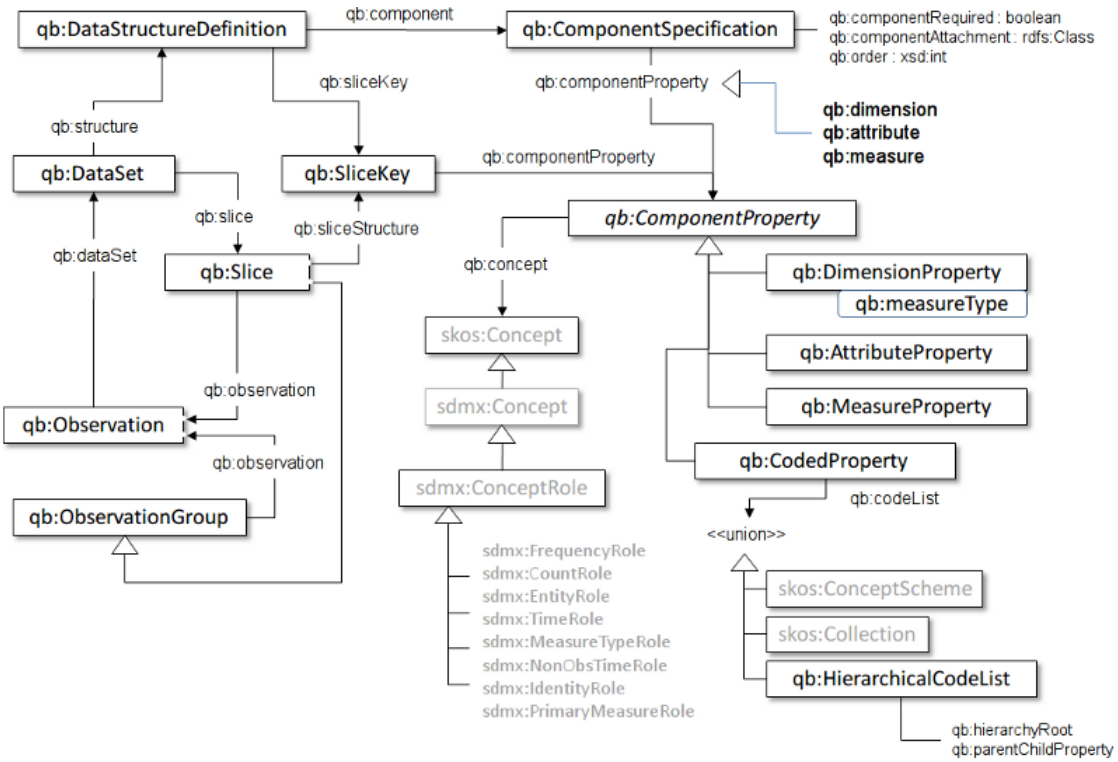


Figure 2.1: Overview of the QB [16]

	2004–2006		2005–2007		2006–2008	
	Male	Female	Male	Female	Male	Female
Newport	76.7	80.7	77.1	80.9	77.0	81.5
Cardiff	78.7	83.3	78.6	83.7	78.7	83.475
Merthyr Tydfil	75.5	79.1	75.5	79.4	74.9	79.6

Table 2.2: Example data set of the life expectancy (excerpt) [16]

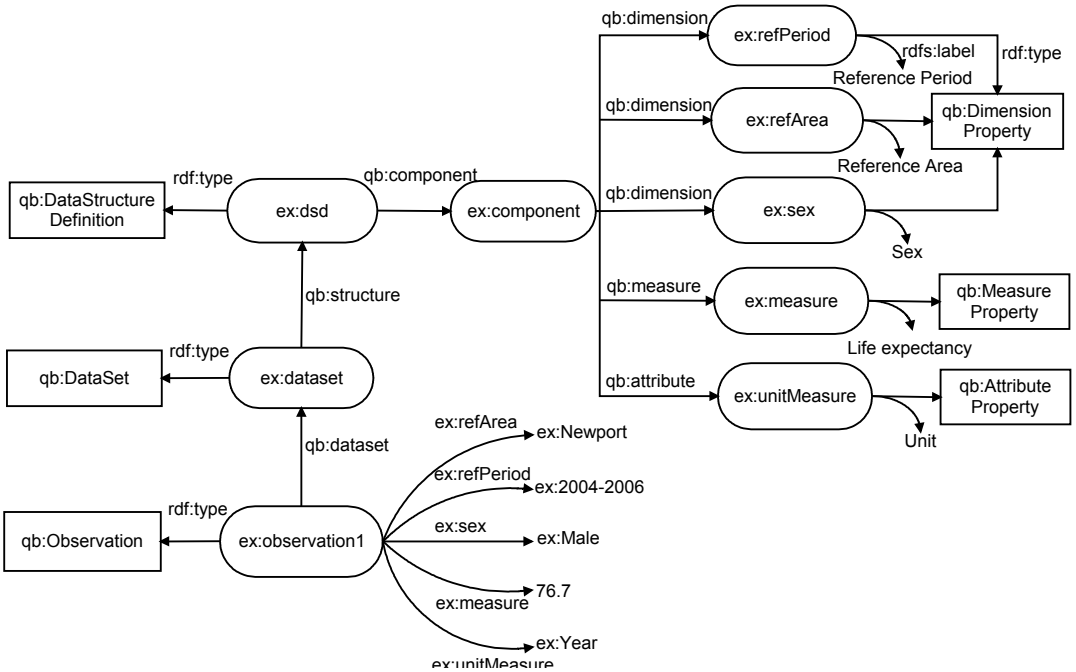


Figure 2.2: A representation according to the QB (excerpt)

2.5 Role of Standards in Statistical Data Integration and Exploration

Data integration is the problem of combining data from multiple heterogeneous sources and providing users with a unified view of data [43, 44, 45]. *Data exploration* is an important step in data analysis that aims to provide a better understanding of data for users [46, 47, 48]. This task involves a wide range of activities such as visualizing data, discovering correlation between variables, identifying invalid values, etc. To facilitate automated data integration and exploration, the use of existing standards including SDMX, RDF, and the QB plays an important role because of the following reasons.

First, SDMX defines a set of cross-domain concepts and code lists to represent concepts and entities used in a statistical data set. These identifiers allow data providers to share the same understanding, which facilitates data integration.

Second, the QB vocabulary provides a standardized representation for statistical data in RDF format. The use of this vocabulary facilitates the creation of applications of data visualization, correlation discovery, etc. However, it does not define a set of shared URIs to link concepts and entities used in heterogeneous data sources. To provide a sound foundation for statistical data integration and exploration, we need to combine existing standards to explicitly capture structure, linking, and semantics of data.

2.6 Existing Architectures for Data Integration and Exploration

Two recent papers [49, 50], published in 2016 and 2017, show a gap in the research of statistical data integration. Our aim in this thesis is to support users in statistical data integration and exploration. Therefore, in this section we first discuss a generic architecture of data integration systems. Next, we present two architectures designed for enterprise information systems and multimedia data sources. We also introduce an architecture for statistical data exploration. Based on these architectures, we show the essential requirements of an architecture for statistical data integration and exploration.

2.6.1 Warehousing and Virtual Integration Architectures

Most data integration systems can be classified into two approaches [51], i.e., (i) *warehousing:* data from individual data sources is loaded and stored in a physical data store called a *warehouse* and (ii) *virtual integration:* data from each source is accessed directly. Figure 2.3 depicts a generic architecture for data integration systems. We first describe four main components in the *virtual integration* approach [52, 51], i.e., *data sources*, *wrappers*, *a mediated schema*, and *source descriptions*.

Data sources refer to input objects that data integration systems need to merge to provide a unified view of data for users. They can differ in size, format, as well as access

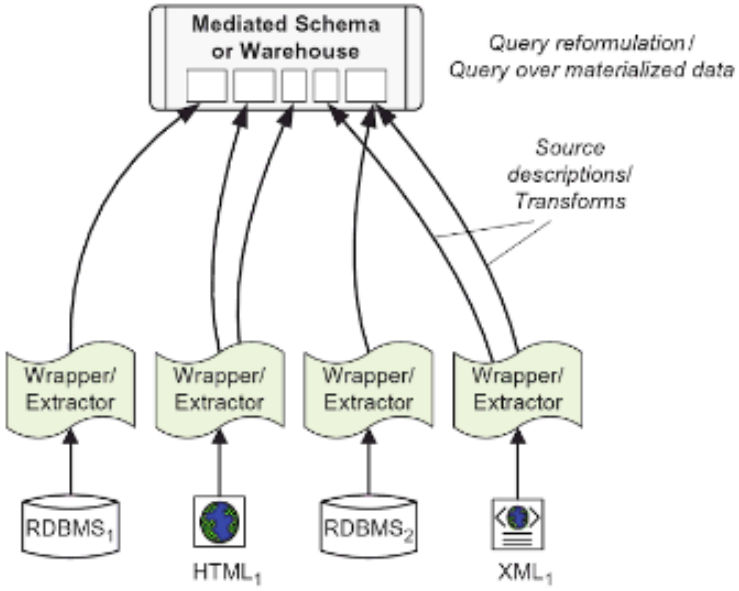


Figure 2.3: A Generic Data Integration Architecture [51]

mechanism. For example, the input of a data integration system can contain relational databases, XML databases, HTML files, etc.

Wrappers are programs whose role is to connect the data sources with requests from users and external applications. Each wrapper is specifically tailored for a data source that allows it to send queries to the data source, receive answers, and perform data transformation on the results.

A mediated schema is a logical schema designed to allow users send queries to data sources. This schema contains only properties which are relevant to applications/users.

Source descriptions contain necessary information about individual data sources that systems need in order to integrate data from these sources. The key component in each *source description* is a *semantic mapping* that provides the relationship between the *mediated schema* and the schema of data source. As a result, a query using terms and relations in the *mediated schema* can be rewritten into suitable queries for individual data sources. Figure 2.3 also shows that data integration systems require *semantic mappings* only between data sources and the *mediated schema*. Therefore, the number of *semantic mappings* is equal to the number of data sources.

In the *data warehousing* approach [53, 51], the *mediated schema* is replaced by a physical schema, i.e., the *warehouse schema*. In addition, instead of using *wrappers*, data integration systems typically make use of Extract Transform Load (ETL) tools to consolidate data from heterogeneous data sources into a common schema. These data transformations, hence, play a role as *semantic mappings* in the *virtual integration* approach.

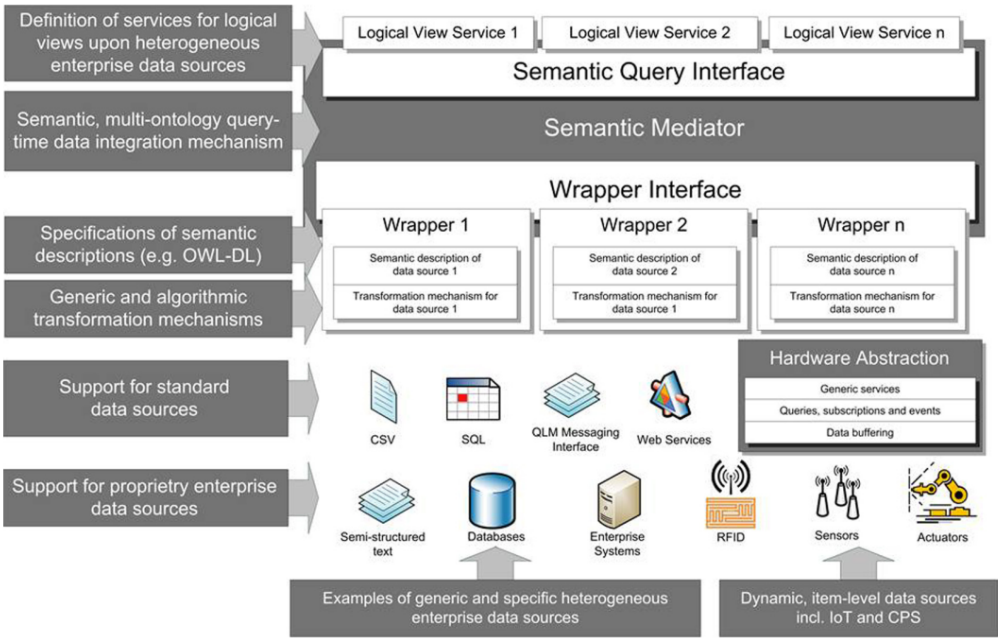


Figure 2.4: A Data Integration Architecture for Enterprise Information Systems [54]

2.6.2 Data Integration Architecture for Enterprise Information Systems

Figure 2.4 depicts an architecture [54] designed to integrate data in enterprise information systems. This architecture was introduced by IFIP Working Group 5.7 on Advances in Production Management[5]. The figure shows a diversity of input data sources including open data sources and proprietary data sources such as web services, databases, RFID data[6], sensor data, etc. Compared to the generic architecture, there are two different points in this architecture, i.e., (i) each *semantic description* of a data source is placed in a *wrapper*; (ii) the *mediated schema* component is placed in a *semantic mediator* that plays a role as a single point of access for users. It receives queries sent by users and translates them to appropriated queries for individual data sources.

2.6.3 Data Integration Architecture for Multimedia Sources

Figure 2.5 shows another data integration architecture [55] tailored to integrate multimedia sources. The authors develop several *wrappers* for relational databases such as Structured Query Language (SQL) database, ORACLE, etc. In addition, there is a special *wrapper* designed to communicate with *MILOS* [56] - a XML database storing and managing multimedia data. In this architecture, the *mediator* performs two main tasks: (i) generates the *mediated schema* (or global schema) and (ii) manages queries sent by users.

[5] `http://www.ifipwg57.org/`, accessed December 30, 2016
[6] `http://www.idautomation.com/barcode-faq/rfid/`, accessed December 30, 2016

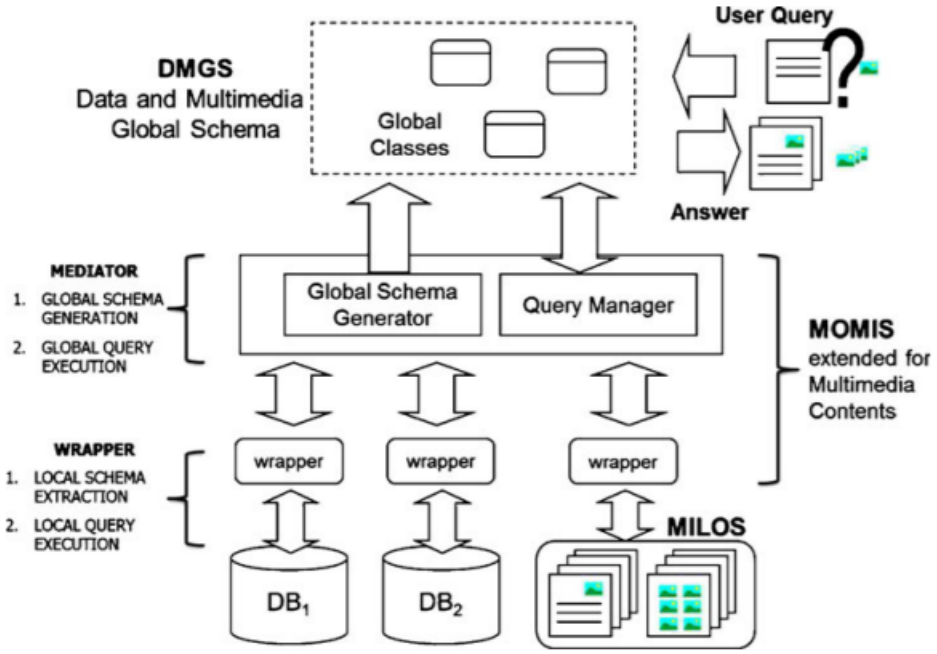


Figure 2.5: A Data Integration Architecture for Multimedia Sources [55]

2.6.4 Statistical Data Exploration Architecture

Salas et al. [57, 58] present a mediation architecture for exploring and publishing statistical data stored in relational databases but their tables can be mapped to concepts in the QB vocabulary. For example, a database contains several tables where one table stores observations (i.e., fact table [59]) and the remaining tables store dimensions and attributes. Although the aim of the authors does not relate to data integration, the architecture depicted in Figure 2.6 has many features of a data integration architecture. Therefore, we review this architecture and point out its three salient characteristics.

- First, the authors construct a *linked data cube description* to model each statistical data source. Each description that is represented in RDF format stores the information needed for accessing each data source (such as database name, host, port, etc.) and triples for describing dimensions and attributes. Compared to the *generic architecture*, *linked data cube descriptions* are similar to *source descriptions*. However, the descriptions in this architecture focus on exploring and locating data.
- Second, this architecture uses a *catalogue* to store all *linked data cube descriptions*. As a result, a *client application* can search and choose suitable data sources before sending requests to the *mediator* to obtain data (Figure 2.7).
- Third, the *mediator* accesses the *catalogue* to identify (i) suitable descriptions with the input keyword and (ii) connection information to a specific data source. Next,

when the *client application* requests data from a database, the *mediator* invokes the relevant *wrapper* to query the database. When receiving data from the *wrapper*, it lifts data into RDF format and returns to the application.

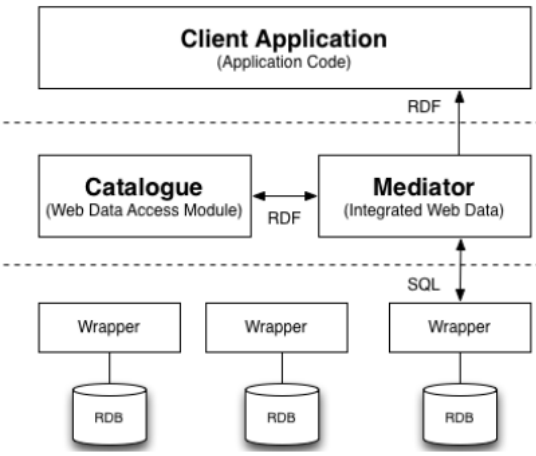


Figure 2.6: An Architecture for Exploring Statistical Data [57, 58]

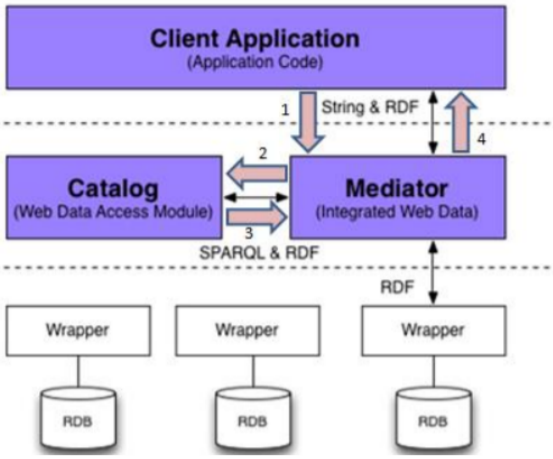


Figure 2.7: Communication between components in search scenario [57]

To sum up, existing architectures show requirements to construct an architecture for statistical data integration and exploration. First, such an architecture should provide uniform access to multiple heterogeneous data sources including relational databases, SPARQL endpoints, spreadsheets, CSV files, etc. Second, each statistical data set should be modelled in a description that includes the characteristics of this data set and necessary information for access. Finally, a collection of descriptions stored in a catalogue allows users to quickly search suitable data sets for their needs.

CHAPTER 3

Architecture for Statistical Data Integration and Exploration

In this chapter, we present our approach for statistical data integration and exploration. We first introduce an architecture in Section 3.1. Next, we describe the role of components in this architecture including the *RML mapping service* (Section 3.2), *URI design patterns* (Section 3.3), *metadata repository* (Section 3.4), *metadata generator* (Section 3.5), *mediator* (Section 3.6), and *explorer* (Section 3.7).

3.1 Architecture

Figure 3.1 introduces our architecture for statistical data integration and exploration and the relationships between its components. At its core, the *metadata repository* contains metadata descriptions of data sets that provide the information needed to query and link individual data sets. Two supporting services are necessary to populate this repository: (i) an *RML mapping service* that converts data from an original raw format into RDF, and (ii) a *metadata generator* that analyzes data sources and uses *URI design patterns* to generate the corresponding metadata. Finally, the *mediator* service and *explorer* application provide uniform access to all data sets in the *metadata repository*. There are three major processes connecting these resources:

Metadata repository building. We use the *metadata generator* to create metadata descriptions for each data set and store them in the *metadata repository*. For raw data sets, we use an *RML mapping service* – a processor of the RDF mapping language (RML) [33] – to transform statistical data from non-RDF formats into RDF following the QB vocabulary before analyzing them. For RDF data sets (that either have been published in SPARQL endpoints or are converted data sets), we use our data source analysis algorithm and mapping algorithms (cf. Section 4.1.3) to generate metadata based on the identified structure and contents of the data set.

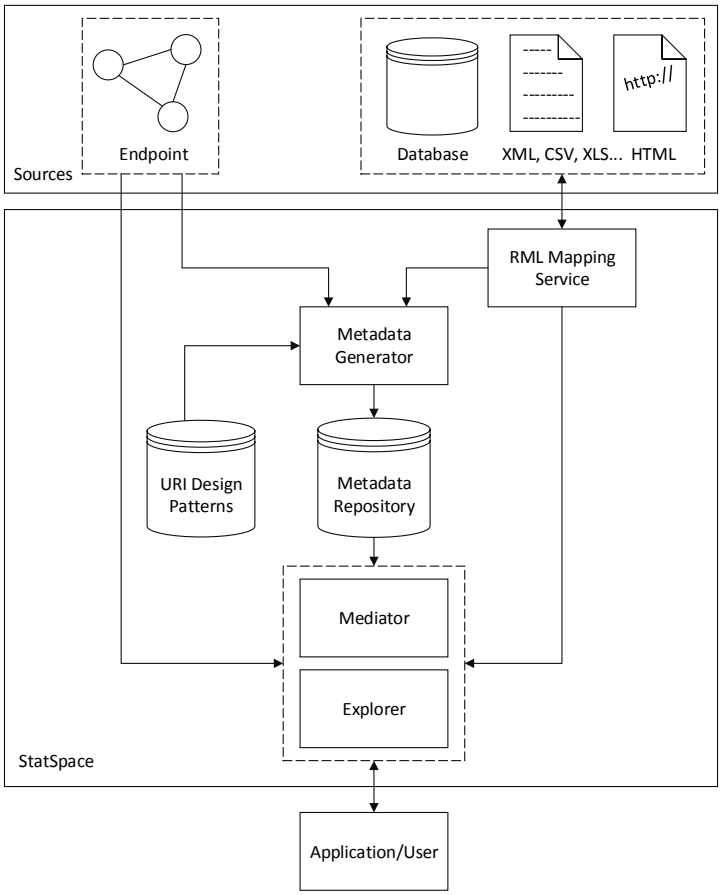


Figure 3.1: Architecture overview

Cross-data set SPARQL querying. The *mediator* provides a single point of access that allows advanced users to query data sets in the *metadata repository* and integrate the results into a consolidated representation. It accepts SPARQL queries that use uniform URIs to refer to data set components and filter values as input. When executing a query, the mediator makes use of the *metadata repository* to first identify relevant data sets and then rewrite the query for each of the identified data sources. Next, it invokes the *RML mapping service* and/or queries the respective SPARQL endpoints using the rewritten queries to obtain the results from the original sources. Finally, the *mediator* uses the *metadata repository* to rewrite the result set obtained from each source and then integrates it into a consolidated result before returning it to the user.

Web-based data set exploration. The *explorer* is a web application designed to support non-expert users who are not familiar with semantic web technologies in exploring statistical data sets. Users can visualize each data set as well as integrate multiple data sets through provided functionalities. Based on users' requirements, the *explorer* will query the *metadata repository* to identify necessary information for access to relevant

data set(s). Next, to obtain the data, the *explorer* sends request(s) to the *RML mapping service* and/or the respective SPARQL endpoint(s). Finally, the *explorer* visualizes the result to users.

3.2 RML Mapping Service

Currently, a large number of statistical data sets are published in raw formats like CSV, XML, JSON, Spreadsheet, etc. In order to establish a uniform access mechanism for all data sources, we transform raw data sets into RDF format. As a result, we can utilize SPARQL queries to access, explore, and analyze data. In our approach, the data transformation process relies on mappings that describe the way for lifting an input data set into RDF format. This approach allows us to execute on-demand conversions and provide up-to-date data to users. In addition, mappings require only very limited space for storage, which makes the approach more scalable and easier to manage.

In our architecture, we make use of RDF mapping language (RML) [33] – a recently-developed language for transforming raw data into RDF format. This language builds upon R2RML [60] – a W3C recommendation for transforming relational databases into RDF. RML not only follows the approach and descriptions of R2RML, but also extends its capability to an arbitrary format.[1]

The on-demand transformation at query time negatively impacts query answering time. To reduce the impact of this effect, we cache RDF data sets that were recently transformed. The caching behaviour is controlled via a URL parameter `cache=[yes|no]`. If the cache is enabled (default behaviour), we reuse the cached RDF data sets for relevant processes, such as data integration and visualization. These RDF data sets are automatically updated every three months or when a user requests the transformation of a new RDF data set. If the cache is disabled, the RDF data set is generated at query time.

3.3 URI Design Patterns

To integrate two arbitrary data sets, we need to identify equivalent URIs used in these data sets. To this end, we identified a set of shared URIs that can be matched to URIs used in existing data sources. This URIs set, hence, plays a role as a central reference point for data linking. Our approach is motivated by the role of DBpedia [61] which is a nuclear in the Linked Open cloud [62]. We organize URIs into the following three sub-groups: (i) *URIs representing components* such as spatial or temporal dimensions, (ii) *URIs representing values of components* (i.e., code lists) such as URIs for geographical areas or URIs for intervals, and (iii) *URIs representing subjects of data sets* such as GDP or Population. To enhance the acceptance and reusability of these URIs, we considered reusing existing URIs and defining a number of new patterns to generate the lacking URIs. To this end, we first reviewed the existing standard, i.e., SDXM and discovered how

[1] A detailed comparison of mapping languages which we reviewed, is provided in Section 6.1

major organizations such as the World Bank and the European Commission represent components and values in their data sets. Based on the usage of URIs in data sources, we then defined a set of shared URIs for data linking:

- First, SDMX defines a list of cross-domain concepts[2] in text format to represent components in a statistical data set. This list contains concepts of many dimensions (e.g., reference area – REF_AREA, education level – EDUCATION_LEVEL), one measure (OBS_VALUE), and one unit of measure (UNIT_MEASURE). Furthermore, it also presents code lists in raw format to describe values of many dimensions including economic activity, age, civil status, expenditure, currency, frequency, reference area, occupation, and sex.[3]

- Second, we reviewed the World Bank – one of the largest statistical data organizations in the world – with regard to its data encoding. At present, the World Bank provides more than 1,400 statistical data sets on various subjects such as agriculture, climate change, health, etc. To describe data sets semantically, it defines a list of subjects for data sets and also provides the relationship between a specific subject and its sub-subjects. In addition, the World Bank defines units that can appear in a statistical data set such as currencies (Euro, Dollar), metric units (km, meter), etc.

- Finally, data providers typically define particular URIs to represent their data in RDF format. As a result, the same entity can be represented by different URIs in different sources. However, many organizations such as the European Union Open Data Portal (EUODP), the European Environment Agency (EEA), the Central Statistics Office of Ireland (CSO) share a large number of common URIs. These URIs are defined by the United Kingdom (UK)'s time reference service[4] and the *Publishing statistical linked data* workshop[5]. While the former set provides a code list for temporal dimension, the latter set contains four name spaces, i.e., *sdmxd, sdmxm, sdmxa*, and *sdmxcode* which are semantic versions of dimensions, measures, attributes, and code lists defined in SDMX version 2009. For instance, the concept "reference area" in SDMX is represented by `http://purl.org/linked-data/sdmx/2009/dimension#refArea` in *sdmxd* namespace. Table 3.1 summarizes the components and code lists reused in our approach.

[2] `https://sdmx.org/wp-content/uploads/SDMX_Glossary_Version_1_0_February_2016.docx`, accessed December 30, 2016

[3] `https://sdmx.org/?page_id=3215`, accessed December 30, 2016

[4] The base URI is `http://reference.data.gov.uk/`, accessed December 30, 2016

[5] `https://code.google.com/archive/p/publishing-statistical-data/`, accessed December 30, 2016

	Components			Code lists			
	Dimension	Measure	Attribute	Unit	Subject	Temporal value	Other code lists
Source	SDMX	SDMX	SDMX	WB	WB	UK	SDMX
Orginal format	Raw	Raw	Raw	Raw	URI	URI	Raw
Available as URIs	sdmxd	sdmxm	sdmxa	No	Yes	Yes	Partly in sdmxcode

Table 3.1: Summary of components and code lists reused

A statistical data set can contain multiple measures whereas SDMX contains only one concept of measure (i.e., observed value). This leads to two approaches for data publication into RDF format as follows:

- Data providers can define local URIs or they may reuse URIs available in *sdmxm* to represent measures. Currently, in *sdmxm*, there are seven URIs referring to six specific measures, i.e., *sdmxm:age, sdmxm:civilStatus, sdmxm:currency, sdmxm:educationLev, sdmxm:occupation, sdmxm:sex*, and one generic measure, i.e., *sdmxm:obsValue*. However, the number of these concepts is too limited when compared to the number of existing raw data sets. For example, the World Bank has more than 1,400 different raw data sets. Therefore, to transform these data sets into RDF format, for this approach we would need to use more than 1,400 different URIs to represent measures. A large number of new URIs would need to be defined, such as URIs referring to GDP, population, health spending, etc.

- Any multi-measure data set can be split into multiple single-measure data sets [16], hence, we can assign a corresponding subject to each single-measure data set and use the existing generic measure in *sdmx-m*, i.e., *sdmxm:obsValue* to represent measures in all statistical data sets. Compared to the first approach, this approach allows us to reuse the list of subjects defined by the World Bank and does not require to define new URIs. Therefore, we make use of the second approach.

3.4 Metadata Repository

In SDMX-XML, each statistical data set is attached with a description file in order to provide information regarding the structure of data set and the values of each component. These description files support developers in creating applications for data integration and exploration. Therefore, we generate a metadata description for each statistical data set. Compared to description file in SDMX, our metadata description has differences as follows:

- Due to the variety of formats and access methods of statistical data sets, the metadata description contains information not only about the data structure, but also about the access mechanism to each data set in its original format.
- Due to the inconsistency in using URIs, we need to map URIs used in data sources to a set of shared URIs. In SDMX, because organizations use standardized concepts and values to represent data, the alignment of entities is not necessary.
- We assign a specific subject (e.g., GDP, population) for each data set. This information may be used to compare data sets published by different providers but they refer the same subject. In addition, based on the relationship between subjects, it allows users to identify related data sets to a given data set.

Figure 3.2 illustrates the generic structure of the metadata. It presents a standardized model to represent the data structure and access mechanism of an arbitrary statistical data set. The root node in this tree represents the metadata of a statistical data set. It is linked to URIs of the data set, components in the data structure, and their values. We reuse predicates in existing vocabularies including *QB, dc, void, dcat, owl, rdf,* and *rdfs* to represent relationships in metadata. An example of a metadata description is illustrated in Figure 3.3. Each metadata description consists of four main parts:

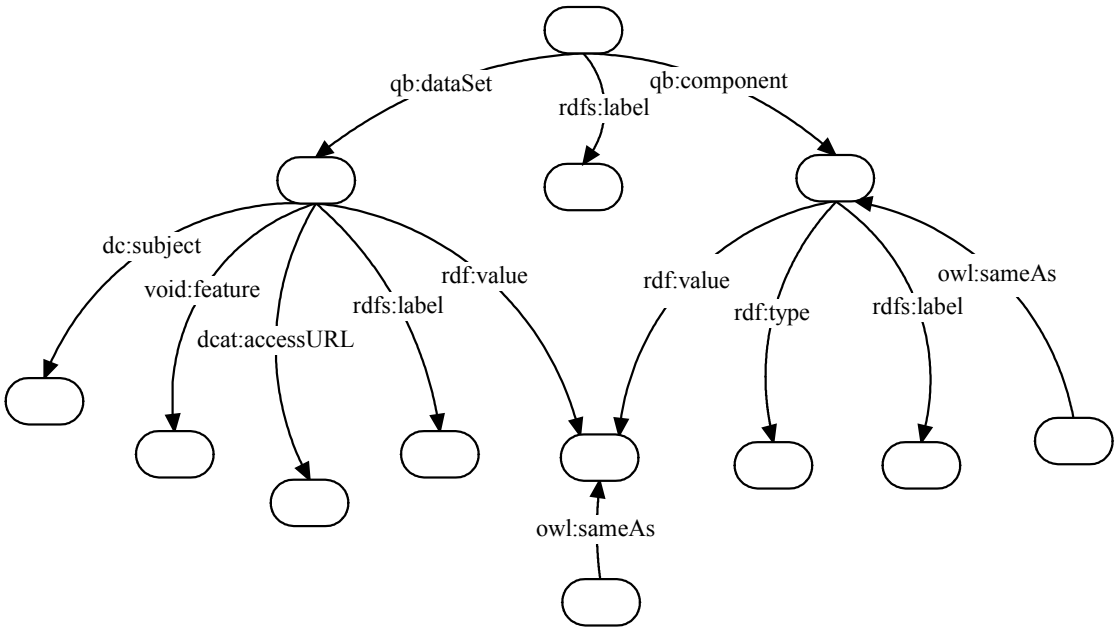


Figure 3.2: Structure of metadata

(i) *Descriptions of the data set, which focus on subject and access mechanism.* We use *qb:dataSet* to link URI of the metadata (the root node) to a corresponding

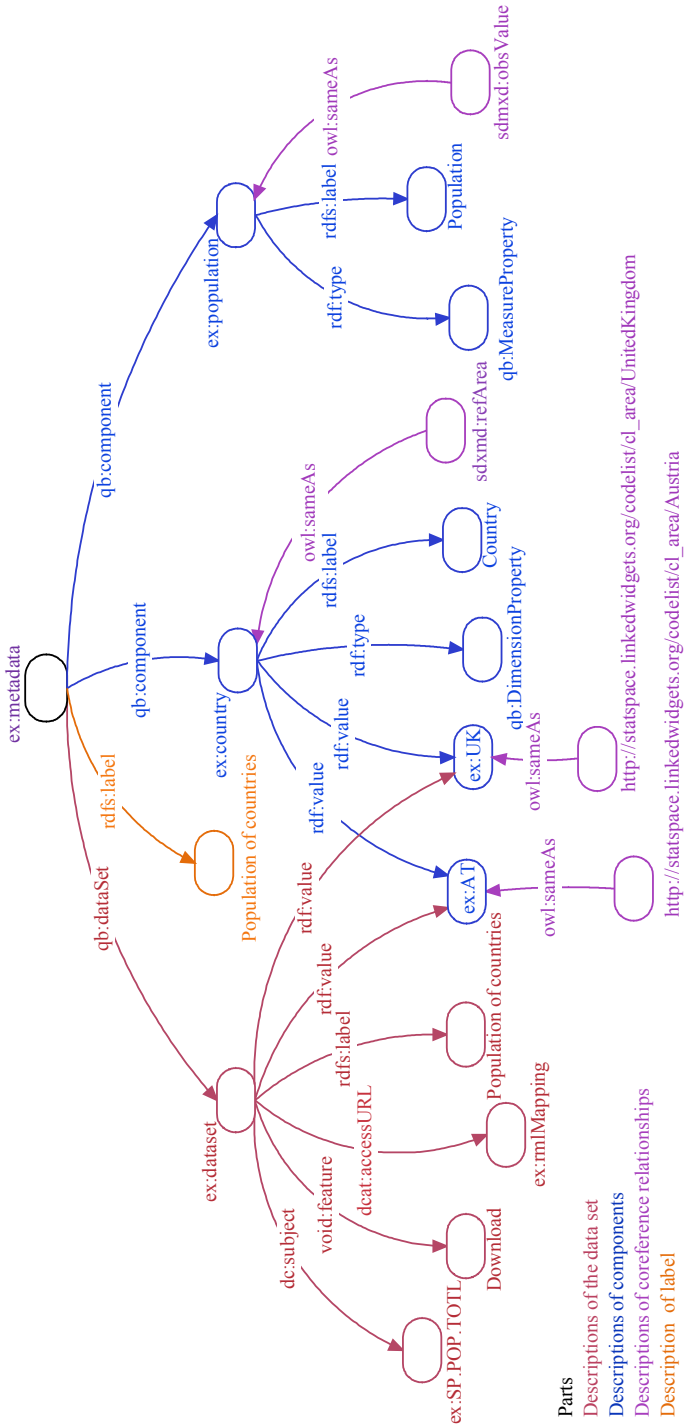


Figure 3.3: An example of metadata description (excerpt)

data set. In addition, we describe characteristics of this data set, including its subject (*dc:subject*), access mechanism (i.e., Download, API, or SPARQL via *void:feature*), access URL (i.e., the URL of the RML mapping or the endpoint via *dcat:accessURL*), label (*rdfs:label*), and the values of each component of this data set (*rdf:value*). Subject information is built upon a hierarchical pattern (*Topic.General Subject.Specific Subject. Extension*) that follows the structure used by World Bank (cf. Section 3.3). It plays a key role for data integration and exploration as it allows users to automatically identify data sets that contain particular indicators. Compared to keyword-based search, subject information can enhance search results and allows to classify and group results.

(ii) *Descriptions of components, which focus on data structure and values of each component.* We use *qb:component* to connect the metadata and each component in the data set. Each metadata description can link to multiple dimensions, but it has only one connection to a measure, and only one connection to an attribute that describes the unit of observed values. For each component, we describe its type (e.g., dimension, measure, and attribute via *rdf:type*), values of this component (only for dimensions and attribute via *rdf:value*), and its label (*rdfs:label*). Different data sets can use the same URI to represent a component (e.g., *sdmxd:refArea*). To model the range of values of each data set, this URI is linked to its possible values. Thereby, this URI will be linked to all other possible representations in different data sets. In order to indicate the values that occur in a specific data set, we use *rdf:value* to link the data set to the values of each component.

(iii) *Coreference relationships. owl:sameAs* is used to establish coreference relationships for components and their values with our shared URIs. Ideally, data providers use URIs to represent the values of each component. However, in some data sources, the values of a component are literals (e.g., *2016-05-01*). Therefore, we identify the coreference relationships and represent them via the triple ⟨ coreferenceURI, owl:sameAs, URI/literal ⟩.

(iv) *Label of the metadata.* In case multiple measure data sets are split into single measure data sets, we need to differentiate their respective metadata via labels of individual metadata descriptions, which are derived from the original label of the data set and of each measure. We create links between each metadata definition and a respective label via *rdfs:label*. In the case of single measure data sets, labels of the data set and its corresponding metadata are the same.

3.5 Metadata Generator

This service accepts either a URL of an RML mapping or a URL of a SPARQL endpoint as the input, analyzes the data sets, and returns the respective metadata descriptions. The most challenging issue in this process is the identification of the subject, the unit of measure, and coreference relationships between the URIs used in the data sets and

shared URIs. For example, the title of a data set is typically not sufficient to identify its subject. Hence, missing information needs to be added manually in some cases. In Sections 4.1.3 we present relevant algorithms for data source analysis and coreference resolution of components and their values.

3.6 Mediator

The *mediator* service supports advanced users who are familiar with the SPARQL query language in exploring statistical data sets. The service receives a SPARQL query as the input, analyzes it to identify matching data sets, rewrites the input query into appropriate queries for individual data sets, rewrites a set of results into a consolidated representation, then integrates this set in a single result before returning it to users. We describe each of these steps in the following.

Query analysis. The input query should be formulated based on shared URIs to represent subjects, components, and filter values. The *mediator* will analyze the input to identify filter conditions and graph patterns for matching. Next, the *mediator* will query the *metadata repository* to identify matching data sets. Access methods, coreference information regarding filter conditions, and data structures of these data sets are identified to perform the next steps.

Query rewriting. The *mediator* rewrites the query to translate it into appropriate queries for matching data sets. There are three cases: (i) if the query method of the data set is `SPARQL`, it will generate a new SPARQL query for the relevant endpoint; (ii) if the value is `API`, the mediator combines the RML mapping from the metadata with the parameters of the query; (iii) finally, if the value is `Download`, the mediator calls the *RML mapping service* to analyze the RML mapping and transform the data set into RDF.

Result rewriting. The *mediator* receives results from the data sources and rewrites the results into a consolidated representation. Each result may use distinct URIs and, hence, cannot be integrated immediately. The *mediator* therefore queries the *metadata repository* to identify coreference information of each data set, and then rewrites results into a new representation. Next, if relevant data sets use different units or scales to describe observed values, the mediator transforms these values into a common unit or scale.

Result integration. The *mediator* integrates results following the temporal dimension in an increasing order. As a result, users can compare different values of multiple sources for the same observation (e.g., statistics at the same time and same location) over time. Figure 3.4 shows the communication between components to answer an input SPARQL query.

3.7 Explorer

Explorer is a web application that enables non-expert users to visually interact with statistical data. Rather than formulating SPARQL queries, users can explore the data in three steps:

Search data sets. The *explorer* queries the *metadata repository* to provide metadata information of each data set to users including label, subject, and publisher.

Discover metadata descriptions. Users can discover the metadata through many ways such as (i) identify relatable data sets; (ii) visualize the data set; (iii) access the original data source; and (iv) display metadata via an interface.

Integrate multiple data sets. Assuming that users select an arbitrary metadata description, the application will provide a list of relatable data sets with the selected data set. Users then can choose data sets from this list to compare with the initial data set. Next, the *explorer* will invoke the *RML mapping service* and/or query relevant SPARQL endpoints to obtain results, integrate them, and visualize the result. Figure 3.5 shows the communication between components.

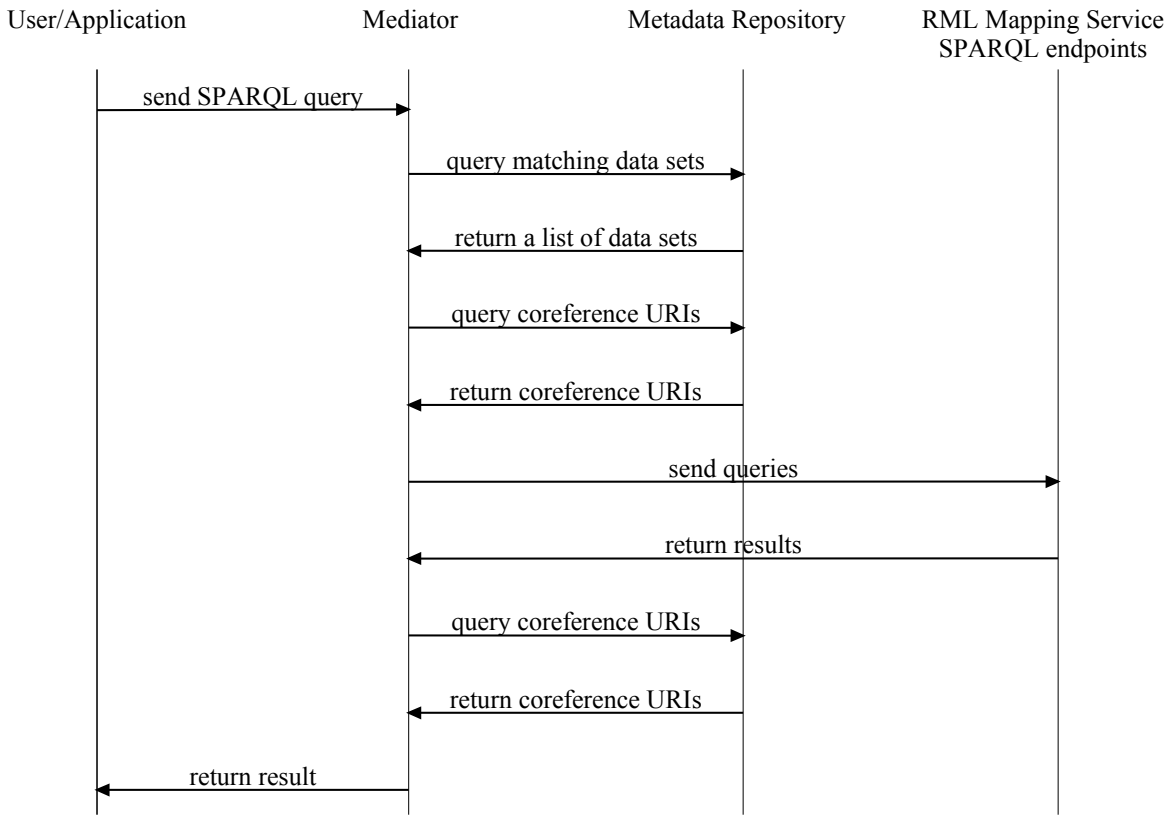


Figure 3.4: Communication between components to answer an input SPARQL query

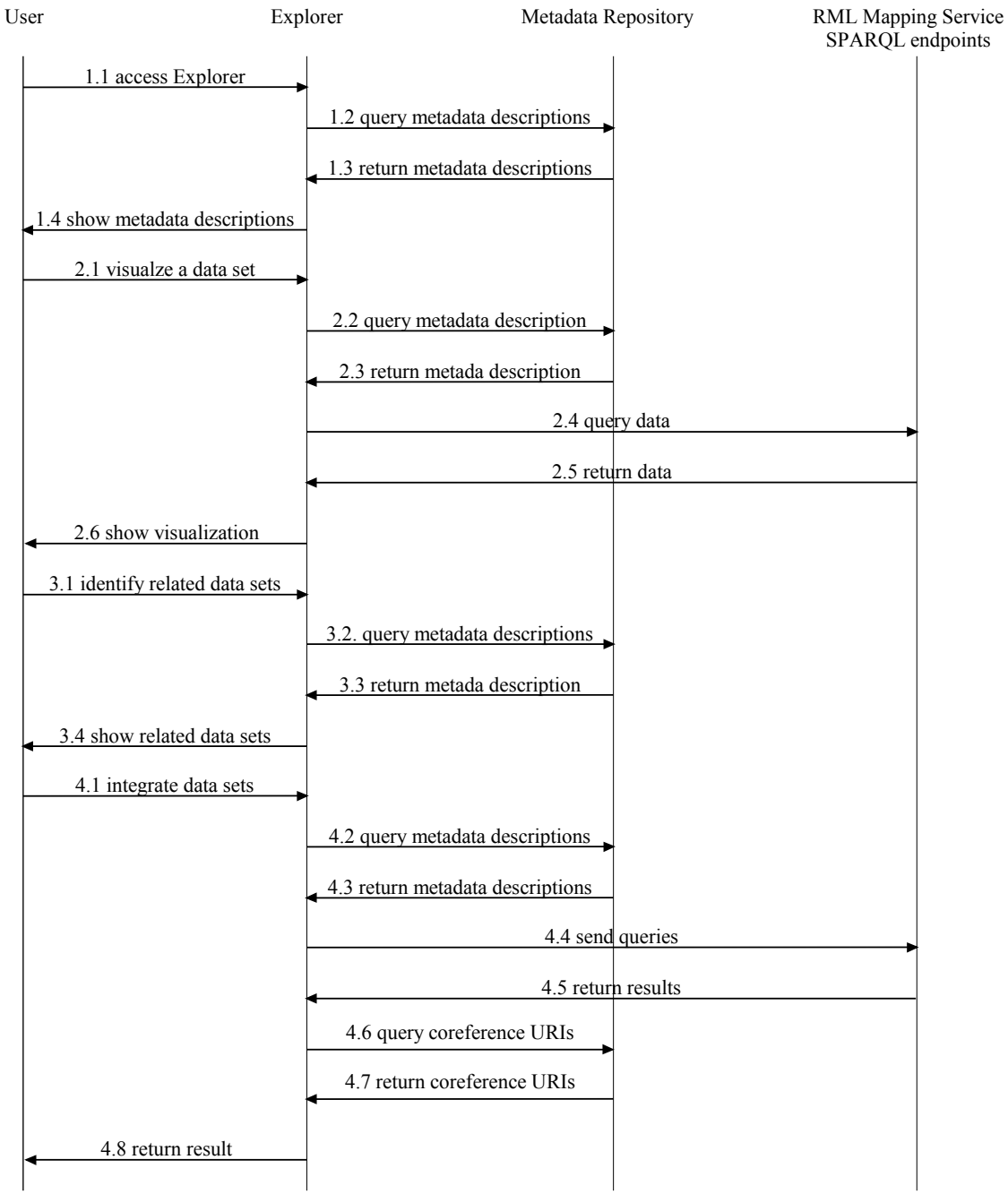


Figure 3.5: Communication between components to satisfy users' access needs

CHAPTER 4

Implementation

In this chapter, we first introduce the implementation of six elements of StatSpace architecture, i.e., the RML mapping service (Section 4.1.1), URI design patterns (Section 4.1.2), metadata generator (Section 4.1.3), metadata repository (Section 4.1.4), mediator (Section 4.1.5), and explorer (Section 4.1.6). Next, in Section 4.2 we illustrate StatSpace's usefulness by means of examples of data integration (Section 4.2.1), data quality assessment (Section 4.2.2), correlation mining (Section 4.2.3), and spatial data visualization (Section 4.2.4).

4.1 Elements

4.1.1 RML Mapping Service

In StatSpace, the *RML mapping service* relies on the RDF Mapping Language (RML) [33] to transform data from raw formats into RDF. Listings 4.1 and 4.2 illustrate an input data set and its corresponding RML mapping. In the mapping content, we need to declare the address of input data set (via *rml:source*), its format (via *rml:referenceFormulation*), iteration method (via *rml:iterator*), patterns to coin URIs (via *rr:tempate and rr:reference*), and the relationships between components in the data set (via *rr:predicate*).

At the beginning of building StatSpace [35] (in 2015), we made use of the RML mapping service[2] developed by our colleague, Peb Ruswono Aryan. This service is suitable for use in our initial example [35] of statistical data integration. However, its processor has a limitation regarding the size of input data sets. It cannot completely transform the World Bank data sets (approximately three Megabyte for each data set) into RDF. To address this issue, we created a new service based on the use of the open source RML

[1] `http://rml.io/`, accessed December 30, 2016
[2] `http://pebbie.org/mashup/rml`, accessed December 30, 2016

```
{ ... ,
  "Performance" :
  {
    "Perf_ID": "567",
    "Venue": {
      "Name": "STAM",
      "Venue_ID": "78"
    }
    "Location": {
      "lat": "51.043611",
      "long":"3.717222"
    }
  },
  ...
}
```

Listing 4.1: An example input data set

```
rml:logicalSource [
  rml:source "http://ex.com/performances.json"
      ;
  rml:referenceFormulation ql:JSONPath;
  rml:iterator "$.Performance.[*]"
];
rr:subjectMap [
  rr:template "http://ex.com/{Perf_ID}"
];
rr:predicateObjectMap [
  rr:predicate ex:name;
  rr:objectMap [
    rr:template "http://ex.com/{Name}"
  ]
].
```

Listing 4.2: Excerpt from RML mapping[1]

```
http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/
    mapping/wb.ttl&indicator=SP.POP.TOTL
```

Listing 4.3: An example query sent to the RML mapping service

processor[3] published by the authors of this language. At present, our service is available at `http://statspace.linkedwidgets.org/rml`. Listing 4.3 introduces a request sent to this service in order to translate the World Bank population data set into RDF.

The original RML processor published by the RML authors can transform data from three formats, i.e., CSV, JSON, and XML. We extend its capability to spreadsheet format and enhance the reusability of mappings by supporting variable declarations. The details of these extensions are presented in our paper [35] and implemented in Peb's RML mapping service. In our current service, we continue to improve the extensions and apply them to the original RML processor.

Data sources (e.g., the ONS) can publish data sets in spreadsheet formats such as XLS and XLSX, which are not supported by the original RML processor. Therefore, we extend this processor to support additional formats. The original RML processor consists of three separate sub-processors designed to analyze three input formats. Based on the particular format of an input data set, a sub-processor is used to analyze the mappings and then generate RDF triples. This mechanism enables developers to create new sub-processors for other formats. In our service, we implement a new sub-processor for XLS and XLSX formats and integrate it into the original RML processor. Instead of label-based iteration (such as "Performance" label) as in existing sub-processors (cf. Listings 4.1 and 4.2), users need to declare the range of cells and relevant sheets. We make use of the following two syntaxes for the iteration.

(i) *<Sheet name>!<beginning data cell>:<ending data cell>:<beginning header*

[3] `https://github.com/RMLio/RML-Processor/`, accessed December 30, 2016

cell>:<ending header cell>. The first syntax is designed for spreadsheet layouts whose headers are labels. For example, Table 4.1 and Listing 4.4 show a spreadsheet data set from the ONS data source and its corresponding declaration. The full RML mapping for this data set is provided in Listing C.2.

(ii) *<Sheet name>!<beginning data cell>:<ending data cell>:<beginning header cell>:<ending header cell>:<beginning header cell referring to value>:<ending header cell refering to value>*. The second syntax is intended for layouts whose headers can be values. For example, Table 4.2 shows another spreadsheet data set from the ONS data source where its headers from cells D2 to U2 are temporal values. Listing 4.5 presents an excerpt of the RML mapping for this data set. The full mapping is provided in Listing C.3.

Cell	**A**	**B**
3	Mid-Year	Population (millions)
5	1964	54.0
6	1965	54.3
...	...	...
54	2013	64.1

Table 4.1: An example spreadsheet data set uses the first layout

```
rml:logicalSource [
  rml:referenceFormulation  ql:Spreadsheet;
  rml:source "http://www.ons.gov.uk/ons/rel/pop-estimate/population-estimates-for-uk--
      england-and-wales--scotland-and-northern-ireland/2013/chd-1-for-story.xls";
  rml:iterator "Table!A5:B54:A3:B3";
];
```

Listing 4.4: An example of iteration declaration used for the first layout

Cell	**C**	**D**	**E**	**...**	**U**
2	Age	Mid-1953	Mid-1954	...	Mid-1970
3	All Ages	50,592.9	50,764.9	...	55,632.2
4	0-4	3,959.8	3,894.8	...	4,617.7
5	5-9	4,203.0	4,255.5	...	4,676.6
...	...	...	...	...	...
21	85+	248.0	263.2	...	443.9

Table 4.2: An example of spreadsheet data set uses the second layout

```
rml:logicalSource [
  rml:referenceFormulation  ql:Spreadsheet2;
  rml:source "http://www.ons.gov.uk/ons/about-ons/business-transparency/freedom-of-
      information/what-can-i-request/published-ad-hoc-data/pop/july-2015/uk-population
      -estimates-1851-2014.xls";
  rml:iterator "UK␣Quinary␣1953-1970!C3:U21:C2:C2:D2:U2";
];
```

Listing 4.5: An example of iteration declaration used for the second layout

Furthermore, we support users in declaring and using variables in their mappings. Thereby, an RML mapping can be reused for multiple data sets that share the same structure. This extension is motivated by the feature of APIs that allows users to use parameters in order to query parts of a data set as well as different data sets. In our service, the RML mapping for the World Bank data source consists of two variables, i.e., *indicator* (represents subject of data set) and *refArea* (represents ISO codes of countries). We use this mapping for data transformation of more than 1,400 World Bank data sets. Users can pass values for these variables through URL parameters. Listing 4.6 shows an excerpt of this mapping (cf. B.1) regarding declaration and use of variables.

```
<#Parameters>
  rmlx:defaultValue
        [rmlx:varName "indicator"; rr:constant "SP.POP.TOTL"],
        [rmlx:varName "refArea"; rr:constant "all"]
.
<#Observation>
  rml:logicalSource [
    rml:source  "http://api.worldbank.org/countries/{refArea}/indicators/{indicator}?
        format=json&page=1&per_page=15000";
    rml:referenceFormulation ql:JSONPath;
    rml:iterator "$[1].*"
  ];
```

Listing 4.6: An RML mapping contains variables (excerpt)

4.1.2 URI Design Patterns

The aim of URI design patterns is to coin a set of shared URIs which play a role as a central reference point for the linking of URIs used in different data sources. In this section we provide descriptions regarding shared URIs that can be organized into two groups: (i) URIs represent components and (ii) URIs represent values in code lists.

Tables 4.3, 4.4, and 4.5 present URIs in the first group. To construct these URIs, we first reuse existing URIs defined in *sdmxd*, *sdmxm*, and *sdmxa* name spaces. Next, we make use of a pattern based on the base URI, i.e., `http://statspace.linkedwidgets.org/`

dimension/ in order to represent new dimensions that are not available in *sdmxd*. The detailed results of this approach are as follows.

First, we reuse the nine URIs in *sdmxd* to represent dimensions including reference area (*sdmxd:refArea*), reference period (*sdmxd:refPeriod*), age (*sdmxd:age*), education level (*sdmxd:educationLev*), occupation (*sdmxd:occupation*), currency (*sdmxd:currency*), civil status (*sdmxd:civilStatus*), frequency (*sdmxd:frequency*), and sex (*sdmxd:sex*). There are two additional dimensions, i.e., *expenditure* and *economic activity*. Second, regarding measure component, we match URIs of measures used in data sets to the generic concept defined in *sdmxm*, i.e., *sdmxm:obsValue*. Finally, for attributes that describe the unit of observed values, we match them to the corresponding URI in *sdmxa*, i.e., *sdmxa:unitMeasure*.

No.	URI	Label
1	http://purl.org/linked-data/sdmx/2009/dimension#refArea	Reference Area
2	http://purl.org/linked-data/sdmx/2009/dimension#refPeriod	Reference Period
3	http://purl.org/linked-data/sdmx/2009/dimension#age	Age
4	http://purl.org/linked-data/sdmx/2009/dimension#educationLev	Education Level
5	http://purl.org/linked-data/sdmx/2009/dimension#occupation	Occupation
6	http://purl.org/linked-data/sdmx/2009/dimension#currency	Currency
7	http://purl.org/linked-data/sdmx/2009/dimension#civilStatus	Civil Status
8	http://purl.org/linked-data/sdmx/2009/dimension#freq	Frequency
9	http://purl.org/linked-data/sdmx/2009/dimension#sex	Sex
10	http://statspace.linkedwidgets.org/dimension/economicActivity	Economic Activity
11	http://statspace.linkedwidgets.org/dimension/expenditure	Expenditure

Table 4.3: Shared URIs used to represent dimensions

No.	URI	Label
1	http://purl.org/linked-data/sdmx/2009/measure#obsValue	Observation Value

Table 4.4: Shared URI used to represent measure

No.	URI	Label
1	http://purl.org/linked-data/sdmx/2009/attribute#unitMeasure	Unit of Measure

Table 4.5: Shared URI used to represent attribute

Code list	|URI design patterns|	|URIs|	Hierarchy Support
CL_Area	1	Unlimited	Yes
CL_Period	11	Unlimited	Yes
CL_Economic_Activity	1	996	Yes
CL_Age	1	209	Yes
CL_Education_Level	1	9	No
CL_COICOP	1	230	Yes
CL_COFOP	1	188	Yes
CL_COPP	1	51	Yes
CL_COPNI	1	65	Yes
CL_Occupation	1	619	Yes
CL_Currency	1	180	No
CL_Civil_Status	1	8	No
CL_Frequency	1	9	No
CL_Sex	1	5	Yes
CL_Unit_Measure	1	43	No
CL_Subject	1	1613	Yes

Table 4.6: Characteristics of code lists

Table 4.6 provides an overview of URI design patterns and associated URIs in the second group. We can break this group into the following four sub-groups.

Code list for reference area dimension. To cover spatial areas, we define a consolidated pattern as follows: `Country/Administrative_Area_Level_1/.../Administrative_Area_Level_n`.[4] For example, the city of Linz can be represented by the URI, i.e., `Austria/UpperAustria/Linz`.

Code list for reference period dimension. We make use of seven patterns defined by the UK time reference service to describe arbitrary time spans. Table 4.7 provides a complete list of patterns used to coin URIs for this code list.

[4]The base URI is `http://statspace.linkedwidgets.org/codelist/cl_area/`

No.	Pattern	Description
1	`/id/gregorian-year/{year}`	Year
2	`/id/gregorian-half/{year}-{half}`	One-half year
3	`/id/gregorian-quarter/{year}-{quarter}`	Quarter
4	`/id/gregorian-month/{year}-{month}`	Month
5	`/id/gregorian-day/{year}-{month}-{day}`	Day
6	`/id/gregorian-week/{year}-{week}`	Week
7	`/id/gregorian-interval/{dateTime}/{duration}`	Duration

Table 4.7: Patterns designed for code list of temporal dimension[5]

Code lists for the remaining dimensions. SDMX presents code lists[6] in text format for the following nine dimensions: economic activity, age, education level, expenditure, occupation, currency, civil status, frequency, and sex.[7] We first reuse the URIs defined in *sdmxcode* to represent values of the *frequency* and *sex* dimensions. Next, we coin new URIs to represent values of the remaining dimensions that are not defined in *sdmxcode*. Moreover, we build hierarchies for age, economic activity, expenditure, occupation, and sex based on *skos:narrower* and *skos:broader* relationships. This allows StatSpace to return aggregated values when data sets that use different levels of granularity are integrated. For instance, the top concept in the age code list is *total* (i.e., `http://statspace.linkedwidgets.org/codelist/cl_age/total`) which is split into various age groups such as 0-4, 5-9, ... , 105-109 (type: *age-group*), and special values, e.g., 70+, 75+, 80+ (type: *age-plus*). Each age group is split into individual ages (type: *age-individual*), and each special value is split into age groups. Appendix A provides detailed descriptions about all code lists.

Subjects and units of measure. Each World Bank data set has a subject which follows a hierarchical structure, i.e., *Topic.General Subject.Specific Subject.Extensions*. We reuse the available World Bank code list[8] and extend it with 180 new codes for subjects and 14 new codes for units of measure. Furthermore, we define a pattern to represent orders of magnitude, i.e., *unit.power*, to extend the code list towards units that likely appear in statistical data sets.

[5] The base URI is `http://reference.data.gov.uk`

[6] `https://sdmx.org/?page_id=3215`, accessed December 30, 2016

[7] The expenditure dimension consists of four code lists, i.e., classification of individual consumption by purpose (COICOP), classification of functions of government (COFOG), classification of purposes of non-profit institutions serving households (COPNI), and classification of outlays of producers by purpose (COPP).

[8] `http://databank.worldbank.org/data/download/site-content/WDI_CETS.xls`, accessed December 30, 2016

4.1.3 Metadata Generator

In this section, we introduce a *data source analysis* algorithm and *mapping algorithms* used to generate descriptions for each statistical data set. To illustrate the implementation of the algorithms, we choose a data set[9] from EUODP[10], that represents the average European Commission funding per participation in FP7-ICT projects[11]. Table 4.8 shows a set of observations in this data set which has two dimensions, i.e., *ex:country* and *ex:year*, one measure, i.e., *ex:value*, and one attribute, i.e., *ex:unit*.

ex:country	ex:year	ex:value	ex:unit
ex:Austria	ex:2007	358,279	euro
ex:Germany	ex:2007	414,531	euro
ex:Austria	ex:2008	358,133	euro

Table 4.8: An excerpt from the example data set[12]

Data Source Analysis Algorithm

This algorithm receives either a SPARQL endpoint following the QB or an RDF file generated by the *RML mapping service* as the input. The algorithm then analyzes triples in the input data source to identify: (i) a list of data sets in the data source; (ii) dimensions, measures, and attributes associated with each data set; and (iii) a list of values for dimensions and attributes. This information will be used to construct metadata descriptions for each data set.

Data Set Identification. The QB vocabulary allows us to identify a data set (e.g., *ex:dataset*) through one of the following triple patterns: (i) $ex : observation - qb : dataSet \rightarrow ex : dataset$, (ii) $ex : dataset - rdf : type \rightarrow qb : DataSet$, and (iii) $ex : dataset - qb : structure \rightarrow ex : dsd$ (cf. Figure 2.1). The first pattern is typically available in all SPARQL endpoints published on the web, because it describes the relationship between a data set and its observations. However, as an endpoint such as the EEA contains millions of observations, the related query (the third query in Listing 4.7), which is essential for data set detection, may return a time out error. To alleviate this issue, we can make use of the two latter triple patterns. They represent a 1:1 relationship between a data set and its type as well as between a data set and its data structure definition, respectively. However, data sources such as the Open Data Communities (ODC) can focus on describing observations without representing the remaining relationships. Therefore, an effective combination of these triple patterns not only results in high recall for detecting

[9] `http://data.lod2.eu/scoreboard/ds/indicator/FP7ICT_afxp_All_partners_euro`, accessed December 30, 2016

[10] `http://data.europa.eu/euodp/en/linked-data`, accessed December 30, 2016

[11] `http://cordis.europa.eu/fp7/ict/home_en.html`, accessed December 30, 2016

[12] We changed specific prefixes used in this data set by a generic prefix, i.e., ex

data sets but also can diminish time for query. Listing 4.7 shows a list of queries that we use to identify data sets and their labels in statistical data sources.

```
//1st query
SELECT DISTINCT ?ds ?l
WHERE{
    ?ds rdf:type qb:DataSet.
    FILTER EXISTS{?o qb:dataSet ?ds}
    optional{?ds rdfs:label ?l.}
}

//2nd query
SELECT ?ds ?l
WHERE{
   ?ds rdf:type qb:DataSet.
   optional{?ds rdfs:label ?l.}
}
```

```
//3rd query
SELECT DISTINCT ?ds ?l
WHERE{
    ?o qb:dataSet ?ds.
    optional{?ds rdfs:label ?l.}
}

//4th query
SELECT ?ds ?l
WHERE{
   ?ds qb:structure ?dsd.
   optional{?ds rdfs:label ?l.}
}
```

Listing 4.7: Queries used to identify statistical data sets

Source	**SPARQL endpoint**
EUODP	`http://data.europa.eu/euodp/en/linked-data`
EEA	`http://semantic.eea.europa.eu/sparql`
Scottish Statistics (ScotStat)	`http://statistics.gov.scot/sparql`
CSO	`http://data.cso.ie/sparql`
ODC	`http://opendatacommunities.org/sparql`
Vienna Open Government Data (VOGD)	`http://ogd.ifs.tuwien.ac.at/sparql`

Table 4.9: SPARQL endpoints used for testing algorithms

Components Identification. We make use of predicates *qb:dimension, qb:measure, qb:attribute* and classes containing *qb:Dimension, qb:Measure, qb:Attribute* to indicate the role of each component in a data set. If these predicates and classes are not used in a data source, we derive dimensions, measures, and attributes based on their URIs and values in the observations. In a statistical data set, the values of a measure are typically represented in numerical format (e.g., 358,279 in Table 4.8) whereas the values of dimensions are often either temporal values or URLs (e.g., 2013-12-31, ex:Austria, ex:Germany, etc.) and the values of attributes typically refer to units of observed measure (e.g., currency, distance, scale, etc.). Furthermore, URIs of components may indicate their role, such as *http://.../dimension/..* may refer to a dimension.

Values and Labels Identification. A complete list of values for each dimension and attribute should be obtained, because it is an important part in metadata description of each data set. In addition, we need to identify labels of components in a data set because they support users in understanding the meaning of these components. For example, in a data set of the VOGD, the measure `http://.../Pas` does not make sense for users, while its label – *"Number of passengers"* is a better meaningful description.

Mapping Algorithms

In a statistical data set, each component is represented by a URI, e.g., *ex:country*, *ex:year*, *ex:unit*, etc. In addition, this component is attached to a set of specific values, for instance, the dimension *ex:country* in Table 4.8 contains two values, i.e., *ex:Austria*, *ex:Germany*. Therefore, the task of mapping algorithms is to identify equivalent relationships between components and their values with respectively shared URIs introduced in Section 4.1.2.

We built eleven algorithms for the mapping of eleven dimensions (cf. Table 4.3) and one algorithm for the mapping of the unit attribute. The validation of these mappings is provided in Section 5.4. In order to map URIs of components (e.g., *ex:country, ex:year*) used in an arbitrary data set to correspondingly shared URIs (e.g., *sdmxd:refArea, sdmx:refPeriod*), we collect keywords used to represent these components and their labels in existing data sources. Table 4.10 presents keywords that we use for mapping. We do not have many keywords for five dimensions including *education level, occupation, currency, economic activity*, and *expenditure* because these dimensions do not appear in the data sources collected. However, we still build related algorithms to be able to generate mappings for new data sources that we will collect in the future. Next, to map values of dimensions and the unit attribute to their shared URIs, each mapping algorithm uses one of the following approaches: (i) we use Google geocoding service to map spatial values to shared URIs; (ii) we use patterns to recognize temporal values and ages, because in a statistical data set these values typically follow a common design; (iii) we can use keywords for the mapping of the remaining components because the number of values in these components is typically limited. For example, the sex dimension contains only five values, i.e., *male, female, total, unknown*, and *not applicable*. In the following, we introduce two key *mapping algorithms* including *spatial dimension mapping* and *temporal dimension mapping* because of the frequent appearance and important role of these two dimensions in statistical data.

Spatial Dimension Mapping Algorithm

The spatial dimension (e.g., *ex:country*, *sdmxd:refArea*) describes geographical area(s) where statistical observations were made. Although this dimension appears in most statistical data sets, it still can be missing if a data set describes data of a unique geographical area. For example, the EEA contains data on landings of fishery products in Germany[13], but this data set does not describe a spatial dimension, because it implicitly

[13] `http://rdfdata.eionet.europa.eu/page/eurostat/data/fish_ld_de`

No.	Component	Keywords
1	Reference area	ref-area, refarea, ref area, country, state, place, geocode, region, reference area
2	Reference period	ref-period, refperiod, ref period, reference period, date, year, time-period, time period, timeperiod, time period
3	Age	/age, _age, #age, refage
4	Civil status	civil status
5	Frequency	freq
6	Sex	sex, gender
7	Education level	education level
8	Occupation	occupation
9	Currency	currency
10	Economic activity	economic
11	Expenditure	function of government, individual consumption, outlays of producer, puproses of non-profit institution
12	Attribute (unit-of-measure)	unit

Table 4.10: Keywords used for the recognition of components

refers to Germany via its label. In this case, we need to explicitly add a spatial dimension to the structure of the data set as well as identify the geographical area attached to it (e.g., *Germany* in this case). This addition allows users to compare and integrate data of Germany in this data set with other data sets. To this end, if a data set does not have spatial dimension, we compare the label and URI of this data set with a list of countries and their ISO codes to identify the country which it represents. As a result, after this pre-processing step, all statistical data sets in SPARQL endpoints that we collected, contain a spatial dimension. Next, in order to consolidate different URIs used to represent spatial dimension, we match them to *sdmxd:refArea*.

Statistical data sources typically only contain data of a limited number of geographical areas. For instance, two data sources, i.e., EUODP and European Commission (EC) contain statistical data of European countries, whereas CSO includes only the data of one country (Ireland in this case). As a result, no data source contains a large number of geographical URIs that can be reused in all other data sources. The reuse of URIs is also limited by inconsistent geographical classification. For example, the EEA use Nomenclature of territorial units for statistics (NUTS) to gather data at three levels (i) major socio-economic regions - NUTS1 such as Ostösterreich (Eastern Austria), Südösterreich (Southern Austria), Westösterreich (Western Austria); (ii) basic geographical regions - NUTS2 such as Burgenland, Vienna, Salzburg; and (iii) smaller regions - NUTS3 such as Nordburgenland (Northern Burgenland), Mittelburgenland

(Central Burgenland), and SüdBurgenland (Southern Burgenland). The VOGD, by contrast, organizes statistical data based on so-called administrative areas. This means that it does not contain, for instance, regions classified by cardinal direction such as Eastern Austria. A mapping of geographical areas from multiple data source is therefore a challenging issue.

To solve this issue, our approach is to use a consolidated pattern to coin URIs for any geographical area in the world. The URIs generated by this pattern therefore play a role as a common coreference point to link URIs used in existing data sources. To this end, we consider available geographical data sources and services. At present, Google Maps[14] is one of the largest data sources regarding global spatial information. Google also provides Geocoding APIs [15] to convert the name of areas into geographical coordinates and hierarchical information. Furthermore, this service returns the same result for different names of the same area such as "Vienna Austria", "Wien Österreich". Listings 4.8 and 4.9 present a request to Google Maps and its respective result. Using this service, our approach is described as follows: (i) our pattern relies on hierarchical information that Google Maps provides to coin a unique URI for each area. For instance, Vienna city will be represented by "Austria/Vienna/Vienna"; (ii) we use labels and URIs of spatial values as the input of the mapping algorithm. Next, we make use of Google's geocoding APIs to obtain spatial information; and (iii) we use the results and the original URIs of spatial values to identify their coreference URIs.

```
https://maps.googleapis.com/maps/api/geocode/json?address=Vienna%20Austria
```

Listing 4.8: An example concerning a geocoding request

There are two main issues which need to be overcome when using this service: (i) for one input, the service is likely to return a large number of different areas due to its ambiguity; and (ii) the service works with administrative areas, hence, regions classified based on cardinal directions (e.g., Eastern Austria, Southern Austria) do not yield correct results.

In order to solve these issues, we make use of the original geographical classification that each data source relied on to form URIs. For example, the EEA uses the NUTS classification where lower territorial levels (e.g., nuts:AT1 – Eastern Austria, nuts:AT2 – Southern Austria) are formed based on the URIs of areas at one level higher than them (i.e., nuts:AT – Austria) plus one character (i.e., '1' and '2', respectively). Another example, in the CSO and VOGD, URIs of lower territorial levels (e.g., cso:CTY/C01, vogd:AUT/Vienna) are formed based on their higher level URIs (i.e., cso:CTY and vogd:AUT, respectively) plus their names or codes (i.e., C01 and Vienna, respectively). To determine whether area_i is one level lower than area_k in a geographical classification, we define an evaluation function based on existing classifications to identify this relationship. Using this function, our approach to address two issues is as follows: (i) to solve the first

[14] https://maps.google.com/, accessed December 30, 2016

[15] https://developers.google.com/maps/documentation/geocoding/start, accessed December 30, 2016

```
{
        "results" : [
        {
                "address_components" : [
                {
                        "long_name" : "Vienna",
                        "short_name" : "Vienna",
                        "types" : [ "locality", "political" ]
                },
                {
                        "long_name" : "Vienna",
                        "short_name" : "Vienna",
                        "types" : [ "administrative_area_level_1", "political" ]
                },
                {
                        "long_name" : "Austria",
                        "short_name" : "AT",
                        "types" : [ "country", "political" ]
                }
        ],
                "formatted_address" : "Vienna, Austria",
                "geometry" : {
                        "bounds" : {
                                "northeast" : {
                                "lat" : 48.3230999,
                                "lng" : 16.5774999
                        },
                        "southwest" : {
                                "lat" : 48.1182699,
                                "lng" : 16.18262
                        }
                },
                "location" : {
                        "lat" : 48.2081743,
                        "lng" : 16.3738189
                }
        }
}
```


Listing 4.9: An example about geocoding response (excerpt)

issue, we first order the areas such that broader areas are positioned before narrower areas. Next, we use this ordering to reduce the ambiguity by adding the label of the broader area to the queries of its narrower areas; (ii) to address the second issue, we create a new URI combining the URI of its broader area and its label. Figure 4.1 shows an example of spatial value mapping between two data sources the EEA and VOGD, and our shared URIs. We match different URIs regarding Austria including *nuts:AT* and *vogd:AUT* to a unique URI, i.e., *ssarea:Austria*. Furthermore, we use predicates *skos:narrower* and *skos:broder* to represent the relationships between URIs in the code list.

Assume that $L = \{l_1, \ldots, l_n\}$ is a set of spatial values associated with the spatial dimension in a data source and $G = \{g_1, \ldots, g_n\}$ is the output of the algorithm (Algorithm 1).

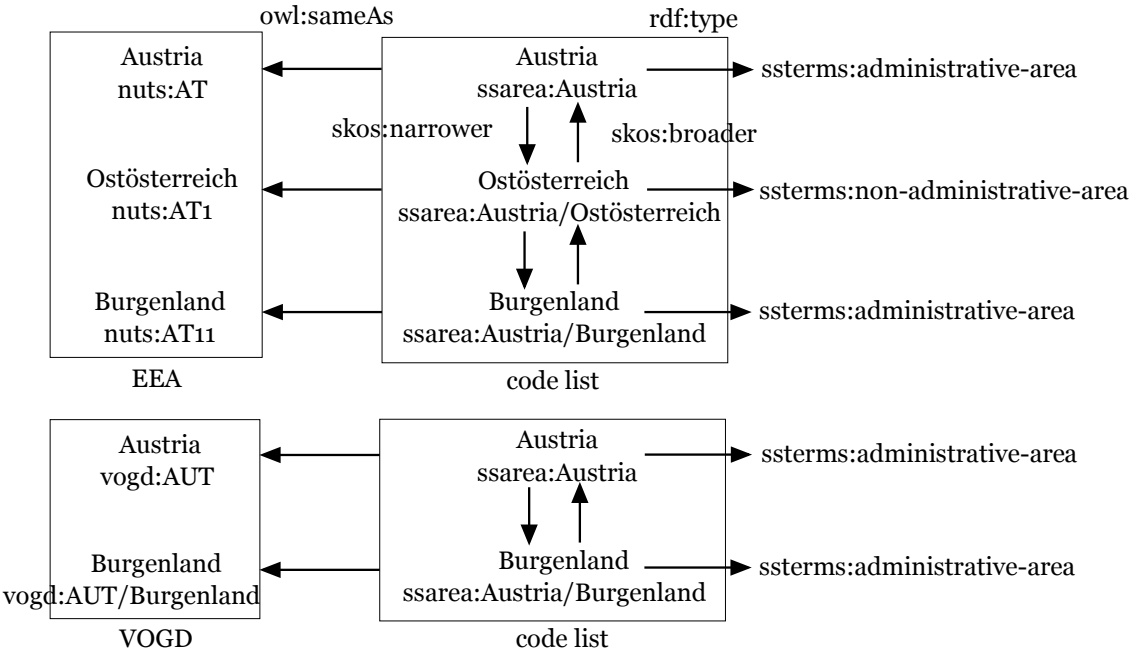


Figure 4.1: An example of spatial value mapping

Each $l_i \in L$ is a pair (uri_l_i, label_l_i) that contains its URI and label. Our algorithm solves the inconsistency of spatial URIs by mapping each area $l_i \in L$ to g_i based on Google geocoding service. Given l_i as an input, the service can return a set of probable areas. The algorithm returns one area from this result. In addition, we use the predicates *skos:broader* and *skos:narrower* to build relationships between areas in the resulting set G. In our algorithm, both triples *(area$_k$, skos:narrower, area$_i$)* and *(area$_i$, skos:broader, area$_k$)* mean that in a geographical classification, area$_k$ is one level higher than area$_i$. We designed Algorithm 1 to match areas in L to G as follows:

- To query an area l_i, we combine its label with the label of area l_k, in which l_k is the broader area of l_i.
- We filter the results in g_i in two steps (assume that $g_i = \{r_1, \ldots, r_m\}$). First, because l_k is a broader area of l_i, we remove results in g_i that are not a broader area of g_i. Second, we use latitude and longitude information (cf. Listing 4.9) in each result to select a unique result r_j which has a minimal distance to adjacent areas g_{i-1} and g_{i-2} (if $i \geq 2$).
- Direction-classified regions are assigned new URIs based on the URI of its broader area (i.e., uri_k) and its label (i.e., $label_i$). To distinguish these regions, we set their type to *non-administrative area*. As a result, data can be aggregated from narrower areas through two distinctive groups: administrative areas (e.g., Burgenland, Vienna) and non-administrative areas (e.g., Östosterreich, Südösterreich).

Temporal Dimension Mapping

The temporal dimension (e.g., *ex:year*, *sdmxd:refPeriod*) represents the time period(s) in which data publishers collected observations such as day, month, quarter, or year. It allows us to compare changes of observed values over time. Despite this importance, some data sets lack a temporal dimension. For example, VOGD contains statistical data of election results at different areas of Austria in 2013, but this data set does not have a temporal dimension in its data structure. By matching URIs and labels of such data sets with time patterns, we add a temporal dimension to their structures. As a result, after the pre-processing step, all the statistical data sets in data sources that we collected contain a temporal dimension. Next, to consolidate different URIs referring to this dimension, we map them to a shared URI, i.e., *sdmxd:refPeriod*.

Data sources follow different approach to represent temporal values, For example, EUODP defines its own URIs such as `http://data.lod2.eu/scoreboard/year/2012`, whereas the EEA uses literal values, e.g., *2012-12-31*, and the EC and VOGD make use of a Gregorian URI set provided by the `data.gov.uk` time reference service, e.g., `http://reference.data.gov.uk//id/gregorian-year/2012`. This time reference service provides semantics for a wide range of temporal values. Therefore, we choose URIs defined by this service to construct the code list for the temporal dimension.

Algorithm 2, which maps temporal values used in data sources to shared URIs, receives either a URI or a literal value as input. Then, by using time patterns, it identifies contained intervals and relates them to the corresponding URIs according to the Gregorian URI scheme. Furthermore, using the semantics provided by the service, we create *time:intervalContains* relationships between this URI and its related intervals in the code list. Figure 4.2 illustrates an example of temporal value mapping.

Algorithm 1 Geographical Area Mapping

Input: $L = \{l_1, \ldots, l_n\}$, $l_i = (uri\_l_i, label\_li)$
Output: Mapping L to G, $G = \{g_1, \ldots, g_n\}$,
$g_i = (uri\_g_i, label\_gi, lat\_g_i, lng\_gi, type\_gi)$
procedure GeographicalAreaMapping(L)
  sortInAscendingOrder(L)
  for each area $l_i \in L$ **do**
    $k \leftarrow indexOfBroaderArea(L, l_i)$
    if k!=-1 **then**
      if $type\_g_k$ *is administrative-area* **then**
        $queryString \leftarrow label\_l_i + label\_l_k$
      else
        $queryString \leftarrow label\_li$
        $uri\_b \leftarrow uriOfBroaderArea(uri\_g_k)$
      end if
    else
      $queryString \leftarrow label\_li$
    end if
    $g_i \leftarrow queryGoogleAPI(queryString)$
    ▷ $g_i = \{r_1, \ldots, r_m\}$, $r_j = (uri\_r_j, label\_r_j)$
    for each result $r_j \in g_i$ **do**
      if ($type\_g_k$ is administrative area and
       !isUriOfBroaderArea(uri_g_k, uri_r_j)) or
       ($type\_g_k$ is non-administrative area and
       !isUriOfBroaderArea(uri_b, uri_r_j) **then**
        remove r_j from g_i
      end if
    end for
    if size(g_i)>1 **then**
      removeByDistance(G, g_i)
    end if
    if size(g_i)=1 **then**
      set type_g_i is administrative area
    else
      uri_g_i ← uri_g_k + label_l_i
      set type_g_i is non-administrative area
    end if
    if k!=-1 **then**
      set (uri_g_i, *sw:broader*, uri_g_k)
      set (uri_g_k, *sw:narrower*, uri_g_i)
    end if
  end for
end procedure

43: **procedure** sortInAscendingOrder(L)
44: ▷ *sort areas in L in ascending order of uri*
45: **end procedure**
46: **procedure** queryGoogleAPI(queryString)
47: ▷ *return query results of Google Geocoding APIs*
48: **end procedure**
49: **procedure** removeByDistance(G, g_i)
50: ▷ *retain only one result in g_i, that is, the one which has the minimal distance to area(s) g_{i-1} and g_{i-2} (if i>=2)*
51: **end procedure**
52: **procedure** uriOfBroaderArea(uri_g_k)
53: ▷ *return uri of the area which is the broader area of the input uri*
54: **end procedure**
55: **procedure** isUriOfBroaderArea(uri_g_k,uri_g_j)
56: ▷ *return true if uri_g_k is a broader area of uri_g_i, else return false*
57: **end procedure**
58: **procedure** indexOfBroaderArea(L, l_i)
59: ▷ *return index of the area which is a broader area of l_i in list L*
60: **end procedure**

Algorithm 2 Temporal Value Mapping

1: **Input**: $T = \{t_1, \ldots, t_n\}$
2: **Output**: Mapping T to U, $U = \{uri_1, \ldots, uri_n\}$
3: **procedure** TemporalValueMapping(T)
4: uk = "http://reference.data.gov.uk/id/"
5: pYear = "[1-9][0-9]{3}"
6: pHYear = "[1-9][0-9]{3}-H[1-2]"
7: pQuarter = "[1-9][0-9]{3}-Q[1-4]"
8: pMonth = "[1-9][0-9]{3}-[0-1][0-9]"
9: pWeek = "[1-9][0-9]{3}-W[1-52]"
10: pDate = "[1-9][0-9]{3}-[0-1][0-9]-[0-3][0-9]"
11: pDuration = "[1-9][0-9]{3}-[0-1][0-9]-[0-1][0-9]T"

```
12:     for each value t_i ∈ T do
13:         if pDuration match t_i then
14:             v = getValue(pDuration, t_i)
15:             uri_i = uk + "gregorian-interval/"+v
16:         else
17:             if pDate match t_i then
18:                 v = getValue(pDate, t_i)
19:                 uri_i = uk + "gregorian-date/"+v
20:             else
21:                 if pWeek match t_i then
22:                     v = getValue(pWeek, t_i)
23:                     uri_i = uk + "gregorian-week/"+v
24:                 else
25:                     if pMonth match t_i then
26:                         value = getValue(pMonth, t_i)
27:                         uri_i = uk + "gregorian-month/"+v
28:                     else
29:                         if pQuarter match t_i then
30:                             value = getValue(pQuarter, t_i)
31:                             uri_i = uk + "gregorian-quarter/"+v
32:                         else
33:                             if pHYear match t_i then
34:                                 value = getValue(pHYear, t_i)
35:                                 uri_i = uk + "gregorian-half/"+v
36:                             end if
37:                             if pYear match t_i then
38:                                 value = getValue(pYear, t_i)
39:                                 uri_i = uk + "gregorian-year/"+v
40:                             end if
41:                         end if
42:                     end if
43:                 end if
44:             end if
45:         end if
46:         if uri_i !=null then
47:             queryMeaning(uri_i)
48:         end if
49:     end for
50: end procedure
51: procedure getValue(P, t)
52:     ▷ return contained interval in t through using pattern P
53: end procedure
54: procedure queryMeaning(uri)
55:     ▷ return semantics of this uri through the use of time service
56: end procedure
```

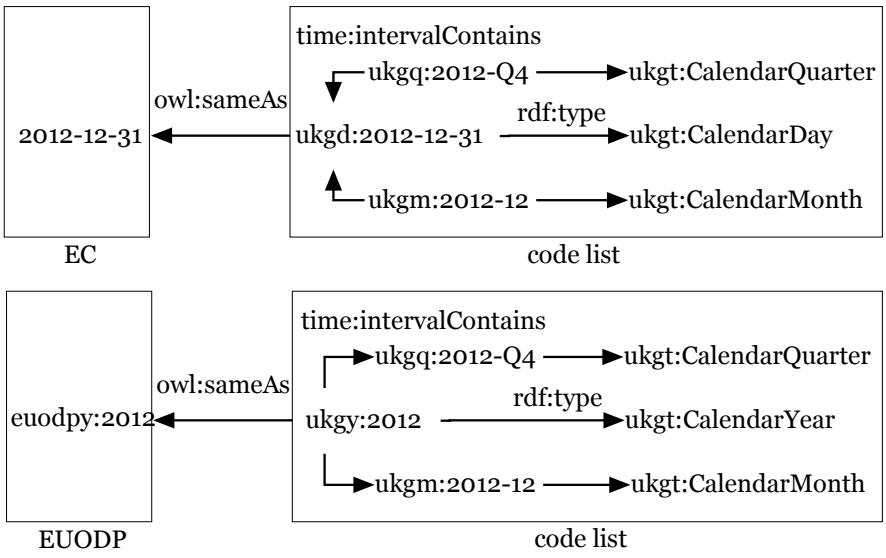


Figure 4.2: An example of temporal value mapping

4.1.4 Metadata Repository

The metadata repository is a collection of metadata descriptions of statistical data sets that we collected. Users can access this repository at our space[16] in datahub[17] as well as from a SPARQL endpoint[18]. Listing 4.10 and 4.11 introduce two example queries used to list information in this repository.

```
select distinct ?ds ?p
where{
  ?md qb:dataSet ?ds.
  optional{?md dcterms:publisher ?p}
}Order by ?ds
```

Listing 4.10: Query all data sets in the metadata repository

```
select *
where{
  ?md qb:dataSet ?ds.
  ?md qb:component ?cp.
  ?cp rdf:type ?t.
  ?cp rdf:value ?v.
  ?ds rdf:value ?v.
  FILTER(?md = <http://statspace.linkedwidgets.org/metadata/WorldBank-SP.POP.TOTL>)
}
```

Listing 4.11: Query information concerning a specific metadata description

[16] https://datahub.io/dataset/statspace, accessed December 30, 2016
[17] https://datahub.io, accessed December 30, 2016
[18] http://ogd.ifs.tuwien.ac.at/sparql, accessed December 30, 2016

4.1.5 Mediator

Mediator is a service developed to support advanced users in exploring StatSpace through SPARQL queries. Figure 4.3 presents a prototype of the *mediator* that can be accessed at `http://statspace.linkedwidgets.org/mediator/`. We support three different output formats for a query including HTML, JSON, and XML. Furthermore, users can choose either cached RDF data sets in our server or completely new RDF data sets that will be generated by the *RML mapping service* if the user's query relates to raw data sets.

Query

```
PREFIX qb:    <http://purl.org/linked-data/cube#>
PREFIX sdmx-dimension:   <http://purl.org/linked-data/sdmx/2009/dimension#>
PREFIX sdmx-measure:     <http://purl.org/linked-data/sdmx/2009/measure#>
PREFIX sdmx-attribute:   <http://purl.org/linked-data/sdmx/2009/attribute#>
SELECT *
WHERE {
      ?ds dc:subject  <http://statspace.linkedwidgets.org/codelist/cl_subject/SP.POP.TOTL>.
      ?o qb:dataSet ?ds.
      ?o sdmx-measure:obsValue  ?obsValue.
      ?o sdmx-dimension:refPeriod ?refPeriod.
      ?o sdmx-dimension:refArea   ?refArea.
      ?o sdmx-attribute:unitMeasure ?unit.
      Filter(?refArea= <http://statspace.linkedwidgets.org/codelist/cl_area/UnitedKingdom>)
}
```

Result Formats HTML ▾ Use cache

Figure 4.3: Query interface of the mediator

To provide a uniform view of data for users, the *mediator* relies on equivalent mappings between URIs defined in individual data sets and shared URIs. Providing mappings that the *mediator* used for data integration is necessary to allow users to evaluate the final result. Figure 4.4 shows an excerpt of the equivalent relationship used for the example query in Figure 4.3.

Equivalent relationships used

Shared URIs	Particular URIs used in the dataset
http://purl.org/linked-data/sdmx/2009/dimension#refArea	http://purl.org/linked-data/sdmx/2009/dimension#refArea
http://statspace.linkedwidgets.org/codelist/cl_area/UnitedKingdom	http://dd.eionet.europa.eu/vocabulary/worldbank/country/GB
http://purl.org/linked-data/sdmx/2009/dimension#refPeriod	http://purl.org/linked-data/sdmx/2009/dimension#timePeriod
http://reference.data.gov.uk/id/gregorian-year/2010	2010-12-31^^http://www.w3.org/2001/XMLSchema#date

Figure 4.4: An example of equivalent relationships used by the mediator

4.1.6 Explorer

To support non-expert users who do not have knowledge of semantic technologies in exploring StatSpace, we provide a prototype explorer at `http://statspace.linkedwidgets.org/explorer`. This web application provides functionalities for faceted search, metadata description discovery, and data integration.

Faceted search. End users can enter search keywords such as *"Austria 2010 GDP"* to explore data sets in the *metadata repository*. When receiving a user's request, the *explorer* first will classify the keywords into three parts including: (i) a list of geographical areas, e.g., "Austria". To identify these areas, we make use of DBPedia spotlight [63] – a tool for named entity recognition; (ii) temporal values, e.g., "2010". These values are recognized based on the use of temporal patterns listed in the temporal value mapping algorithm (cf. Algorithm 2); and (iii) the remaining keywords, e.g., "GDP" that typically refer to labels of data sets. Next, the *explorer* queries the *metadata repository* to find matching data sets with three described requirements. Metadata descriptions of resulting data sets will be provided to users including information of data publishers, subjects, and labels. Users can filter the list of metadata through names of providers and subjects of data sets. Figure 4.5 presents the search interface of the *explorer*.

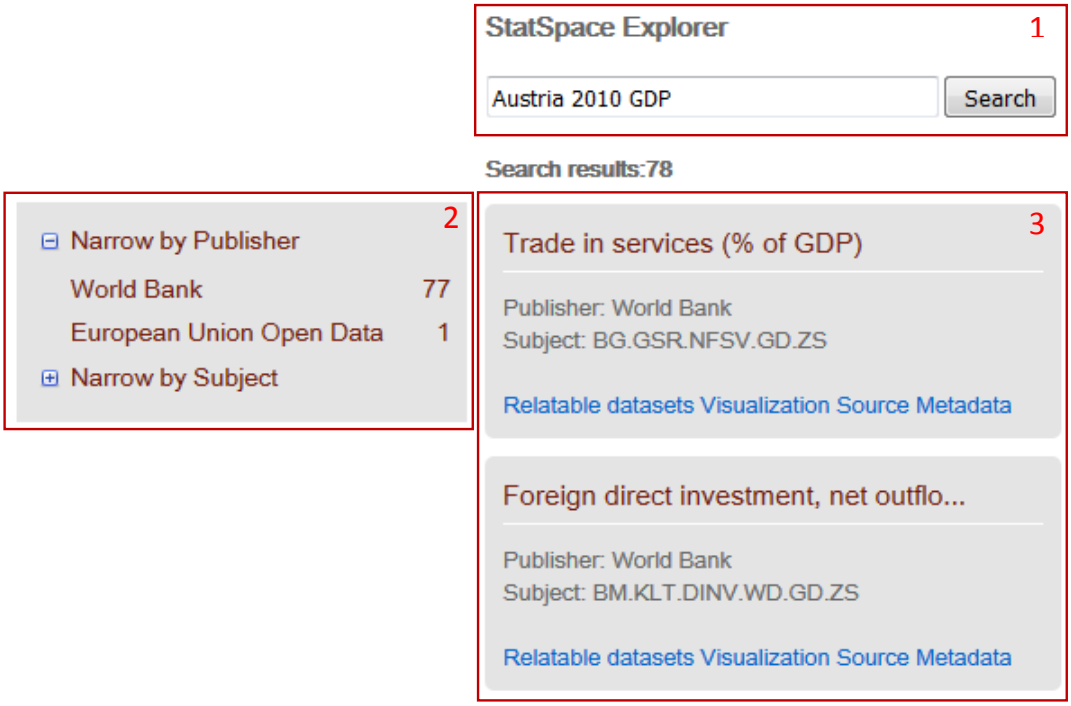


Figure 4.5: The explorer interface: Search field (1); Filters (2); List of results (3)

Metadata descriptions discovery. We provide four options for users to discover the metadata description of each data set.

(i) *Identification of relatable data sets.* For a given data set, we identify a list of relatable data sets and display metadata descriptions of these data sets to users. Two data sets are relatable if their data structures can be mapped to each other. The details of this comparison are described in Section 4.2.1 (Requirements for data integration).

(ii) *Visualization of the data set.* The *explorer* provides data visualization for all statistical data sets described in the *metadata repository* including raw data sets and RDF data sets. For RDF data sets, their access mechanism relies on SPARQL, that allows advanced users to quickly and easily query data. We make use of two approaches for data visualization of data sets: (i) for raw data sets, we first use the *RML mapping service* to transform these data sets into RDF format. The resulting data sets are then returned to the client browser in JSON format. Based on users' filter conditions, JQuery functions will be used to identify relevant data. Next, we use the C3js[19] library to provide charts to users. Figure 4.6 illustrates data visualization of a raw data set; (ii) for RDF data sets, we generate a SPARQL query whenever a user changes filter conditions. We model and expose each RDF data set in a representation named *widget*. Each widget has two functionalities, i.e., *query building* and *chart generation*. *Query building* means that end users can establish filter conditions via an interactive interface. Next, the widget formulates a corresponding query and sends it to the SPARQL endpoint. Based on the returned result, the widget will generate *different charts* such as column, line, pie, bubble, and geo map charts[20] to provide meaningful views on the data set. Figure 4.7 shows a sample widget. We provide a prototype application for widget generation at `http://statspace.linkedwidgets.org/generation/`.

(iii) *Access to the original data source.* Based on the information of the address of the data set that is stored in the metadata description, we direct users to the original data source that may be a SPARQL endpoint or a data portal.

(iv) *Visualization of metadata.* Figure 4.8 present the interface for visualization of a metadata description via Pubby[21].

Data Integration. The *explorer* allows users to integrate data from two relatable data sets. Similarity to the *mediator*, the *explorer* relies on coreference relationships to rewrite individual results to a consolidated result. Next, it generates charts to visualize the final result to users. Figure 4.9 shows an example of data integration.

[19] `http://c3js.org/`, accessed December 30, 2016
[20] `https://developers.google.com/chart/interactive/docs/gallery`
[21] `http://wifo5-03.informatik.uni-mannheim.de/pubby/`, accessed December 30, 2016

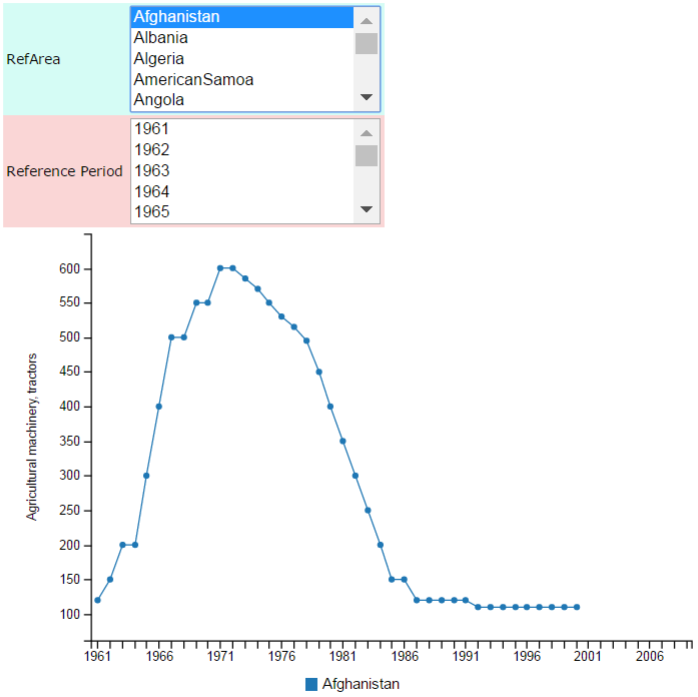


Figure 4.6: Visualization of a raw data set

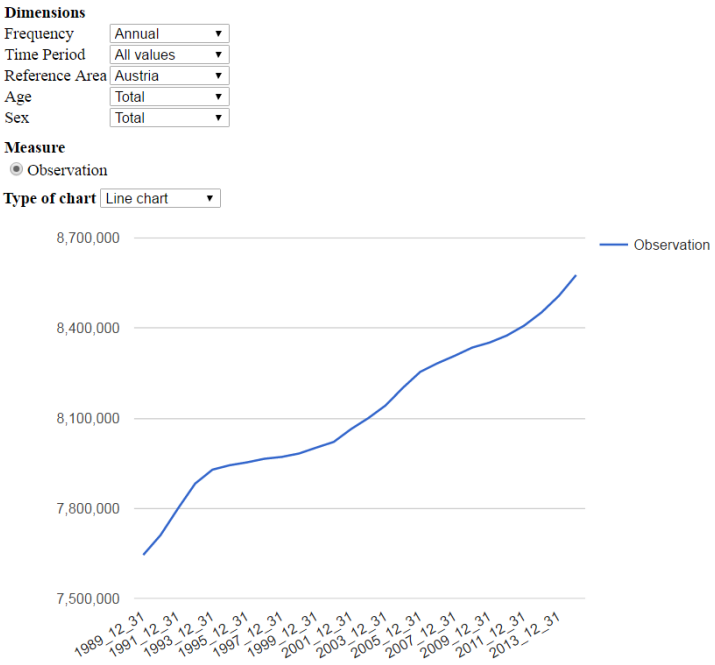


Figure 4.7: Visualization of an RDF data set

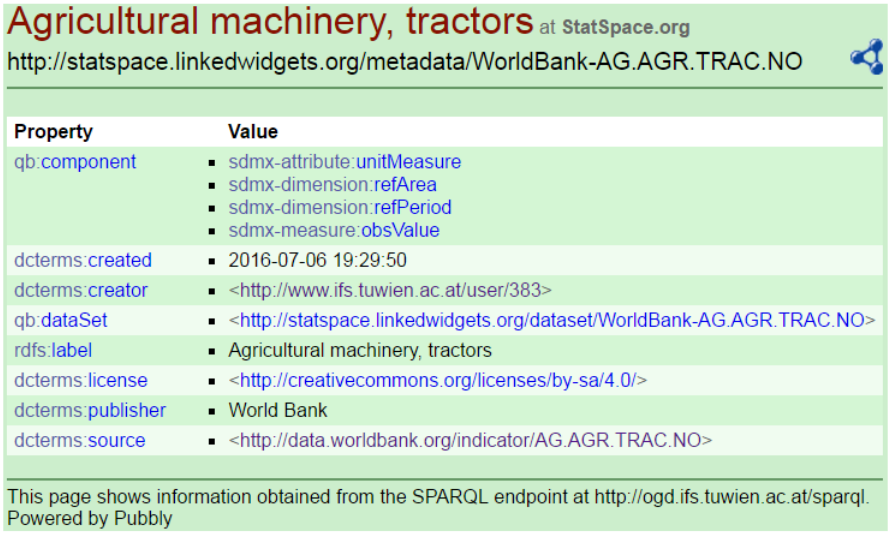


Figure 4.8: Interface for visualization of a metadata description

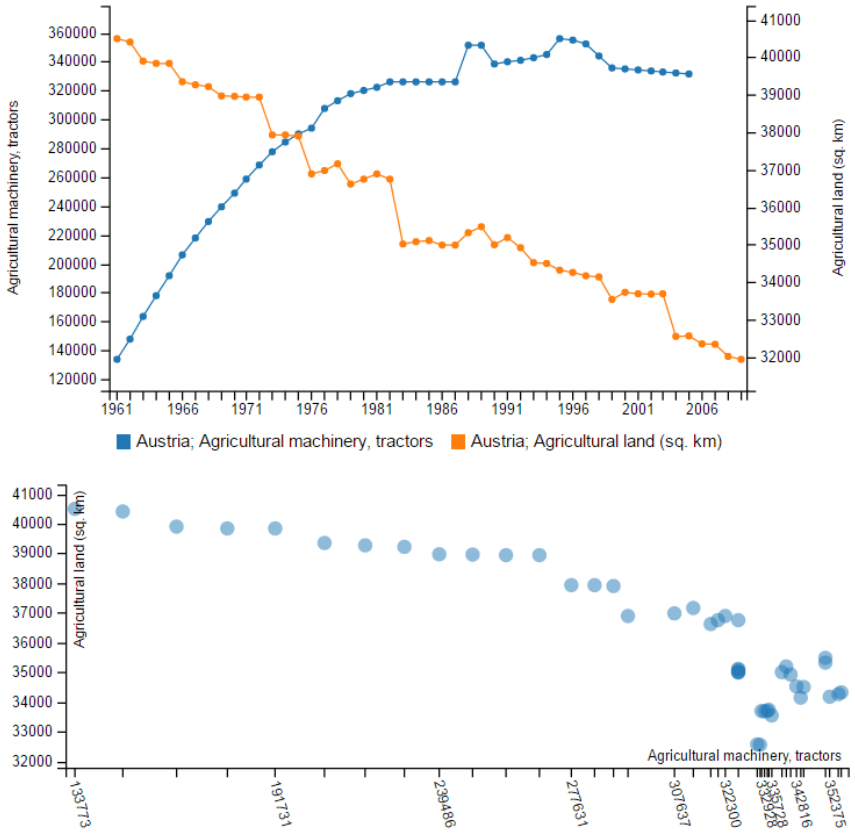


Figure 4.9: Visualization used in data integration

4.2 Example Use Cases

4.2.1 Statistical Data Integration

To illustrate statistical data integration, we present an example of *"comparing the population of the UK"* based on three data sources, each using different formats, structures, and access mechanisms. These data sources are: (i) the ONS[22], (ii) the World Bank, and (iii) the EEA. The ONS data source is published in Excel spreadsheet format. It includes population data[23] from 1964 to 2013. The World Bank provides population data[24] via Application Programming Interfaces (APIs). Finally, the EEA data source provides population data[25] in RDF format, which is organized by criteria such as time, sex, and age group. Table 4.11, Table 4.12, and Listing 4.12 show excerpts from the data sets.

Mid-Year	Mid-Year Population (millions)	Annual Percentage Change
2011	63.3	0.84
2012	63.7	0.66
2013	64.1	0.63

Table 4.11: UK population data in spreadsheet format (excerpt)

```
<wb:data>
  <wb:indicator id="SP.POP.TOTL">
      Population, total
  </wb:indicator>
  <wb:country id="GB">
    United Kingdom
  </wb:country>
  <wb:date>2014</wb:date>
  <wb:value>64,510,376</wb:value>
  <wb:decimal>1</wb:decimal>
</wb:data>
```

Listing 4.12: World Bank population data for the UK in XML format (excerpt)

sdmxd: freq	sdmxd: timePeriod	sdmxd: refArea	sdmxd:age	sdmxd: sex	sdmxm: obsValue
sdmx-code:freqA	2014	geo:UK	ag:TOTAL	sdmxcode: sex-T	64,308,261
sdmx-code:freqA	2014	geo:UK	ag:Y_LT15	sdmxcode: sex-T	11,333,471
sdmx-code:freqA	2014	geo:UK	ag:Y15–64	sdmxcode: sex-F	20,929,655

Table 4.12: European population data in RDF format (excerpt)

[22] http://www.ons.gov.uk/, accessed December 30, 2016

[23] http://www.ons.gov.uk/ons/rel/pop-estimate/population-estimates-for-uk--england-and-wales--scotland-and-northern-ireland/2013/chd-1-for-story.xls, accessed December 30, 2016

[24] http://api.worldbank.org/countries/GB/indicators/SP.POP.TOTL, accessed December 30, 2016

[25] http://rdfdata.eionet.europa.eu/eurostat/data/demo_pjanbroad, accessed December 30, 2016

Requirements for Statistical Data Integration

Because any multi-measure data set (e.g., the UK population data set in Table 4.11) can be split into multiple single-measure data sets [16], without loss of generality, we focus on data integration requirements for single-measure data sets. In the *metadata repository*, the data structure of each data set is represented by a list of dimensions, a measure, and a unit. Although the transformation of the unit attribute is necessary if different units (e.g., kilometer vs. meter) or different scales (absolute number vs. millions) are used, this component does not affect data integration. We can still integrate two statistical data sets using completely different units (e.g., USD and percentage) in order to identify the correlation between them. For example, a combination of two economic indicators including GDP (unit is USD) and inflation rate (unit is percentage) has been studied in many papers [64, 65, 66]. In the following, we describe data integration requirements in terms of dimensions used.

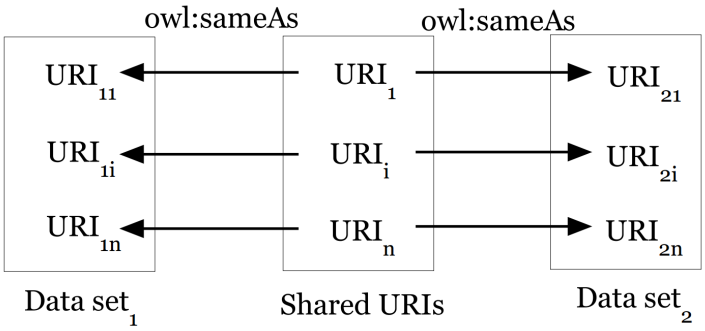
Assume that $L_1 = \{URI_{11}, \ldots, URI_{1n}\}$ and $L_2 = \{URI_{21}, \ldots, URI_{2m}\}$ are two lists of dimensions used in two single-measure data sets. Because data sets can make use of different URIs to represent dimensions and their values, we identify requirements based on coreference relationships. There are two cases:

(i) *They have the same set of dimensions.* In this case, two lists must have the same number of components (i.e., $n = m$). In addition, each URI in a list can be mapped to only one URI in the other list. That means, *for each* $URI_{1i} \in L_1$ if the URI_i is its equivalent URI then we should have $URI_{2j} \in L_2$ that URI_i is also equivalent URI of this URI. For instance, we can compare population figures of the ONS and World Bank data sets, because these data sets use the same set of dimensions including one spatial dimension and one temporal dimension. Figure 4.10a provides an illustration of data integration requirement in the first case.

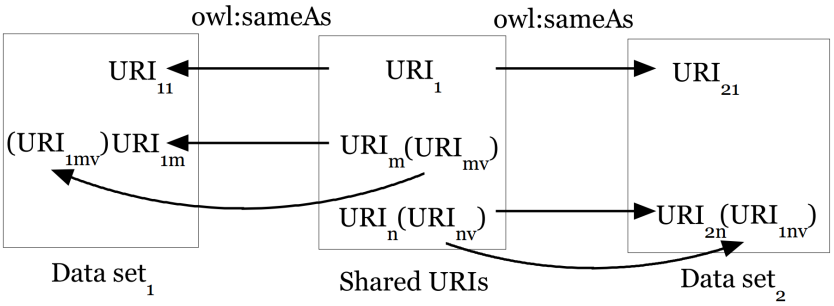
(ii) *They have different sets of dimensions.* We use code lists of dimensions to assign values at top concept levels to dimensions that do not appear in both data sets. For instance, to integrate the EC and World Bank data sets, we assign *ag:TOTAL* to the *age* dimension, *sdmxcode:sex-T* to the *sex* dimension, and *sdmxcode:freqA* to the *frequency* dimension because these dimensions are not available in the World Bank data set. Figure 4.10b explains the data integration requirement in the second case. In this figure, URI_{mv} and URI_{nv} are two values at the top concepts in the code lists of dimensions URI_m and URI_n, respectively. In addition, URI_{1mv} and URI_{2nv} are two respectively specific values used in data sets.

Mediator-based Data Integration

To compare population figures for the UK in different data setsc, users can create a SPARQL query and then invoke the *mediator* to perform this query. Listing 4.13 depicts an example query containing conditions of the subject (Population), data structure, and



(a) Two data sets have the same set of dimensions



(b) Two data sets have the different sets of dimensions

Figure 4.10: Illustrations of requirements for data integration

the value of the spatial dimension (United Kingdom). The *mediator* identifies seven suitable data sets in the *metadata repository*. Next, it rewrites the input query into specific queries for individual data sets. Listings 4.14, 4.15 and 4.16 show three queries used to obtain data from the data sets listed in Table 4.11, Table 4.12, and Listing 4.12. Figure 4.11 presents the result in HTML for the input query. We use different colours to highlight each data set in order to support users in distinguishing them. The figure shows that each data set gives a different population figure for the UK in 2010.

```
SELECT *
WHERE {
 ?ds dc:subject   <http://statspace.linkedwidgets.org/codelist/cl_subject/SP.POP.TOTL>.
 ?o qb:dataSet ?ds.
 ?o sdmxm:obsValue  ?obsValue.
 ?o sdmxd:refPeriod ?refPeriod.
 ?o sdmxd:refArea   ?refArea.
 ?o sdmxa:unitMeasure ?unit.
 FILTER (?refArea= <http://statspace.linkedwidgets.org/codelist/cl_area/UnitedKingdom>)
}
```

Listing 4.13: Example input query for cross-data set population comparison

```
SELECT * WHERE {
 ?o qb:dataSet <http://rdfdata.eionet.europa.eu/eurostat/data/demo_pjanbroad>.
 ?o sdmxm:obsValue ?obsValue.
 ?o sdmxd:timePeriod ?timePeriod.
 ?o sdmxd:refArea ?refArea.
 ?o sdmxa:unitMeasure ?unit.
 ?o sdmxd:freq <http://purl.org/linked-data/sdmx/2009/code#freq-A>.
 ?o sdmxd:age <http://dd.eionet.europa.eu/vocabulary/eurostat/age/TOTAL>.
 ?o sdmxd:sex  <http://purl.org/linked-data/sdmx/2009/code#sex-T>.
 FILTER(?refArea= <http://dd.eionet.europa.eu/vocabulary/eurostat/geo/UK>)
```

Listing 4.14: EEA data set query generated by the mediator

```
http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/
    mapping/wb.ttl&indicator=SP.POP.TOTL&refArea=GB
```

Listing 4.15: World Bank data set query generated by the mediator

```
http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/
    mapping/uk7.ttl
```

Listing 4.16: UK data set query generated by the mediator

Number of matching datasets: 7

http://rdfdata.eionet.europa.eu/worldbank/dataset/wdi/SP.POP.TOTL	http://reference.data.gov.uk/id/gregorian-year/2010	62641000
http://rdfdata.eionet.europa.eu/eurostat/data/tps00001	http://reference.data.gov.uk/id/gregorian-year/2010	62510197
http://rdfdata.eionet.europa.eu/eurostat/data/demo_pjanbroad	http://reference.data.gov.uk/id/gregorian-year/2010	62510197
http://rdfdata.eionet.europa.eu/who/data/WHS9_86	http://reference.data.gov.uk/id/gregorian-year/2010	62066000.0
http://statspace.linkedwidgets.org/dataset/ONS-Population-change	http://reference.data.gov.uk/id/gregorian-year/2010	62800000.0
http://statspace.linkedwidgets.org/dataset/WorldBank-SP.POP.TOTL	http://reference.data.gov.uk/id/gregorian-year/2010	62766365
http://statspace.linkedwidgets.org/dataset/ONS-Population-1851-2014	http://reference.data.gov.uk/id/gregorian-year/2010	627594560.0

Figure 4.11: Result of the example query

4.2.2 Data Quality Assessment

Data quality assessment is defined as the process of identifying the ability of data to satisfy users' needs [67, 68, 69]. It is also considered to be the process of finding out

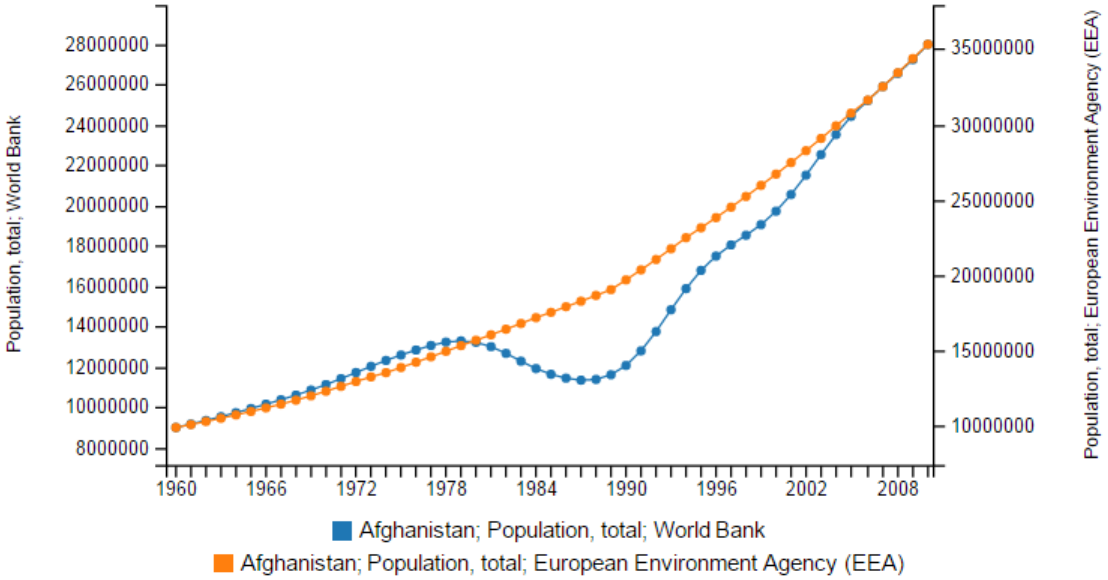


Figure 4.12: An example of out-of-date data of the EEA

data errors and estimating their impacts on services and business processes [70]. The *metadata repository* in StatSpace contains descriptions of each data set such as label and values of dimensions. Therefore, it can be used to evaluate data accuracy.

Data accuracy refers to the discrepancy of observed values in different data sets. In particular, we can focus on the comparison between raw data sets and RDF data sets because these data sets can originate from the same data source. For example, the EEA transforms a raw data set[26] of the World Bank into RDF format and then stores it[27] in a SPARQL endpoint[28]. However, when the original data source (e.g, the World Bank) updates its data, relevant sources (e.g., EEA) may not recognize this update to make the necessary changes. Figure 4.12 shows an inconsistency of population data and fluctuations for Afghanistan during the period from 1960 to 2010.

4.2.3 Correlation Mining

Correlation mining is a task in data mining [71] that aims to discover interesting hidden dependencies between variables from large amounts of data [72, 73]. StatSpace can support users in this task because it provides a single point of access to a large number of heterogeneous data sets. In addition, based on semantics of data represented by subjects, data structures, and coreference relationships, StatSpace can identify relatable data sets

[26] `http://api.worldbank.org/en/countries/all/indicators/SP.POP.TOTL?format=xml`, accessed December 30, 2016

[27] `http://rdfdata.eionet.europa.eu/worldbank/dataset/wdi/SP.POP.TOTL`, accessed December 30, 2016

[28] `http://semantic.eea.europa.eu/sparql`, accessed December 30, 2016

for an arbitrary data set. As a result, it can improve the *feature selection step* used in existing statistical data mining applications [74, 75].

At present, StatSpace provides a preliminary implementation of correlation mining through relatable data sets identification and data visualization. As an example use case, consider the common macroeconomic topic of the relationship between inflation and unemployment. To investigate this relationship, users can explore the *metadata repository* and retrieve the inflation and unemployment figures of various countries and from various sources. The provided charts, e.g., scatter plot and line chart, visualize the development and relations between the individual indicators. Figure 4.13 shows the resulting plot for Japan during the period from 1991 to 2014, which exhibits the expected relationship [76], i.e., a negative correlation between unemployment and inflation.[29]

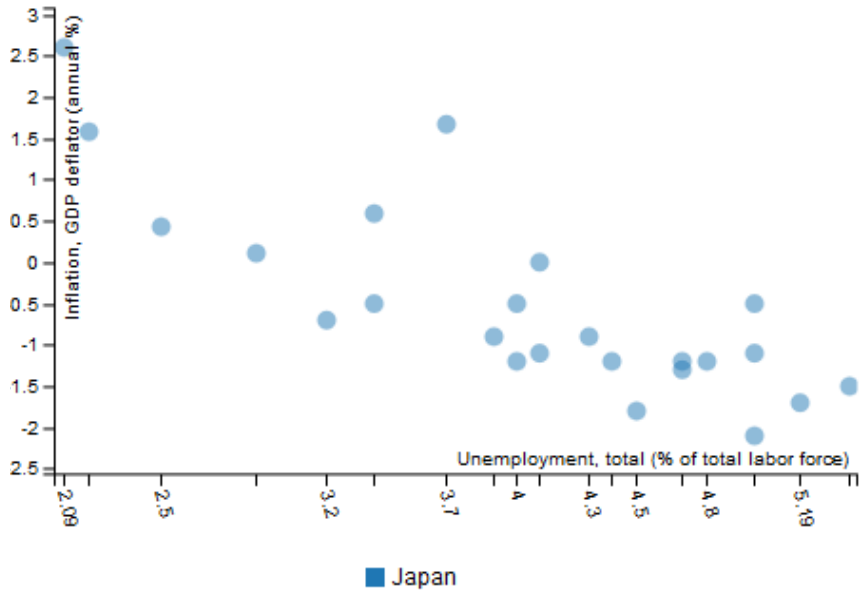


Figure 4.13: Relationship between Inflation and Unemployment indicators for Japan

4.2.4 Spatial Data Visualization

All StatSpace resources are available on the web. Therefore, developers can use these resources to support users in exploring statistical data in various ways. In this section, we introduce two examples that we developed in an existing platform outside StatSpace.

In metadata descriptions, spatial URIs of data sets are linked to shared URIs, hence, users are able to explore data for the same geographical area in different data sets. To illustrate this capability, we introduce the following two use cases:

[29] http://statspace.linkedwidgets.org/compareDataSet?&id1=http://statspace.linkedwidgets.org/metadata/WorldBank-SL.UEM.TOTL.ZS&id2=http://statspace.linkedwidgets.org/metadata/WorldBank-NY.GDP.DEFL.KD.ZG, accessed December 30, 2016

(i) The first use case relates to exploration of a geographical area. Users can provide an address, e.g., *Donaufelder Strasse 54, Austria*, or simply define a point on a map, for which we identify the corresponding administrative areas, e.g., *Country: Austria*, *City: Vienna*, and *District: Floridsdorf*. Next, we query the *metadata repository* to find relevant data sets. Finally, we provide data visualization for each data set based on the use of data visualization functionality of the *explorer*.

(ii) The second use case allows users to compare data of different areas. First, users choose multiple areas on a map, e.g., Germany and Austria. Next, we identify data sets that contain data of both countries and provide data visualization to users.

We implement the example use cases based on the use of an existing platform outside StatSpace. The key element of this platform is the so-called linked widget [77, 78, 79], which represents an extension of a standard web widget backed by a semantic model. This model describes data input/output and the information of data provenance and licensing terms. In this platform, linked widgets are grouped into widget collections. Each collection addresses a different problem domain. We developed a collection[30] for spatial data exploration, which consists of the following three widgets.

(i) *Spatial Entity Recognizer.* This widget receives an address text or a user-defined location as its input and uses Google's geocoding APIs to obtain corresponding spatial entities at different levels, e.g., country level, or administrative area levels.

(ii) *Spatial Data Locator.* This widget queries the *metadata repository* to identify a list of matching data sets to the input entities. It contains an option to filter the data sets based on their labels.

(iii) *Spatial Data Visualization.* This widget provides the visualization for chosen area(s).

Sample visualizations created from theses widgets are shown in Figures 4.14 and 4.15: (i) discovery of statistical information on an area based on a user-provided location[31]; and (ii) comparison of geographical areas[32].

[30] http://linkedwidgets.org/MashupPlatform.html?widgetCollectionId=SpatialStatisticalCollection

[31] http://linkedwidgets.org?id=MashupSpatialDataLocator

[32] http://linkedwidgets.org?id=MashupSpatialDataComparator

Figure 4.14: Exploration of a geographical area

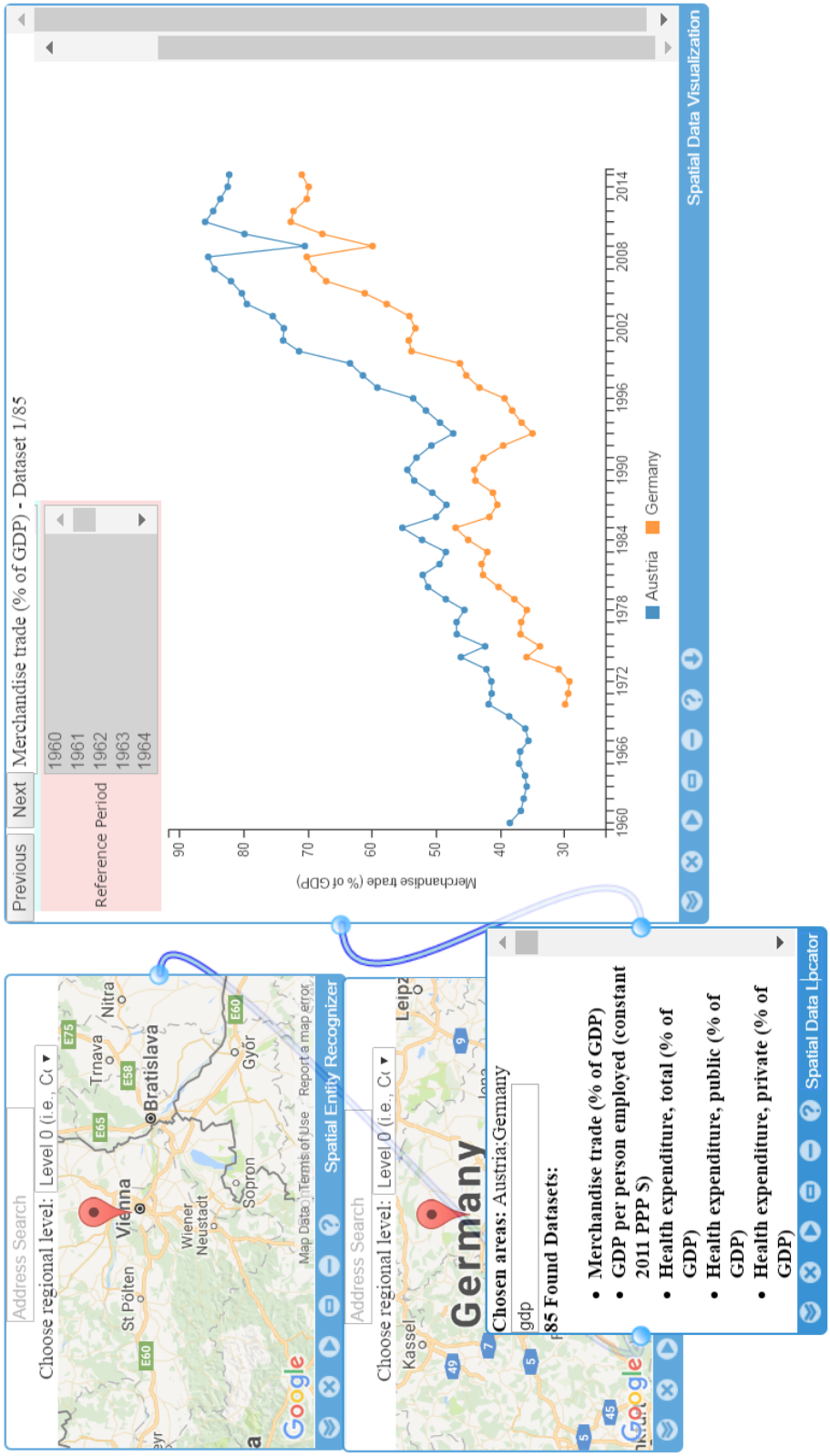


Figure 4.15: Statistical data comparison of different areas

CHAPTER 5

Evaluation

In this chapter, we report the evaluation results of four components in StatSpace including the *RML mapping service*, *URI design patterns*, the *metadata repository*, and the *metadata generator*. First, in Section 5.1 we present an experiment to evaluate the runtime of the *RML mapping service* in raw data transformation into RDF. This experiment is performed on a personal computer with the configuration as follows: Operation System - Windows 7 Enterprise 64 bit, Ram - 8.00 GB of DDR3, Processor - Intel Core i5-3470 CPU@3.20 GHz. The computer's internet connection speed is around 800 Mbps for download and 900 Mbps for upload (tested by `http://www.speedtest.net/`). Next, we analyze the coverage of *URI design patterns* and *metadata repository* in Sections 5.2 and 5.3, respectively. Finally, Section 5.4 provides users' evaluation results about equivalent relationships identified by the *metadata generator*.

5.1 RML Mapping Service

One of the most demanding tasks in StatSpace is the transformation of raw data sets into RDF at query time. Therefore, we focus on evaluating the *RML mapping service* in terms of time consumption. The performance of this service depends on four main factors including: (i) the internet connection speed for downloading raw data sets; (ii) the number of requests that servers where the data sets are stored receive at the time of the experiment; (iii) the speed of the RML processor in analyzing the mappings and generating RDF triples; and (iv) the format and size of the input data set. In these factors, the first two factors are not stable and are hard to predict with accuracy. In order to alleviate this issue, we needed to repeat the same experiment at different times.

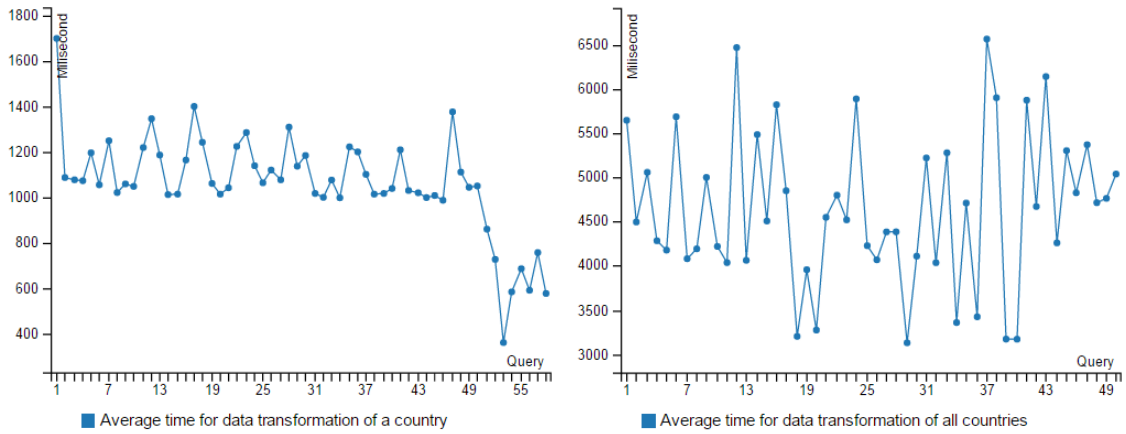
We randomly generated 50 queries that called upon the *RML mapping service* to transform raw data of a single country from the World Bank data source into RDF and 50 queries that invoked this service to transform the raw data of all countries into RDF. In addition, we used eight queries that transform raw data of the ONS into RDF. We calculate the

time between the invocation of the *RML mapping service* and the generation of the corresponding output RDF data set. Furthermore, we repeated the whole experiment at three different times to provide a more comprehensive view of the performance of the *RML mapping service*. Table 5.1 provides an overview of characteristics of these queries. The full list of queries and results of evaluation are available in the Appendixes D – E.

Query type	**Data source**	**Data format**	**Size of data set**
One country	World Bank	JSON	8 – 10 KB
	ONS	XLS	35 – 56 KB
All countries	World Bank	JSON	2.4 – 3 MB

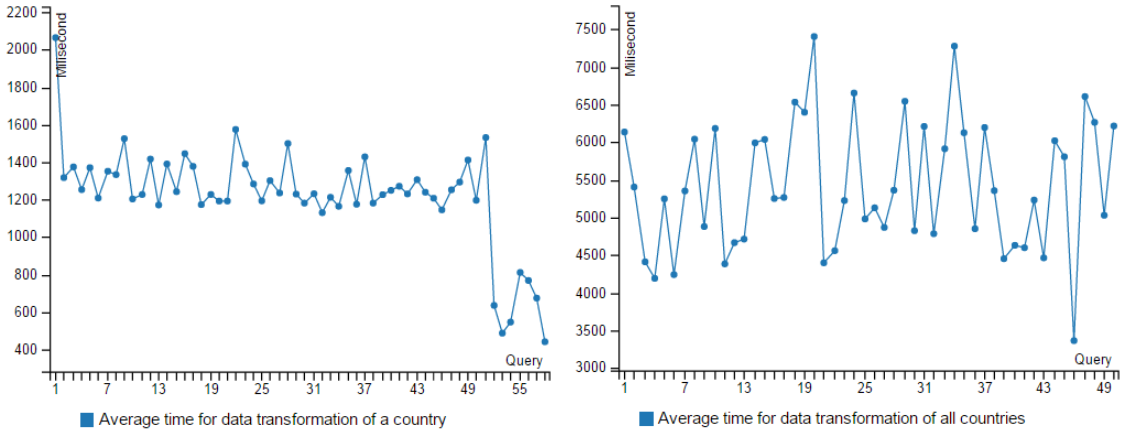
Table 5.1: Characteristics of queries

Figures 5.1 – 5.3 show the time needed by the *RML mapping service* for each query in three experiments. In figures 5.1a – 5.3a, we see that the data transformation for a single country from the World Bank data source into RDF typically takes between 1 and 2 seconds. For the ONS data source (queries 51 to 58), the *RML mapping service* is significantly faster. Furthermore, due to unpredicted and outside effects from data source servers and the quality of the internet connection, the difference of the time needed for data transformation of the same query at different experiments is up 0.3 seconds. Finally, figures 5.1b – 5.3b show that data transformation for all countries from the World Bank needs from 3.1 to 7.4 seconds. In addition, the disparity of time for data transformation of the same query increases to 4.1 seconds.



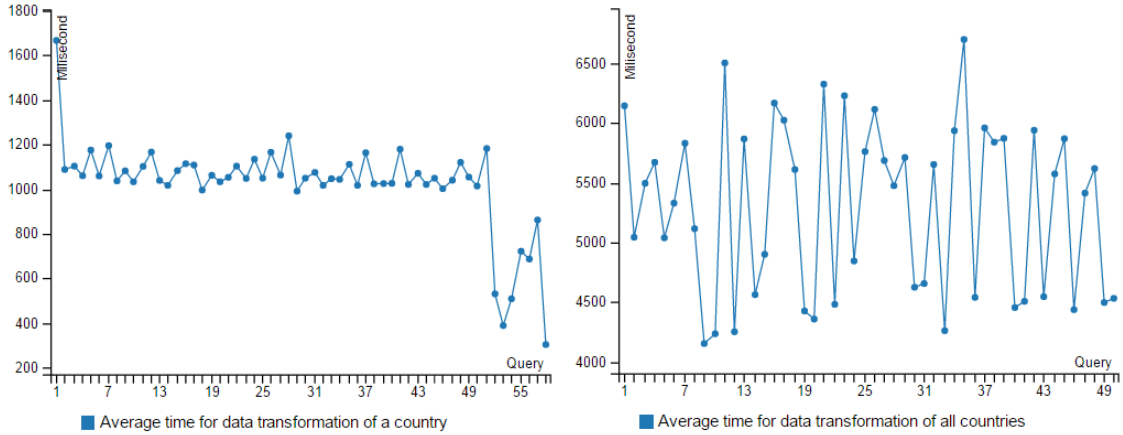
(a) Data transformation of one country (b) Data transformation of all countries

Figure 5.1: Elapsed time for raw data transformation in the first experiment



(a) Data transformation of one country (b) Data transformation of all countries

Figure 5.2: Elapsed time for raw data transformation in the second experiment



(a) Data transformation of one country (b) Data transformation of all countries

Figure 5.3: Elapsed time for raw data transformation in the third experiment

5.2 URI Design Patterns

Table 5.2 shows a comparison of the size between seven code lists used in some existing data sources and the respective code lists coined from our patterns. These code lists are selected for evaluation because they appear in most statistical data sources. Therefore, through this comparison, we can recognize the benefits of identifying common code lists. At present, each statistical data source organizes data by many criteria such as geographical area, time, age, sex, etc. The diversity becomes more complex when each criterion can be attached to a large number of potential values. For example, the World Bank provides statistical data about more than 300 countries and regions. The number of European areas in the EEA is over 3,100 areas, and in the CSO more than 33,200

areas in the United Kingdom are shown. Similarity, for temporal dimension, the number of times represented in the EEA approaches 609 values. Dealing with this issue, our code lists not only cover all values used in existing data sources but also can harmonize different values to a common value. As a result, users can integrate statistical data of the same area from different data sources.

Source	**|Area|**	**|Time|**	**|Age|**	**|Sex|**	**|Frequency|**	**|Occupation|**
EUODP	31	12	0	0	0	0
EEA	3,114	609	132	0	3	0
ScotStat	16,718	167	83	3	0	0
CSO	4,806	1	40	3	0	11
ODC	33,286	167	0	0	4	0
VOGD	2,542	823	113	3	0	0
World Bank	304	57	–	–	–	–
URI Design Patterns	Unlimited	Unlimited	209	5	9	619

Table 5.2: A comparison of size of code lists in data sources

5.3 Metadata Repository

Table 5.3 provides an overview of the coverage of the *metadata repository* w.r.t. data sources and the number of data sets. This component contains descriptions of eight data sources. The total number of metadata descriptions in the *metadata repository* is 2,060 including 1,459 descriptions of raw data sets and 601 descriptions of data sets published in diverse SPARQL endpoints by six publishers.

Source	**Data format**	**|Data sets used|**	**|Metadata descriptions|**	**Metadata size (KB)**
EUODP	RDF	151	151	826
EEA	RDF	147	147	3,331
CSO	RDF	61	61	413
ScotStat	RDF	23	23	428
ODC	RDF	3	3	268
VOGD	RDF	39	214	2,385
ONS	XLS	2	8	103
World Bank	JSON	1,451	1,451	22,045

Table 5.3: Sources and numbers of data sets covered

We define RML mappings to transform raw statistical data sets into RDF. Because all World Bank data sets share the same structure, we can use a single RML mapping[1] for their transformation. For the ONS data source, the select data sets are published in spreadsheet format. Each sheet refers to different data content. Therefore, we have to define separate RML mappings for each sheet in these data sets.

Data sets published via SPARQL endpoints typically contain highly heterogeneous data which differ in their components and code lists. For example, the ScotStat and ODC endpoints provide statistical data for more than 16,000 areas. This explains the smaller number of RDF data sets in the *metadata repository* compared to raw data sets.

5.4 Metadata Generator

5.4.1 Mapping Endorsement

In StatSpace, isolated data sets are linked to each other based on mappings from a list of shared URIs to particular URIs used in data sets. To ensure the correctness of mappings, we perform two steps: (i) mapping algorithms identify equivalent relationships between the shared URIs and special URIs used in each data set, and then write these mappings to text files; (ii) we check the mappings and manually correct incorrect mappings. Next, only mappings verified by us are used to generate metadata descriptions for data sets. In the following, we introduce our evaluation system and a small-scale user study of identified mappings.

The evaluation system is integrated with our coreference resolution service at `http://statspace.linkedwidgets.org/sameas/`. Figure 5.4 shows the interface of the coreference resolution service. For each mapping, we provide three options for users including "Agree", "Disagree", and "I don't know". The evaluation results are stored in a MySQL database. In the future, this database can be used for two purposes: (i) it allows us to revise provided mappings to edit mappings that most users do not agree with, and (ii) we can integrate each user's evaluation results with the *mediator* and *explorer*. As a result, only mappings accepted by the user are used for data integration.

We also conducted a user study of identified mappings for the EEA. We randomly selected 25 mappings for the spatial dimension and 25 mappings for the temporal dimension. Next, we sent this list[2] (shown in Figure 5.5) to five experts in our institute. For each mapping, we suggested experts use descriptions of two involved URIs, such as labels, types, links to other URIs, etc. to determine the correctness of the mapping. The evaluation results showed that all experts agree that the correctness of 48 mappings from a total of 50. One spatial value mapping and one temporal value mapping were not agreed by all five experts. In the following, we describe these two mappings.

[1] `http://statspace.linkedwidgets.org/mapping/wb.ttl`, accessed December 30, 2016

[2] `http://statspace.linkedwidgets.org/sameas/userstudy.html`, accessed December 30, 2016

- The mapping between `http://dd.eionet.europa.eu/vocabulary/eurostat/geo/NL213` (label: Twente) and `http://statspace.linkedwidgets.org/codelist/cl_area/Netherlands/Overijssel/HofvanTwente` (label: Hof van Twente) received one vote "Disagree" and two votes "I dont know" from experts. Our wrong mapping originated from the incorrect result that Google's geocoding service provided. Receiving the label "Twente, Overijssel" as input, this service returns spatial information for "Hof van Twente, Overijseel". However, "Hof van Twente" is just a municipality of "Twente" region.

- The mapping between `2007^http://www.w3.org/2001/XMLSchema#int` and `http://reference.data.gov.uk/id/gregorian-year/2007` received one vote "I dont know". The expert argues that the first URI does not contain label and semantic information. Therefore, he can not make the decision.

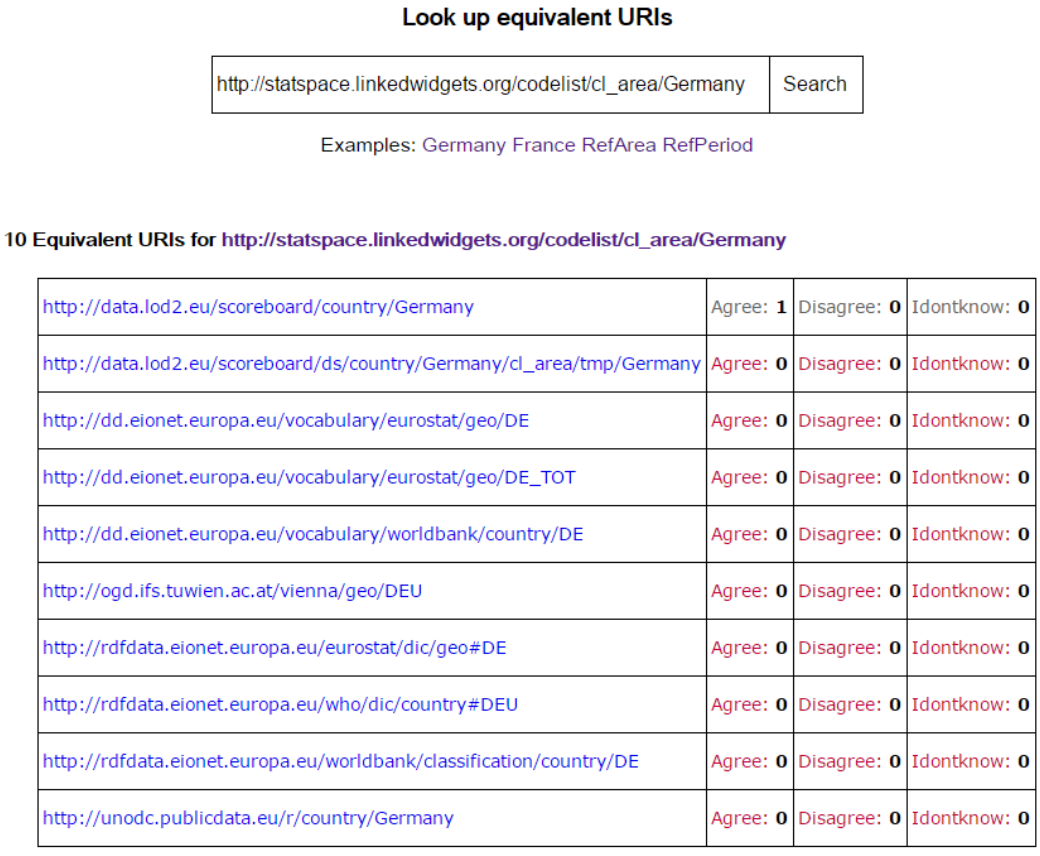
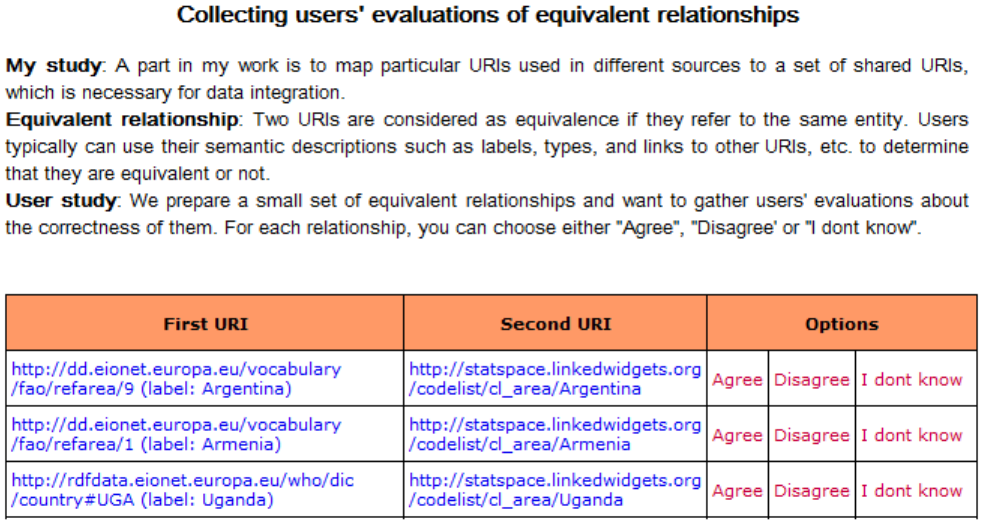


Figure 5.4: The interface of the coreference resolution service



Collecting users' evaluations of equivalent relationships

My study: A part in my work is to map particular URIs used in different sources to a set of shared URIs, which is necessary for data integration.
Equivalent relationship: Two URIs are considered as equivalence if they refer to the same entity. Users typically can use their semantic descriptions such as labels, types, and links to other URIs, etc. to determine that they are equivalent or not.
User study: We prepare a small set of equivalent relationships and want to gather users' evaluations about the correctness of them. For each relationship, you can choose either "Agree", "Disagree' or "I dont know".

First URI	Second URI	Options		
http://dd.eionet.europa.eu/vocabulary /fao/refarea/9 (label: Argentina)	http://statspace.linkedwidgets.org /codelist/cl_area/Argentina	Agree	Disagree	I dont know
http://dd.eionet.europa.eu/vocabulary /fao/refarea/1 (label: Armenia)	http://statspace.linkedwidgets.org /codelist/cl_area/Armenia	Agree	Disagree	I dont know
http://rdfdata.eionet.europa.eu/who/dic /country#UGA (label: Uganda)	http://statspace.linkedwidgets.org /codelist/cl_area/Uganda	Agree	Disagree	I dont know

Figure 5.5: The interface of application used to collect users' evaluations

5.4.2 Limitation of Endorsement

Data providers can use different methods to collect their data. This inhibits direct comparison of the data. At present, StatSpace does not model the data collection methodology used by the various sources and it does not consider complex problems in the spatio-temporal mapping domain (e.g., territorial changes or differences in fiscal years). So far, we focus on manually checking mappings which are automatically identified by our algorithms and deploying an evaluation collecting system from users. In the future, we need to obtain experts' evaluation from data sources that StatSpace covers.

In Linked Data context, ensuring the correctness of equivalent relationships is still a challenge. One principle of Linked Data is that data publishers need to create links between their URIs and existing URIs. A relationship is popularly used, i.e., *owl:sameAs*. At present, data providers often rely on link discovery tools, e.g., LIMES [80] and SILK [81] to perform this task. These tools typically use string similarity metrics to propose equivalent relationships between two data sources. However, they cannot ensure the correctness of proposed mappings. It is a task of data publishers. Supposing that somehow data publishers can validate proposed mappings and then establish *owl:sameAs* links. At present, we have a *sameAs* service [82] available at `http://sameas.org/`, which craws equivalent relationships used in a limited number of data sources. The author of this service explains his requirement for selecting input data sources , i.e., that these data sources are "sufficiently accurate" following his opinion[3], for instance, `http://dbpedia.org/`, `http://linkedgeodata.org/`. However, Brenninkmeijer et al. [83] show that this service still contains a lot of incorrect mappings. This shows that in a small number of "quality" data sources, data publishers may still include incorrect links.

[3] `http://sameas.org/about.php`, accessed December 30, 2016

CHAPTER 6

Related Work

In this chapter, we present work done in relation to our research. We organize the related work into four categories including mapping languages (Section 6.1) research on coreference resolution in Linked Data environment (Section 6.2), statistical data integration research (Section 6.3), and research on statistical data exploration (Section 6.4).

6.1 Mapping Languages

To transform raw data sets into RDF format, users can make use of mapping definition languages which define customized rules for generating RDF triples from an input data set. In this section, we first introduce mapping languages and then we explain the reasoning behind our choice of the RML language in StatSpace architecture.

CSV/Spreadsheet-to-RDF mapping languages

XLWrap[1] [84] supports the data transformation of spreadsheets and CSV files into RDF format. This language makes use of TriG[2], an extension syntax of Turtle, to describe mappings. It is able to process complex data layouts such as cross tables which may relate to multiple columns, rows, sheets, and files. *XLWrap* also supports a wide range of operators and core functions including all standard arithmetic operators, logical operators, string manipulation functions, and aggregated functions.

M^2 (Mapping Master) [85] is a mapping language for spreadsheet data transformation into Web Ontology Language (OWL). The distinction of this transformation is that it requires the generation of many classes and properties. In this context, mapping languages such as *XLWrap* need a verbose syntax for expressing mappings [85]. To overcome this issue, O'Connor et al. developed a new mapping language based on an

[1] `http://xlwrap.sourceforge.net/`, accessed December 30, 2016
[2] `https://www.w3.org/TR/trig/`, accessed December 30, 2016

extension of the Manchester OWL syntax [86]. This language not only supports arbitrary spreadsheet layouts but also provides a concise representation for mappings.

XML-to-RDF mapping languages

XSPARQL [21] is a combination of two languages, i.e., SPARQL and XQuery[3] to allow data transformation between XML and RDF formats. In addition, *XSPARQL* also supports RDF-to-RDF transformation. To build this language, the authors rely on the syntax of XQuery and define some new components that are available in SPARQL syntax.

Relational Database-to-RDF mapping languages

D2RMap [87] is a declarative language based on XML syntax for expressing mappings between relational databases and RDF. This language can process many database structures including one-to-one, one-to many, many-to-many relationships, normalized table structures, etc. Each mapping file contains two parts: (i) the first part describes the database connection, the desired output format (e.g., RDF, N3, N-Triples), and a namespace; (ii) the second part defines components in a triple (i.e., subject, predicate, object) and the patterns for coining these URIs.

R2RML [60] is the only mapping language that has become a W3C recommendation for data transformation. R2RML mappings need to be written in the Turtle RDF syntax and satisfy the R2RML mapping graph. This graph includes a set of rules about the generation of Internationalized Resource Identifiers (IRIs) and the correct use of defined classes, etc. To transform a table, view, or valid SQL query of the input database, users need to define a *triples map* which includes two components: (i) a subject for all generated RDF triples and (ii) a set of *predicate-object maps* describing pairs of predicates and objects that will be linked to the subject. In addition, users can link two *triples maps* through JOINT operator to express triples that are generated from data of both tables.

M2RML [88, 89] is a mapping language designed to transform multidimensional data in Online Analytical Processing (OLAP) cubes into RDF format. To define this language, the authors use *R2RML* as a reference language. Compared to *R2RML*, there are two salient characteristics to distinguish it, i.e., (i) it follows the QB vocabulary to represent the output RDF data set whereas *R2RML* allow users to follow an arbitrary vocabulary, and (ii) *M2RML* defines special classes and properties which are mapped to concepts of the QB vocabulary. The special classes allow users to eliminate the declaration of several *predicate-object maps* whereas the special properties remove either the predicate element or the object element in a *predicate-object map*. Although the use of these definitions reduces the number of mappings that users have to define, it may break the intelligibility and consistency of the whole mapping file.

RML [90, 33]. A large number of mapping languages have been introduced by researchers to lift raw data sets into RDF. However, these languages are typically limited to only one input format. Therefore, to transform raw data sets published in multiple heterogeneous

[3] `https://www.w3.org/TR/xquery/`, accessed December 30, 2016

formats into RDF format, users need to combine different languages. To address this issue, Dimou et al. present the *RML* language based on *R2RML* to provide a uniform mapping description for data transformation of an arbitrary input format. Database-related concepts in *R2RML* such as *logical table*, *table name*, *column*, etc. are replaced by new generic concepts i.e., *logical source*, *source name*, *reference*, respectively. In addition, *triples maps* can be linked to each other in order to create the link between resources in the same data set or in different data sets. *RML* is designed to support data transformation of every format. However, its current processor[4] is limited to three input formats, i.e., CSV, JSON, and XML.

To sum up, *RML* has advantages compared to other mapping languages. The use of this language in StatSpace architecture is therefore a rational approach. However, we still need to extend the existing RML processor to support not only new input formats, such as spreadsheets, databases, etc. but also operators and manipulation functions that are supported in *XLWrap* language.

	Supported formats					Support operators and manipulation functions
	CSV	Spreadsheet	JSON	XML	Database	
XLWrap	✓	✗	✗	✗	✗	✓
M^2	✗	✓	✗	✗	✗	✗
XSPARQL	✗	✗	✗	✓	✗	✗
D2RMap	✗	✗	✗	✗	✓	✗
R2RML	✗	✗	✗	✗	✓	✗
M2RML	✗	✗	✗	✗	✓	✗
RML	✓	✓*	✓	✓	✓*	✗

Table 6.1: Key characteristics of mapping languages
✓: Yes; ✓*: Possible; ✗: No

6.2 Coreference Resolution in Linked Data context

Coreference resolution [91, 92] is the task of identifying all expressions that denote the same entity in a text. In the Linked Data context, we classify research on coreference resolution into two groups: (i) link discovery frameworks (Section 6.2.1) that determine the equivalent relationships between URIs used in two given data sources and (ii) coreference resolution services (Section 6.2.2) that crawl and manage equivalent relationships in data sources.

6.2.1 Link Discovery Frameworks

KnoFuss [93, 94] presents an architecture for link discovery. This architecture consists of three main components, i.e., (i) *a set of tasks* needed to be finished; (ii) *a library*

[4] `https://github.com/RMLio/RML-Processor/`, accessed December 30, 2016

of methods for solving tasks such as string similarity algorithms, aggregated functions, threshold values, etc.; and (iii) *a mechanism* for selecting and invoking appropriate methods based on a specific task and input data.

Silk [81, 95] is a framework designed to support data publishers in establishing links to existing data sources. It uses a declarative language named "Silk - Link Specification Language" to support users in describing their requirements such as: (i) address of SPARQL endpoints that they want to link to; (ii) type of relationship such as *owl:sameAs*; and (iii) link conditions need to be satisfied. To describe these conditions, users can use string similarity metrics such as Jaro distance [96], Q-Grams distance [97], etc. as well as aggregated functions such as MAX, MIN, AVG, etc. The discovered links will be returned to users in an RDF file.

LIMES (Link Discovery Framework for Metric Spaces) [80] is a time-effective approach for link discovery between data sources. The authors make use of the "triangle inequality" theorem [98] to quickly eliminate a large number of candidate pairs that cannot fulfil the link conditions. To evaluate this framework, the authors perform experiments on real data in order to discover links between DBpedia and other data sources such as Drugbank[5] and LinkedCT[6]. The results show that LIMES outperforms the Silk framework in all experiments.

Our approach for equivalent relationships discovery is similar to the approach used in existing frameworks, i.e., based on the use of link conditions. The difference lies in the implementation of these conditions. For example, in our approach we rely on geographical information to match spatial URIs used in a data source to shared URIs, whereas Silk and LIMES typically rely on string distance metrics to perform the comparison. Table 6.2 provides a comparison of key characteristics between our approach and the existing frameworks.

6.2.2 Coreference Resolution Services

Glaser et al. [100, 82, 22] present an approach to manage equivalent relationships in and between data sources. The authors argue that the semantics of *owl:sameAs* are too strict whereas the meaning of URIs may depend on specific contexts [101, 102]. A URI of Austria, for instance *ex:Austria*, may refer to a political entity, a geographical area, as well as a sport team. Therefore, data publishers should manage equivalent URIs in *coreference bundles* instead of using *owl:sameAs* links. Each bundle is a set of URIs that refers to the same entity in terms of semantics. In addition, each bundle has its own URI, which allows it to be reused in different sources. The authors applied this approach to 20+ data sources generated in ReSIST project[7] [103], which aims to publish data from

[5] `http://wifo5-03.informatik.uni-mannheim.de/drugbank/sparql`
[6] `http://data.linkedct.org/sparql/`
[7] `http://www.rkbexplorer.com/`, accessed December 30, 2016

	KnoFuss	Silk	LIMES	Our approach
Domains used	all	all	all	statistical data
Input	RDF, SPARQL	RDF, SPARQL	RDF, SPARQL	RDF, SPARQL
Supported link	owl:sameAs	owl:sameAs and other link types	owl:sameAs and other link types	owl:sameAs
Configuration	manual, unsupervised learning	manual, supervised learning	manual, supervised learning, unsupervised learning	automatic
String similarity measures	✓	✓	✓	✗
Other similarity measures	–	date similarity, numeric similarity	–	similarity of eleven dimensions and unit attribute e.g., date, age, area, etc.

Table 6.2: A comparison between approaches of link discovery (based on the survey of link discovery frameworks [99]). ✓: Yes; ✗: No; –: Undefined

	Glaser et al.	Schlegel et al.	Our approach
Input sources	SPARQL	SPARQL	SPARQL, RDF
Size of input sources	50+	112	8
Method used to discovery link	special predicates	special predicates	keywords, patterns, and geographical information
Use a set of shared URIs for linkage	✗	✗	✓

Table 6.3: A comparison between coreference resolution services ✓: Yes; ✗: No

ACM[8], DBLP[9], IEEE[10], etc. into RDF format.

In addition, the authors also present a procedure [82] for identifying a complete set of equivalent relationships for a given URI. The mechanism of this procedure relies on

[8] `https://www.acm.org/`, accessed December 30, 2016
[9] `http://dblp.uni-trier.de/`, accessed December 30, 2016
[10] `https://www.ieee.org/index.html`, accessed December 30, 2016

the traversal of synonymous URIs in the same bundle and the "following your nose" principle. Although this procedure assumes that data sources use *coreference bundles*, it can be applied to other data sources that use "owl:sameAs" links. As a result, the authors introduce a coreference resolution service at `http://sameas.org/` based on the crawling of equivalent relationships from the data sources in RESIST project and more than 30 other data sources which the authors think are "sufficiently accurate"[11] such as `http://dbpedia.org/`, `http://linkedgeodata.org/`. Equivalent relationships are identified through various predicates including:

```
http://www.w3.org/2002/07/owl#sameAs
http://www.rkbexplorer.com/ontologies/coref#coreferenceData
http://umbel.org/umbel/sc/isLike
http://www.w3.org/2004/02/skos/core#exactMatch
http://www.w3.org/2004/02/skos/core#closeMatch
http://open.vocab.org/terms/similarTo
http://www.geneontology.org/formats/oboInOwl#hasExactSynonym
```

However, Brenninkmeijer et al. [83] show the inaccuracy of this service in some cases. For example, when querying equivalent URIs for a URI referring to a protein entity, i.e., `http://www.uniprot.org/uniprot/P06213`, the service returns more than 33,000 results. However, the first 1,292 URIs from DBpedia refer to gene concepts, hence, they are not correct results.

Schlegel et al. [23] use special predicates such as *owl:sameAs*, *skos:exactMatch*, etc. to identify equivalent relationships in the Linked Open Data cloud [62]. In addition, provenance information of those data sources that define the relationship is also gathered. Therefore, users can restrict equivalent URIs from some special endpoints that they want to focus on. At present, users can explore the coreference resolution service which is a part of query rewriting service at `http://zaire.dimis.fim.uni-passau.de:8080/balloon/demo`. Compared to the `http://sameas.org` service, the distinction of this service is that the authors crawl data from a large number of SPARQL endpoints registered at DataHub platform[12]. As a result, 17.6M coreference statements are collected and clustered from 112 endpoints (237 in total) at DataHub. However, no SPARQL endpoint of statistical data that we collected (cf. Table 4.9) is analyzed because these endpoints are either not registered at DataHub platform, such as `http://data.europa.eu/euodp/en/linked-data` or ignored by the crawler, such as `http://semantic.eea.europa.eu/sparql`.

To sum up, the two existing coreference resolution services introduced by Glaser et al. [100, 82, 22] and Schlegel et al. [23] do not focus on discovering equivalent relationships used in statistical data sources. To deal with this, we match URIs used in disparate data sources to a set of shared URIs. Compared to existing approaches, instead of using special predicates like *owl:sameAs, skos:exactMatch, etc.*, we rely on patterns, keywords, and

[11] `http://sameas.org/about.php`, accessed December 30, 2016
[12] `http://datahub.io/`, accessed December 30, 2016

Google's geocoding service. This approach allows us to identify equivalent relationships between URIs even when they may not be linked to each other. Our coreference resolution service is available at `http://statspace.linkedwidgets.org/sameas/`.

6.3 Data Integration Research

In this section, we contrast our work to related statistical data integration approaches that have been introduced in recent years.

Capadisli et al. [104, 31] transform statistical data of major organizations such as the World Bank, the International Monetary Fund, the European Central Bank, etc. into RDF format following the QB and then store them in endpoints (each endpoint stores data from a data source). Next, based on the use of the LIMES link discovery framework [80] (cf. Section 6.2.1), equivalent relationships between data sources as well as with external data sources including DBpedia[13] and GeoNames[14] are established. To illustrate the benefits of the data transformation and link relationships, the authors provide a web-based interface to allow users to compare two statistical indicators. In the first step, users select two data sets in existing sources, such as *mortality rate* from the World Bank and *corruption perceptions index* from Transparency International. The application then generates SPARQL queries to gather data from related endpoints. Next, the results are analyzed by Shiny[15] – a framework that supports R language and is visualized to users via a scatter plot in order to express the relationship between two select indicators.

Sabou et al. [105, 28, 106] transform tourist indicators stored in the TourMIS database [107] into RDF format. Next, the authors use the Silk platform [81] (cf. Section 6.2.1) to establish links to DBpedia[16] and GeoNames[17] for URIs referring to cities and countries. The authors also establish *owl:sameAs* links to URIs used in two existing SPARQL endpoints, i.e., `http://worldbank.270a.info/sparql` (stores data of the World Bank) and `http://ecb.270a.info/sparql` (stores data of the European Central Bank). To illustrate the capability of the converted data, the authors provide a visual dashboard at `http://etihq.weblyzard.com/`, which allows users to compare tourist indicators and economic indicators. For instance, users can evaluate the influence of Germany GDP on the number of tourists coming from Germany to Prague.

Kämpgen et al. [24, 108] establish mappings between concepts in the QB vocabulary and elements in a *multidimensional model* used in OLAP. These mappings are then used by an ETL pipeline to translate data from Linked Data sources into tables in a data warehouse. Next, the authors store data in an OLAP server and use Multidimensional Expression Language (MDX) to analyze and integrate statistical data from multiple sources. A drawback of this approach is that the data transformation process needs to be repeated

[13] `http://wiki.dbpedia.org/`, accessed December 30, 2016
[14] `http://sws.geonames.org/`, accessed December 30, 2016
[15] `http://shiny.rstudio.com/`, accessed December 30, 2016
[16] `http://wiki.dbpedia.org/`, accessed December 30, 2016
[17] `http://sws.geonames.org/`, accessed December 30, 2016

when the original data sources are updated. Therefore, in [109, 108] the authors propose another approach. They first define operations of basic OLAP operators [110] including slice, dice, projection, and roll-up on individual RDF data cubes and drill-across on multiple data cubes. Next, the authors introduce a mechanism for transforming an OLAP query based on the MDX language into a SPARQL query. However, this transformation needs additional information because in an OLAP query, the role of each element (e.g., dimension, measure, etc.) is not clearly expressed.

Identifying the relationship between different statistical data sets has also attracted a great interest in recent years. Capadisli et al. [111] show that the similarity of titles is not closely related to the correlations between data sets. Kämpgen et al. [25] rely on an extension of drill-across operator and conversion functions to identify the capability of data integration for two arbitrary data sets. The authors show two essential requirements for data integration, i.e., (i) the equivalent relationships between URIs in two data sets need to be predefined and (ii) two datasets must have same number of dimensions. Bayerl et al. [112] introduce a similarity measure to rank compatible data sets for a given data set to prepare for data integration. This measure is calculated based on the data structures in two data sets, such as the number of dimensions in each data set, the number of common dimensions, etc. Based on calculated values, a ranked list of data sets regarding an input data set is provided to users. Meimaris et al. [113, 114] introduce two properties named *containment* and *complementarity* to describe the relationships between two observations. In particular, *containment* reflects the case in which an observation includes aggregated information of the remaining observation. For example, the first observation contains the population figure of Austria in 2016 and the second one stores population data of Vienna city in the same year. *Complementarity* refers to the capability of combining two observations in order to extend the information and allow users to identify the correlation between observed measures. For instance, two observations storing different indicators of Austria in 2016 will have a *complementarity* relationship.

To sum up, salient characteristics that distinguish our work from these related efforts are as follows:

- Whereas existing research makes use of the *warehousing approach* to integrate statistical data from multiple heterogeneous data sources, we introduce an architecture following the *virtual integration approach*. As a result, we can provide up-to-date data from original sources to users. In addition, we can avoid challenges of data storing space when we integrate new data sources in the future.
- We identify and introduce to users a large number of equivalent relationships, which can be used to create interconnection between heterogeneous data sources. Existing work integrates data from sources which typically have simple data structures. For instance, SPARQL endpoints created by Capadisli et al. [111] have only two dimensions including one spatial dimension and one temporal dimension. Therefore, to integrate these data sources researchers typically need to identify equivalent relationships for a limited number of URIs. In our work, due to the diversity

	Capadisli et al.	Sabou et al.	Kämpgen et al.	Our approach
Original format	SDMX-XML	database	SPARQL	raw formats, SPARQL
Approach	warehousing	warehousing	warehousing	virtual integration
Tools for coreference discovery	LIMES	Silk	manual	matching algorithms
Equivalent relationship types	values of spatial and temporal dimensions	values of spatial and temporal dimensions	values of spatial and temporal dimensions	values of eleven dimensions and unit attribute
Requirements of data structure	same structure	–	same structure	same structure or different structures
Support unit conversion	–	–	✓	✓
Support uniform access	✗	✗	✗	✓

Table 6.4: A comparison between approaches of statistical data integration
✓: Yes; ✗: No; –: Undefined

of data sources and their consistent data structures, we need to identify a large amount of mappings between URIs used in data sources and shared URIs. This work leads to a coreference resolution service available at `http://statspace.linkedwidgets.org/sameas/`.

- Based on identified mappings, hierarchical code lists of components, and subjects of data sets we can identify relatable data sets to a given data set. In addition, we present flexible requirements for data integration. For example, we show that two data sets do not need to have the same structure if we can assign values at top concepts levels to dimensions that do not appear in both data sets.

- Based on metadata descriptions and shared URIs, we can provide a uniform and integrated access to separate statistical data sets. As a result, users can integrate data of different data sources without understanding data structures, access mechanisms, formats, and encoding of individual data sources. Table 6.4 introduces a comparison between our approach and related efforts of statistical data integration.

6.4 Data Exploration Research

Research on data exploration has received significant interest from researchers. A large number of applications have been developed to support users in exploring statistical data in various ways.

CubeViz[18] [26, 115, 116] is an application developed based on the OntoWiki Framework [117] to provide visualization for statistical data. This application receives a SPARQL endpoint or an RDF file adhering to the QB vocabulary as its input and provides visual presentations via five different types of charts including pie, bar, column, line, and polar charts. The visualization process is performed in three main steps. First, *CubeViz* utilizes SPARQL ASK queries to evaluate wherever the integrity constraints[19] of the QB are obeyed. Therefore, only well-formed data cubes are passed to the next step and visualized by this tool. In the second step, users select an arbitrary data set, and then they choose values for its dimensions to construct filter conditions. Finally, *CubeViz* generates a corresponding SPARQL query to obtain relevant data. The result will be analyzed and visualized to users through the use of the D3js[20] and the HighCharts[21] libraries.

Linked Data Query Wizard (LDQW)[22] [118, 29, 119] is a web-based application allowing users to access, filter, and analyze data in SPARQL endpoints. It offers two input options: (i) users provide keywords for searching over a specific SPARQL endpoint; (ii) users select an arbitrary RDF data cube stored in a specific SPARQL endpoint. In both options, the application will generate corresponding SPARQL queries and represent the resulting data via a 2-dimensional tabular interface. Therefore, even non-expert users who are familiar with search engines and spreadsheet tools can easily utilize this application. The table interface is organized as follows: (i) each row represents a subject; (ii) each column describes a predicate; and (iii) each cell stores the object that is identified based on its subject (in row) and predicate (in column). In addition, *Linked Data Query Wizard* provides a large number of functionalities for users to facilitate data exploration such as: add/remove column, set/remove filter value, aggregated values, etc.

Linked Data Vis Wizard (LDVW)[23] [120, 121, 119] is a web-based application developed to provide suitable visualizations for RDF data cubes stored in SPARQL endpoints. To obtain this goal, the authors first develop a visualization vocabulary named *VA* to represent information of visualization semantically. In particular, this vocabulary will describe axes of a chart, suitable data types that a specific axis supports, the number of allowed instances for an axis, e.g., one for x-axis in bar chart but many for this axis in parallel coordinates, etc. Next, the authors build an algorithm relying on the compatibility of the structure and data types to map a specific data set to the

[18] `http://cubeviz.aksw.org/`, accessed December 30, 2016
[19] `https://www.w3.org/TR/vocab-data-cube/#wf-rules`, accessed December 30, 2016
[20] `https://d3js.org/`, accessed December 30, 2016
[21] `http://www.highcharts.com/`, accessed December 30, 2016
[22] `http://code.know-center.tugraz.at/search`, accessed December 30, 2016
[23] `http://code.know-center.tugraz.at/vis`, accessed December 30, 2016

corresponding visualizations. The result of the algorithm is a set of suitable charts for a given data set. At present, the authors make use of Google chart[24] and the D3js[25] libraries to provide a list containing nine different charts. In addition, simultaneous visualizations of many data sets are possible if these data sets make use of the same URIs to represent their dimensions.

OpenCube Toolkit [6, 122, 8, 10] contains a set of three applications developed to support users in exploring statistical data: (i) *OpenCube Browser* is a web application that represents a statistical data set by a two-dimensional table. Users can change the provided configuration (i.e., change two select dimensions, change fixed values for other dimensions) to explore the data set according to new slices. In addition, the browser also supports two OLAP operators including drill-down and roll-up. (ii) *The OpenCube Map View* is a visualization application that is specifically tailored to represent the spatial dimension. It assumes that this dimension is represented by the property, i.e., *sdmx:refArea* or a sub-property of this property. This application assigns fixed values for all other dimensions whereas geographical areas (i.e., values of spatial dimension) are visualized on a map based on OpenStreetMap[26], Mapbox[27] and Leaflet[28] library. (iii) *The R statistical analysis* is a web service relying on Rserve package[29] to provide analysis for data cubes. For example, it can give forecasts based on the existing data. The results of the analysis will be visualized to users through charts.

Payola [123, 124, 30] is a framework supporting users in analysis and visualization of Linked Data sources. The key component of this framework is a set of plugins which is classified into two categories, i.e., *analyzers* and *visualizers*. *Analyzers* are responsible for selecting the input data sources according to a specific requirements, choosing properties, performing data transformation, etc. The results of *Analyzers* will be passed to *Visualizers* in order to generate visual charts for users. Currently, existing *Visualizers* make use of the HighCharts[30] library and provide five different types of charts including line, pie, column, bar, and area charts. Regarding statistical data, *Payola* can receive an arbitrary RDF source and transform it to RDF format conforming to the QB Vocabulary. This process is conducted in a special *analyzer* plugin that requires users to select a target data structure from a list of structures and map properties in the input source to dimensions, measures, and attributes of the select structure.

Compared to existing efforts, two contributions from our work are as follows: (i) whereas most existing applications provide only data visualization for individual data sets, we can identify relatable data sets to a given data set based on semantic descriptions of these data sets. As such, we provide users with not only data visualization but also

[24] `https://developers.google.com/chart/`, accessed December 30, 2016
[25] `https://d3js.org/`, accessed December 30, 2016
[26] `http://wiki.openstreetmap.org/wiki/Main_Page`, accessed December 30, 2016
[27] `https://www.mapbox.com/`, accessed December 30, 2016
[28] `http://leafletjs.com/`, accessed December 30, 2016
[29] `https://cran.r-project.org/web/packages/Rserve/index.html`, accessed December 30, 2016
[30] `http://www.highcharts.com/`, accessed December 30, 2016

correlation mining and data quality assessment between multiple data sets; (ii) we can provide visualization for raw data sources while existing applications typically focus on data visualization for sources published in RDF format. Table 6.5 highlights differences between our approach and the related work.

	CubeViz	LDQW	LDVW	OpenCube	Payola	Our approach
Provide data visualization	✓	✓	✓	✓	✓	✓
Support search	✗	✓	✗	✗	✗	✓
Discover relatable datasets	✗	✗	✗	✗	✗	✓
Support data quality assessment	✗	✗	✗	✗	✗	✓
Support correlation mining	✗	✗	✗	✗	✗	✓
Provide statistical analysis	✗	✗	✗	✓	✗	✗
Input data	SPARQL, RDF	SPARQL	SPARQL, PDF	SPARQL	SPARQL	SPARQL, Raw data

Table 6.5: Key characteristics of applications of statistical data exploration
✓: Yes; ✗: No

CHAPTER 7

Conclusions and Future Work

7.1 Summary

Recent research [6, 7, 8, 9, 10] reports a significant number of statistical data sets published on the web by various governments and organizations. In this context, the use of statistical data creates new opportunities for interesting applications and facilitates more informed decision-making. However, viable means to integrate and explore statistical data sets available on the web are still scarce.

In this thesis, we address the challenges on how to integrate and explore statistical data published in disparate data sources. To the best of our knowledge, we present the first statistical data integration architecture that makes use of the virtual integration approach to integrate data available on the web. To show the effectiveness of this approach, we implement StatSpace, a linked statistical data space that covers and integrates more than 1,800 heterogeneous data sets from eight different publishers. In addition, we evaluate our approach in terms of coverage, validity, and performance. In the following, we describe the contributions of the thesis through giving answers to the research questions posed in Section 1.3.

R1. *How can we address data heterogeneity in terms of formats?*

First, we separate existing data sources into two categories: (i) RDF data sources that are typically stored in SPARQL endpoints; and (ii) Raw data sources that can be published in various formats such as CSV, JSON, Spreadsheet, etc. Next, we select RDF format as the target format for data representation. To this end, we rely on the use of RDF mapping language (RML) to lift raw data sets into RDF following the Data Cube vocabulary. The transformation is performed on the fly when required by users or applications. We also extend this language to support a new format (i.e., Spreadsheet) and enhance the reusability of mappings through variable declarations.

R2. *How can we establish an interconnection between statistical data sets?*

We follow an approach that contains two steps. In the first step, we review URIs and code lists used in existing data sources to find a set of URIs and code lists that have a large coverage. Based on these identifiers, we define *URIs design patterns* to coin shared URIs for reference. In the second step, we build algorithms to map URIs used in data sources to shared URIs. Because raw data sources are also transformed into RDF format through the RML mapping service, this step is suitable for all data sources.

R3. *How can we provide uniform and integrated access to individual data sets?*

We use *metadata* to describe information needed to query and integrate individual data sets. Each metadata description contains two main parts: (i) information of data structure and access mechanism; and (ii) equivalent relationships between URIs used in each data set to a set of shared URIs. In addition, we create a mediator to answer queries posed by users. The mediator can translate a generic query relying on shared URIs into suitable queries for related data sets. It can also rewrite individual results and integrate them to a final result. As a result, uniform and integrated access across multiple data sets are supported.

To sum up, the central research question, i.e., *How can users be enabled to integrate and explore multiple heterogeneous data sources?* is answered through the contributions in this thesis.

7.2 Future Work

In the future, we plan to extend the present components to enhance the efficiency of StatSpace and create more convenient mechanism for crowd participation

RML Mapping Service

At present, our mapping service consists of four separate sub-processors to transform raw data into RDF format from four input formats, i.e., CSV, JSON, XML, and Spreadsheet. In future work, this can be necessary to build new sub-processors for other formats such as relational database, HTML, Document files (.DOC), DOCX, etc. In addition, we also plan to provide convenient interfaces to support users in defining mappings.

URI Design Patterns

We will consider reusing available code lists [125] defined by data publishers if these code lists offer wide coverage. This extension is necessary to allow StatSpace to provide data integration for new data sources in the future.

Metadata Generator

The *metadata generator* service needs to be improved to allow developers and data providers to contribute new metadata descriptions to the *metadata repository* in

an easier way. In addition, we need to conduct user studies in a broader scope to gather their evaluations of the correctness of provided mappings.

Metadata Repository

We plan to generate metadata for statistical data sets published by major organizations such as the IMF, WHO, OECD, etc. The resulting repository, hence, will enhance its value to users.

Mediator and Explorer

We plan to extend the mediator to support more complex analyses that involve additional query conditions and require the integration of data on multiple subjects. For the explorer, we plan to propose a similarity measure based on features of data sets such as subjects, dimensions, code lists, etc. in order to compute the ranking of each relatable data set.

Appendix

A URI Design Patterns for Code Lists

A.1 Reference Area Dimension (cl_area)

- URI: `http://statspace.linkedwidgets.org/codelist/cl_area`
- Semantic description: `http://statspace.linkedwidgets.org/code/cl_area.ttl`
- Pattern: `http://statspace.linkedwidgets.org/codelist/cl_area/{Country}/{Area in level 1}/{Area in level 2}/...{Area in level n}`
- Example: `http://statspace.linkedwidgets.org/codelist/cl_area/Austria/Vienna` represents Vienna, Austria

A.2 Reference Period Dimension (cl_period)

- URI: `http://statspace.linkedwidgets.org/codelist/cl_period`
- Semantic description: `http://statspace.linkedwidgets.org/code/cl_period.ttl`
- Base URI: `http://reference.data.gov.uk`
- Example: `http://reference.data.gov.uk/id/gregorian-year/2016` represents year 2016 and `http://reference.data.gov.uk/id/gregorian-month/2016-01` represents January, 2016
- Patterns:

No.	Pattern	Description
1	/id/gregorian-year/{year}	Year
2	/id/gregorian-half/{year}-{half}	One-half year
3	/id/gregorian-quarter/{year}-{quarter}	Quarter
4	/id/gregorian-month/{year}-{month}	Month
5	/id/gregorian-day/{year}-{month}-{day}	Day
6	/id/gregorian-week/{year}-{week}	Week
7	/id/gregorian-interval/{dateTime}/{duration}	Duration

Table A.1: Patterns designed for code list of temporal dimension

A.3 Age Dimension (cl_age)

- URI: http://statspace.linkedwidgets.org/codelist/cl_age
- Semantic description: http://statspace.linkedwidgets.org/code/cl_age.ttl
- Base URI: http://statspace.linkedwidgets.org/codelist
- Examples: http://statspace.linkedwidgets.org/codelist/cl_age/Y80 represents age 80 and http://statspace.linkedwidgets.org/codelist/cl_age/Y80T84 represents age group from 80 to 84
- Patterns:

No.	Pattern	Description
1	/cl_age/Y{n}, n=0, 1, 2,...,105	Individual age
2	/cl_age/Y{n}T{n+4}, n=0, 5, 9,...,105	Age group (5 years)
3	/cl_age/Y{n}T{n+9}, n=25, 35,...,95	Age group (10 years)
4	/cl_age/Y_GE_{n}, n=65, 70,...,90	Age group (equal or above a specific age)
5	/cl_age/Y_LE_{n}, n=15, 20	Age group (under a specific age)
6	/cl_age/TOTAL	Top concept
7	/cl_age/UNK	Unknown age

Table A.2: Patterns for the code list of age dimension

A.4 Education Level Dimension (cl_educationLev)

- URI: `http://statspace.linkedwidgets.org/codelist/cl_educationLev`
- Semantic description: `http://statspace.linkedwidgets.org/code/cl_educationLev.ttl`
- Pattern: `http://statspace.linkedwidgets.org/codelist/cl_educationLev/L{code}`
- URIs:

No.	URI	Description
1	`http://statspace.linkedwidgets.org/codelist/cl_educationLev/L0`	Pre-primary education
2	`http://statspace.linkedwidgets.org/codelist/cl_educationLev/L1`	Primary education
3	`http://statspace.linkedwidgets.org/codelist/cl_educationLev/L2`	Lower secondary
4	`http://statspace.linkedwidgets.org/codelist/cl_educationLev/L3`	Upper secondary
5	`http://statspace.linkedwidgets.org/codelist/cl_educationLev/L4`	Post-secondary non-tertiary education
6	`http://statspace.linkedwidgets.org/codelist/cl_educationLev/L5`	Short-cycle tertiary education
7	`http://statspace.linkedwidgets.org/codelist/cl_educationLev/L6`	Bachelor or equivalent
8	`http://statspace.linkedwidgets.org/codelist/cl_educationLev/L7`	Master or equivalent
9	`http://statspace.linkedwidgets.org/codelist/cl_educationLev/L8`	Doctoral or equivalent
10	`http://statspace.linkedwidgets.org/codelist/cl_educationLev/L9`	Not elsewhere classified

Table A.3: URIs in the code list of education level dimension

A.5 Occupation Dimension (cl_occupation)

- URI: `http://statspace.linkedwidgets.org/codelist/cl_occupation`
- Semantic description: `http://statspace.linkedwidgets.org/code/cl_occupation.ttl`

- Pattern: `http://statspace.linkedwidgets.org/codelist/cl_occupation/{code}`

- Examples: `http://linkedwidgets.org/resource/codelist/cl_occupation/OC1` represents Managers and `http://statspace.linkedwidgets.org/codelist/cl_occupation/OC11` represents Chief executives, senior officials and legislators.

A.6 Currency Dimension (cl_currency)

- URI: `http://statspace.linkedwidgets.org/codelist/cl_currency`

- Semantic description: `http://statspace.linkedwidgets.org/code/cl_currency.ttl`

- Pattern: `http://statspace.linkedwidgets.org/codelist/cl_currency/{ISO 4217 code}`

- Examples: `http://statspace.linkedwidgets.org/codelist/cl_currency/AED` represents the United Arab Emirates dirham and `http://statspace.linkedwidgets.org/codelist/cl_currency/EUR` represents Euro.

A.7 Civil Status Dimension (cl_civilStatus)

- URI: `http://statspace.linkedwidgets.org/codelist/cl_civilStatus`

- Semantic description: `http://statspace.linkedwidgets.org/code/cl_civilStatus.ttl`

- Pattern: `http://statspace.linkedwidgets.org/code/cl_civilStatus/{code}`

- URIs:

No.	URI	Description
1	http://statspace.linkedwidgets.org/codelist/cl_civilStatus/D	Divorced person
2	http://statspace.linkedwidgets.org/codelist/cl_civilStatus/E	Person whose registered partnership was legally dissolved
3	http://statspace.linkedwidgets.org/codelist/cl_civilStatus/L	Legally separate person
4	http://statspace.linkedwidgets.org/codelist/cl_civilStatus/M	Married person
5	http://statspace.linkedwidgets.org/codelist/cl_civilStatus/P	Person in registered partnership
6	http://statspace.linkedwidgets.org/codelist/cl_civilStatus/Q	Person whose registered partnership ended with the death of the partner
7	http://statspace.linkedwidgets.org/codelist/cl_civilStatus/S	Single person
8	http://statspace.linkedwidgets.org/codelist/cl_civilStatus/W	Widowed person

Table A.4: URIs in the code list of civil status dimension

A.8 Frequency Dimension (cl_frequency)

- URI: http://statspace.linkedwidgets.org/codelist/cl_frequency
- Semantic description: http://statspace.linkedwidgets.org/code/cl_freq.ttl
- Pattern: http://purl.org/linked-data/sdmx/2009/code#freq-{code}
- URIs:

No.	URI	Description
1	http://purl.org/linked-data/sdmx/2009/code#freq-H	Hourly
2	http://purl.org/linked-data/sdmx/2009/code#freq-D	Daily
3	http://purl.org/linked-data/sdmx/2009/code#freq-N	Minutely
4	http://purl.org/linked-data/sdmx/2009/code#freq-S	Half yearly, semester

5	http://purl.org/linked-data/sdmx/2009/code#freq-A	Annual
6	http://purl.org/linked-data/sdmx/2009/code#freq-Q	Quarterly
7	http://purl.org/linked-data/sdmx/2009/code#freq-M	Monthly
8	http://purl.org/linked-data/sdmx/2009/code#freq-B	Daily-business week
9	http://purl.org/linked-data/sdmx/2009/code#freq-W	Weekly

Table A.5: URIs in the code list of frequency dimension

A.9 Sex Dimension (cl_sex)

- URI: http://statspace.linkedwidgets.org/codelist/cl_sex
- Semantic description: http://statspace.linkedwidgets.org/code/cl_sex.ttl
- Pattern: http://purl.org/linked-data/sdmx/2009/code#sex-{code}
- URIs:

No.	URI	Description
1	http://purl.org/linked-data/sdmx/2009/code#sex-M	Male
2	http://purl.org/linked-data/sdmx/2009/code#sex-F	Female
3	http://purl.org/linked-data/sdmx/2009/code#sex-T	Total
4	http://purl.org/linked-data/sdmx/2009/code#sex-U	Unknown gender
5	http://purl.org/linked-data/sdmx/2009/code#sex-N	Not applicable gender

Table A.6: URIs in the code list of sex dimension

A.10 Economic Activity Dimension (cl_economicActivity)

- URI: http://statspace.linkedwidgets.org/codelist/cl_economicActivity

- Semantic description: `http://statspace.linkedwidgets.org/code/cl_economicActivity.ttl`
- Pattern: `http://statspace.linkedwidgets.org/codelist/cl_economicActivity/{code}`
- Examples: `http://statspace.linkedwidgets.org/codelist/cl_economicActivity/A` represents activities of agriculture, forestry, and fishing and `http://statspace.linkedwidgets.org/codelist//cl_economicActivity /A01` represents activities of crop and animal production, hunting, and related service .

A.11 Expenditure Dimension

This code list contains four smaller code lists including:

- Classification of individual consumption by purpose (COICOP)
- Classification of the functions of governments (COFOG)
- Classification of the purposes of non-profit institutions serving households (COPNI)
- Classification of outlays of producers by purpose (COPP)

COICOP (cl_coicop)

- URI: `http://statspace.linkedwidgets.org/codelist/cl_coicop`
- Semantic description: `http://statspace.linkedwidgets.org/code/cl_coicop.ttl`
- Pattern: `http://statspace.linkedwidgets.org/codelist/cl_coicop/{code}`
- Example: `http://statspace.linkedwidgets.org/codelist/cl_coicop/CP01` represents the expenditure of food and non-alcoholic beverages.

COFOG (cl_cofog)

- URI: `http://statspace.linkedwidgets.org/codelist/cl_cofog`
- Semantic description: `http://statspace.linkedwidgets.org/code/cl_cofog.ttl`
- Pattern: `http://statspace.linkedwidgets.org/codelist/cl_cofog/{code}`
- Example: `http://statspace.linkedwidgets.org/codelist/cl_cofog/GF01` represents the expenditure of government for general public services

COPNI (cl_copni)

- URI: http://statspace.linkedwidgets.org/codelist/cl_copni
- Semantic description: http://statspace.linkedwidgets.org/code/cl_copni.ttl
- Pattern: http://statspace.linkedwidgets.org/codelist/cl_copni/{code}
- Example: http://statspace.linkedwidgets.org/codelist/cl_copni/PN1 represents the expenditure of non-profit organization for housing.

COPP (cl_copp)

- URI: http://statspace.linkedwidgets.org/codelist/cl_copp
- Semantic description: http://statspace.linkedwidgets.org/code/cl_copp.ttl
- Pattern: http://statspace.linkedwidgets.org/codelist/cl_copp/{code}
- Example: http://statspace.linkedwidgets.org/codelist/cl_copp/PP1 represents the expenditure of producer on infrastructure.

A.12 Unit of Measure (cl_unitMeasure)

- URI: http://statspace.linkedwidgets.org/codelist/cl_unitMeasure
- Semantic description: http://statspace.linkedwidgets.org/code/cl_unitMeasure.ttl
- Pattern: http://statspace.linkedwidgets.org/codelist/cl_unitMeasure/{unit.scale}
- Example: http://statspace.linkedwidgets.org/codelist/cl_unitMeasure/P1 represents unit of "People" and http://statspace.linkedwidgets.org/codelist/cl_unitMeasure/P1.6 represents unit of "Millions of People".

A.13 Subject (cl_subject)

- URI: http://statspace.linkedwidgets.org/codelist/cl_subject
- Semantic description: http://statspace.linkedwidgets.org/code/cl_subject.ttl

- Pattern: `http://statspace.linkedwidgets.org/codelist/cl_subject/{Topic.General Subject.Specific Subject.Extension}`
- Example: `http://statspace.linkedwidgets.org/codelist/subject/AG.SRF.TOTL.K2` represents subject of "Surface area (sq. km)" and `http://statspace.linkedwidgets.org/codelist/subject//SP.POP.TOTL` represents subject of "Population in total".

B RDF mapping for World Bank Data

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix rml: <http://semweb.mmlab.be/ns/rml#>.
@prefix ql: <http://semweb.mmlab.be/ns/ql#>.
@prefix rr: <http://www.w3.org/ns/r2rml#>.
@prefix sdmxd: <http://purl.org/linked-data/sdmx/2009/dimension#>.
@prefix sdmxm: <http://purl.org/linked-data/sdmx/2009/measure#>.
@prefix sdmxa: <http://purl.org/linked-data/sdmx/2009/attribute#>.
@prefix sdmxcode: <http://purl.org/linked-data/sdmx/2009/code#>.
@prefix qb: <http://purl.org/linked-data/cube#>.
@prefix dcterms: <http://purl.org/dc/terms/>.
@prefix rmlx: <http://pebbie.org/ns/rmlx#>.
<#Parameters>
  rmlx:defaultValue
    [rmlx:varName "indicator"; rr:constant "SP.POP.TOTL"],
    [rmlx:varName "refArea"; rr:constant "all"];
.
<#Observation>
  rml:logicalSource [
    rml:source "http://api.worldbank.org/countries/{refArea}/indicators/{indicator}?
        format=json&page=1&per_page=15000";
    rml:referenceFormulation ql:JSONPath;
    rml:iterator "$[1].*"
  ];
  rr:subjectMap[
    rr:class qb:Observation;
    rr:template "http://statspace.linkedwidgets.org/dataset/WorldBank-{indicator.id}/
        Obs-{country.id}-{date}"; rr:termType rr:IRI
  ];
  rr:predicateObjectMap [
    rr:predicate sdmxd:refArea;
    rr:objectMap [
      rr:template "http://statspace.linkedwidgets.org/codelist/cl_area/{country.value}"
          ; rr:termType rr:IRI
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate sdmxd:refPeriod;
    rr:objectMap [
      rr:template "http://reference.data.gov.uk/id/gregorian-year/{date}"; rr:termType
          rr:IRI
    ]
  ];

  rr:predicateObjectMap [
    rr:predicate sdmxm:obsValue;
    rr:objectMap [
      rml:reference "value"
```

```
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate sdmxa:unitMeasure;
    rr:objectMap [
      rr:constant <http://statspace.linkedwidgets.org/codelist/cl_unitMeasure/NO>
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate qb:dataSet;
    rr:objectMap [
      rr:template "http://statspace.linkedwidgets.org/dataset/WorldBank-{indicator.id}"
          ; rr:termType rr:IRI
    ]
  ];
.
<#Dataset>
  rml:logicalSource [
    rml:source "http://api.worldbank.org/countries/{refArea}/indicators/{indicator}?
        format=json&page=1&per_page=50";
    rml:referenceFormulation ql:JSONPath;
    rml:iterator "$[1][0]"
  ];
  rr:subjectMap[
    rr:class qb:DataSet;
    rr:template "http://statspace.linkedwidgets.org/dataset/WorldBank-{indicator.id}";
        rr:termType rr:IRI
  ];
  rr:predicateObjectMap [
    rr:predicate rdfs:label;
    rr:objectMap [
      rml:reference "indicator.value"
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate dcterms:subject;
    rr:objectMap [
      rr:template "http://statspace.linkedwidgets.org/codelist/cl_subject/{indicator.id
          }"; rr:termType rr:IRI
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate qb:structure;
    rr:objectMap [
      rr:constant <http://statspace.linkedwidgets.org/dataset/WorldBank/dsd>
    ];
  ];
.
<#DataStructure>
  rr:subjectMap[
    rr:class qb:DataStructureDefinition;
    rr:constant <http://statspace.linkedwidgets.org/dataset/WorldBank/dsd>
  ];
  rr:predicateObjectMap [
    rr:predicate qb:component;
    rr:objectMap [
      rr:constant <http://statspace.linkedwidgets.org/dataset/WorldBank/dsd-refArea>
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate qb:component;
    rr:objectMap [
```

```
      rr:constant <http://statspace.linkedwidgets.org/dataset/WorldBank/dsd-refPeriod>
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate qb:component;
    rr:objectMap [
      rr:constant <http://statspace.linkedwidgets.org/dataset/WorldBank/dsd-unitMeasure
          >
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate qb:component;
    rr:objectMap [
      rr:constant <http://statspace.linkedwidgets.org/dataset/WorldBank/dsd-obsValue>
    ]
  ];
.
<#Component - refArea>
  rr:subjectMap[
    rr:class qb:ComponentProperty;
    rr:constant <http://statspace.linkedwidgets.org/dataset/WorldBank/dsd-refArea>
  ];
  rr:predicateObjectMap [
    rr:predicate qb:dimension;
    rr:objectMap [
      rr:constant sdmxd:refArea
    ]
  ];
.
<#Component - refPeriod>
  rr:subjectMap[
    rr:class qb:ComponentProperty;
    rr:constant <http://statspace.linkedwidgets.org/dataset/WorldBank/dsd-refPeriod>
  ];

  rr:predicateObjectMap [
    rr:predicate qb:dimension;
    rr:objectMap [ rr:constant sdmxd:refPeriod]
  ];
.
<#Component - unitMeasure>
  rr:subjectMap[
    rr:class qb:ComponentProperty;
    rr:constant <http://statspace.linkedwidgets.org/dataset/WorldBank/dsd-unitMeasure>
  ];
  rr:predicateObjectMap [
    rr:predicate qb:attribute;
    rr:objectMap [ rr:constant sdmxa:unitMeasure]
  ];
.
<#Component - obsValue>
  rr:subjectMap[
    rr:class qb:ComponentProperty;
    rr:constant <http://statspace.linkedwidgets.org/dataset/WorldBank/dsd-obsValue>
  ];
  rr:predicateObjectMap [
    rr:predicate qb:measure;
    rr:objectMap [ rr:constant sdmxm:obsValue]
  ];
.
<#refPeriod>
  rr:subjectMap[
```

```
    rr:class qb:DimensionProperty;
    rr:constant sdmxd:refPeriod
  ];
.
<#refArea>
  rr:subjectMap[
    rr:class qb:DimensionProperty;
    rr:constant sdmxd:refArea
  ];
.
<#unitMeasure>
  rr:subjectMap[
    rr:class qb:AttributeProperty;
    rr:constant sdmxa:unitMeasure
  ];
.
<#ObsValue>
  rr:subjectMap[
    rr:class qb:MeasureProperty;
    rr:constant sdmxm:obsValue
  ];
.
```

Listing B.1: RML mapping for the World Bank data source

C RML Mapping for ONS Data

C.1 RML Mapping for the first Spreadsheet Data Set Collected

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix rml: <http://semweb.mmlab.be/ns/rml#>.
@prefix ql: <http://semweb.mmlab.be/ns/ql#>.
@prefix rr: <http://www.w3.org/ns/r2rml#>.
@prefix sdmx-dimension: <http://purl.org/linked-data/sdmx/2009/dimension#>.
@prefix sdmx-measure: <http://purl.org/linked-data/sdmx/2009/measure#>.
@prefix sdmx-attribute: <http://purl.org/linked-data/sdmx/2009/attribute#>.
@prefix sdmx-code: <http://purl.org/linked-data/sdmx/2009/code#>.
@prefix qb: <http://purl.org/linked-data/cube#>.
@prefix dcterms: <http://purl.org/dc/terms/>.
<#Observation>
  rml:logicalSource [
    rml:iterator "Table!A5:B54:A3:B3";
    rml:referenceFormulation ql:Spreadsheet;
    rml:source "http://www.ons.gov.uk/ons/rel/pop-estimate/population-estimates-for-uk
        --england-and-wales--scotland-and-northern-ireland/2013/chd-1-for-story.xls"
  ];
  rr:subjectMap[
    rr:class qb:Observation;
    rr:template "http://statspace.linkedwidgets.org/dataset/ONS-Population-change/Obs-{
        Mid-Year}"; rr:termType rr:IRI
  ];
  rr:predicateObjectMap [
    rr:predicate sdmx-dimension:refArea;
    rr:objectMap [
      rr:constant <http://statspace.linkedwidgets.org/codelist/cl_area/UnitedKingdom>
    ]
```

```
  ];
  rr:predicateObjectMap [
    rr:predicate sdmx-dimension:refPeriod;
    rr:objectMap [
      rr:template "http://reference.data.gov.uk/id/gregorian-year/{Mid-Year}"; rr:
          termType rr:IRI
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate sdmx-attribute:unitMeasure;
    rr:objectMap [
      rr:constant <http://statspace.linkedwidgets.org/codelist/cl_unitMeasure/P6>
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate sdmx-measure:obsValue;
    rr:objectMap [
      rml:reference "Mid-Year␣Population␣(millions)"
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate qb:dataSet;
    rr:objectMap [
      rr:constant <http://statspace.linkedwidgets.org/dataset/ONS-Population-change>
    ]
  ];
.
<#Dataset>
  rr:subjectMap[
    rr:class qb:DataSet;
    rr:constant <http://statspace.linkedwidgets.org/dataset/ONS-Population-change>
  ];
  rr:predicateObjectMap [
    rr:predicate rdfs:label;
    rr:objectMap [
      rr:constant "Mid-year␣population␣estimates␣for␣the␣UK␣mid-1964␣onwards"
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate dcterms:subject;
    rr:objectMap [
      rr:constant <http://statspace.linkedwidgets.org/codelist/cl_subject/SP.POP.TOTL>
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate qb:structure;
    rr:objectMap [
      rr:constant <http://statspace.linkedwidgets.org/dataset/ONS-Population-change/dsd
          >
    ];
  ];
.
<#DataStructure>
  rr:subjectMap[
    rr:class qb:DataStructureDefinition;
    rr:constant <http://statspace.linkedwidgets.org/dataset/ONS-Population-change/dsd>
  ];
  rr:predicateObjectMap [
    rr:predicate qb:component;
    rr:objectMap [
      rr:constant <http://statspace.linkedwidgets.org/dataset/ONS-Population-change/dsd
          -refArea>
```

```
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate qb:component;
    rr:objectMap [
      rr:constant       <http://statspace.linkedwidgets.org/dataset/ONS-Population-
          change/dsd-refPeriod>
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate qb:component;
    rr:objectMap [
      rr:constant       <http://statspace.linkedwidgets.org/dataset/ONS-Population-
          change/dsd-unitMeasure>
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate qb:component;
    rr:objectMap [
      rr:constant <http://statspace.linkedwidgets.org/dataset/ONS-Population-change/dsd
          -obsValue>
    ]
  ];
.
<#Component - refArea>
  rr:subjectMap[
    rr:class qb:ComponentProperty;
    rr:constant <http://statspace.linkedwidgets.org/dataset/ONS-Population-change/dsd-
        refArea>
  ];
  rr:predicateObjectMap [
    rr:predicate qb:dimension;
    rr:objectMap [
      rr:constant sdmx-dimension:refArea
   ]
  ];
.
<#Component - refPeriod>
  rr:subjectMap[
    rr:class qb:ComponentProperty;
    rr:constant <http://statspace.linkedwidgets.org/dataset/ONS-Population-change/dsd-
        refPeriod>
  ];
  rr:predicateObjectMap [
    rr:predicate qb:dimension;
    rr:objectMap [
      rr:constant sdmx-dimension:refPeriod
    ]
  ];
.
<#Component - unitMeasure>
  rr:subjectMap[
    rr:class qb:ComponentProperty;
    rr:constant <http://statspace.linkedwidgets.org/dataset/ONS-Population-change/dsd-
        unitMeasure>
  ];
  rr:predicateObjectMap [
    rr:predicate qb:attribute;
    rr:objectMap [
      rr:constant sdmx-attribute:unitMeasure
    ]
  ];
```

```
.
<#Component - obsValue>
  rr:subjectMap[
    rr:class qb:ComponentProperty;
    rr:constant <http://statspace.linkedwidgets.org/dataset/ONS-Population-change/dsd-
        obsValue>
  ];
  rr:predicateObjectMap [
    rr:predicate qb:measure;
    rr:objectMap [
      rr:constant sdmx-measure:obsValue
    ]
  ];
.
<#refArea>
  rr:subjectMap[
    rr:class qb:DimensionProperty;
    rr:constant sdmx-dimension:refArea
  ];
.
<#refPeriod>
  rr:subjectMap[
    rr:class qb:DimensionProperty;
    rr:constant sdmx-dimension:refPeriod
  ];
.
<#unitMeasure>
  rr:subjectMap[
    rr:class qb:AttributeProperty;
    rr:constant sdmx-attribute:unitMeasure
  ];
.
<#ObsValue>
  rr:subjectMap[
    rr:class qb:MeasureProperty;
    rr:constant sdmx-measure:obsValue
  ];
.
```

Listing C.2: RML mapping for the first data set collected

C.2 RML Mapping for the second Spreadsheet Data Set Collected

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix rml: <http://semweb.mmlab.be/ns/rml#>.
@prefix ql: <http://semweb.mmlab.be/ns/ql#>.
@prefix rr: <http://www.w3.org/ns/r2rml#>.
@prefix sdmx-dimension: <http://purl.org/linked-data/sdmx/2009/dimension#>.
@prefix sdmx-measure: <http://purl.org/linked-data/sdmx/2009/measure#>.
@prefix sdmx-attribute: <http://purl.org/linked-data/sdmx/2009/attribute#>.
@prefix sdmx-code: <http://purl.org/linked-data/sdmx/2009/code#>.
@prefix qb: <http://purl.org/linked-data/cube#>.
@prefix dcterms: <http://purl.org/dc/terms/>.
<#Observation1>
  rml:logicalSource [
    rml:referenceFormulation  ql:Spreadsheet2;
```

```
    rml:source "http://www.ons.gov.uk/ons/about-ons/business-transparency/freedom-of-
        information/what-can-i-request/published-ad-hoc-data/pop/july-2015/uk-
        population-estimates-1851-2014.xls";
    rml:iterator "UK␣Quinary␣1953-1970!C3:U21:C2:C2:D2:U2";
  ];
  rr:subjectMap [
    rr:class qb:Observation;
    rr:template "http://statspace.linkedwidgets.org/dataset/ONS-Population-Total
        -1953-1970/Obs-Total-Y{Age}-{ql:Spreadsheet2!Header}"; rr:termType rr:IRI
  ];
  rr:predicateObjectMap [
    rr:predicate sdmx-dimension:refArea;
    rr:objectMap [
      rr:constant <http://statspace.linkedwidgets.org/codelist/cl_area/UnitedKingdom>
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate sdmx-dimension:refPeriod;
    rr:objectMap [
      rr:template "http://statspace.linkedwidgets.org/codelist/cl_period/tmp/{ql:
          Spreadsheet2!Header}"; rr:termType rr:IRI
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate sdmx-dimension:age;
    rr:objectMap [
      rr:template "http://statspace.linkedwidgets.org/codelist/cl_age/tmp/Y{Age}"; rr:
          termType rr:IRI
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate sdmx-dimension:sex;
    rr:objectMap [
      rr:constant sdmx-code:sex-T
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate sdmx-attribute:unitMeasure;
    rr:objectMap [
      rr:constant <http://statspace.linkedwidgets.org/codelist/cl_unitMeasure/P3>
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate sdmx-measure:obsValue;
    rr:objectMap [
      rml:reference "ql:Spreadsheet2!Value"
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate qb:dataSet;
    rr:objectMap [
      rr:constant <http://statspace.linkedwidgets.org/dataset/ONS-Population-Total
          -1953-1970>
    ]
  ];
.
<#Dataset>
  rr:subjectMap[
    rr:class qb:DataSet;
    rr:constant <http://statspace.linkedwidgets.org/dataset/ONS-Population-Total
        -1953-1970>
  ];
```

```
  rr:predicateObjectMap [
    rr:predicate rdfs:label;
    rr:objectMap [
      rr:constant "Population␣estimates␣for␣United␣Kingdom␣by␣age␣-␣1953␣to␣1970"
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate dcterms:subject;
    rr:objectMap [
      rr:constant <http://statspace.linkedwidgets.org/codelist/cl_subject/SP.POP.AGES>
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate qb:structure;
    rr:objectMap [
      rr:constant <http://statspace.linkedwidgets.org/dataset/ONS-Population-Total
          -1953-1970/dsd>
    ];
  ];
.
<#DataStructure>
  rr:subjectMap[
    rr:class qb:DataStructureDefinition;
    rr:constant <http://statspace.linkedwidgets.org/dataset/ONS-Population-Total
        -1953-1970/dsd>
  ];
  rr:predicateObjectMap [
    rr:predicate qb:component;
    rr:objectMap [
      rr:constant <http://statspace.linkedwidgets.org/dataset/ONS-Population-Total
          -1953-1970/dsd-refArea>
  ]
];
  rr:predicateObjectMap [
    rr:predicate qb:component;
    rr:objectMap [
      rr:constant       <http://statspace.linkedwidgets.org/dataset/ONS-Population-
          Total-1953-1970/dsd-refPeriod>
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate qb:component;
    rr:objectMap [
      rr:constant       <http://statspace.linkedwidgets.org/dataset/ONS-Population-
          Total-1953-1970/dsd-age>
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate qb:component;
    rr:objectMap [
      rr:constant       <http://statspace.linkedwidgets.org/dataset/ONS-Population-
          Total-1953-1970/dsd-sex>
    ]
  ];
  rr:predicateObjectMap [
    rr:predicate qb:component;
      rr:objectMap [
      rr:constant       <http://statspace.linkedwidgets.org/dataset/ONS-Population-
          Total-1953-1970/dsd-unitMeasure>
    ]
  ];
  rr:predicateObjectMap [
```

```
    rr:predicate qb:component;
    rr:objectMap [
      rr:constant <http://statspace.linkedwidgets.org/dataset/ONS-Population-Total
          -1953-1970/dsd-obsValue>
    ]
  ];
.
<#Component - refArea>
  rr:subjectMap[
    rr:class qb:ComponentProperty;
    rr:constant <http://statspace.linkedwidgets.org/dataset/ONS-Population-Total
        -1953-1970/dsd-refArea>
  ];
  rr:predicateObjectMap [
    rr:predicate qb:dimension;
    rr:objectMap [
      rr:constant sdmx-dimension:refArea
    ]
  ];
.
<#Component - refPeriod>
  rr:subjectMap[
    rr:class qb:ComponentProperty;
    rr:constant <http://statspace.linkedwidgets.org/dataset/ONS-Population-Total
        -1953-1970/dsd-refPeriod>
  ];
  rr:predicateObjectMap [
    rr:predicate qb:dimension;
    rr:objectMap [
      rr:constant sdmx-dimension:refPeriod
    ]
  ];
.
<#Component - age>
  rr:subjectMap[
    rr:class qb:ComponentProperty;
    rr:constant <http://statspace.linkedwidgets.org/dataset/ONS-Population-Total
        -1953-1970/dsd-age>
  ];
  rr:predicateObjectMap [
    rr:predicate qb:dimension;
    rr:objectMap [
      rr:constant sdmx-dimension:age
    ]
  ];
.
<#Component - sex>
  rr:subjectMap[
    rr:class qb:ComponentProperty;
    rr:constant <http://statspace.linkedwidgets.org/dataset/ONS-Population-Total
        -1953-1970/dsd-sex>
  ];
  rr:predicateObjectMap [
    rr:predicate qb:dimension;
    rr:objectMap [
      rr:constant sdmx-dimension:sex
    ]
  ];
.
<#Component - unitMeasure>
  rr:subjectMap[
    rr:class qb:ComponentProperty;
```

```
    rr:constant <http://statspace.linkedwidgets.org/dataset/ONS-Population-Total
        -1953-1970/dsd-unitMeasure>
  ];
  rr:predicateObjectMap [
    rr:predicate qb:attribute;
    rr:objectMap [
      rr:constant sdmx-attribute:unitMeasure
    ]
  ];
.
<#Component - obsValue>
  rr:subjectMap[
    rr:class qb:ComponentProperty;
    rr:constant <http://statspace.linkedwidgets.org/dataset/ONS-Population-Total
        -1953-1970/dsd-obsValue>
  ];
  rr:predicateObjectMap [
    rr:predicate qb:measure;
    rr:objectMap [
      rr:constant sdmx-measure:obsValue
    ]
  ];
.
<#refPeriod>
  rr:subjectMap[
    rr:class qb:DimensionProperty;
    rr:constant sdmx-dimension:refPeriod
  ];
.
<#refArea>
  rr:subjectMap[
    rr:class qb:DimensionProperty;
    rr:constant sdmx-dimension:refArea
  ];
.
<#age>
  rr:subjectMap[
    rr:class qb:DimensionProperty;
    rr:constant sdmx-dimension:age
  ];
.
<#sex>
  rr:subjectMap[
    rr:class qb:DimensionProperty;
    rr:constant sdmx-dimension:sex
  ];
.
<#unitMeasure>
  rr:subjectMap[
    rr:class qb:AttributeProperty;
    rr:constant sdmx-attribute:unitMeasure
  ];
.
<#ObsValue>
  rr:subjectMap[
    rr:class qb:MeasureProperty;
    rr:constant sdmx-measure:obsValue
  ];
.
```

Listing C.3: RML mapping for the second data set collected

D RML Queries used in Evaluation

D.1 RML Queries used to transform data of one country into RDF

No.	Query
1	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=TC&indicator=SE.TER.ENRL.TC.ZS&cache=no
2	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=EU&indicator=DC.DAC.DEUL.CD&cache=no
3	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=KR&indicator=DT.TDS.PNGB.CD&cache=no
4	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=MA&indicator=IC.ELC.OUTG&cache=no
5	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=BY&indicator=SL.FAM.0714.MA.ZS&cache=no
6	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=MH&indicator=DT.DOD.OFFT.CD&cache=no
7	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=VN&indicator=SL.TLF.0714.SW.MA.TM&cache=no
8	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=PY&indicator=DT.DXR.DPPG.CD&cache=no
9	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=MV&indicator=DT.MAT.PRVT&cache=no
10	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=UA&indicator=EN.ATM.CO2E.SF.ZS&cache=no
11	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=DK&indicator=FM.LBL.BMNY.GD.ZS&cache=no
12	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=4E&indicator=DT.NFL.UNRW.CD&cache=no

```
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
13  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=SA&indicator=DT.INT.PCBK.CD&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
14  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=WS&indicator=NY.GDP.FCST.CN&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
15  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=ZA&indicator=FS.AST.DOMS.GD.ZS&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
16  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=XE&indicator=SH.STA.OWGH.ZS&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
17  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=BB&indicator=BM.KLT.DINV.CD.WD&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
18  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=KW&indicator=NE.TRD.GNFS.ZS&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
19  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=LR&indicator=DT.DOD.MLTC.CD&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
20  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=GD&indicator=SP.UWT.TFRT&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
21  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=LI&indicator=SH.DYN.MORT.MA&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
22  statspace.linkedwidgets.org/mapping/wb.ttl&refArea=BR&
    indicator=SL.TLF.CACT.FM.NE.ZS&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
23  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=AT&indicator=BM.GSR.GNFS.CD&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
24  statspace.linkedwidgets.org/mapping/wb.ttl&refArea=MP&
    indicator=EN.ATM.METH.EG.KT.CE&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
25  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=FJ&indicator=NE.EXP.GNFS.KN&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
26  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=LU&indicator=SE.XPD.CPRM.ZS&cache=no
```

```
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
27  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=CD&indicator=SP.DYN.IMRT.FE.IN&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
28  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=MA&indicator=TM.VAL.MRCH.R3.ZS&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
29  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=BE&indicator=BN.CAB.XOKA.CD&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
30  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=PA&indicator=DC.DAC.LUXL.CD&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
31  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=OE&indicator=SE.SEC.ENRR.FE&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
32  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=TR&indicator=IC.TAX.METG&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
33  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=UY&indicator=SI.DST.03RD.20&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
34  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=MT&indicator=SH.STA.ACSN.UR&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
35  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=GN&indicator=SL.TLF.ACTI.MA.ZS&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
36  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=XD&indicator=SE.PRM.NENR.MA&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
37  statspace.linkedwidgets.org/mapping/wb.ttl&refArea=ZG&
    indicator=TM.TAX.MANF.SM.FN.ZS&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
38  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=VI&indicator=EN.HPT.THRD.NO&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
39  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=CL&indicator=DT.INT.DLXF.CD&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
40  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=ME&indicator=SE.PRM.TENR.FE&cache=no
```

```
41 http://statspace.linkedwidgets.org/rml?rmlsource=http://
   statspace.linkedwidgets.org/mapping/
   wb.ttl&refArea=CZ&indicator=DT.AMT.DPNG.CD&cache=no
42 http://statspace.linkedwidgets.org/rml?rmlsource=http://
   statspace.linkedwidgets.org/mapping/
   wb.ttl&refArea=SB&indicator=EG.GDP.PUSE.KO.PP&cache=no
43 http://statspace.linkedwidgets.org/rml?rmlsource=http://
   statspace.linkedwidgets.org/mapping/
   wb.ttl&refArea=CV&indicator=NY.GDP.NGAS.RT.ZS&cache=no
44 http://statspace.linkedwidgets.org/rml?rmlsource=http://
   statspace.linkedwidgets.org/mapping/
   wb.ttl&refArea=FR&indicator=NY.ADJ.DMIN.CD&cache=no
45 http://statspace.linkedwidgets.org/rml?rmlsource=http://
   statspace.linkedwidgets.org/mapping/
   wb.ttl&refArea=IE&indicator=NE.GDI.FTOT.KN&cache=no
46 http://statspace.linkedwidgets.org/rml?rmlsource=http://
   statspace.linkedwidgets.org/mapping/
   wb.ttl&refArea=SA&indicator=SH.STA.OWGH.ZS&cache=no
47 http://statspace.linkedwidgets.org/rml?rmlsource=http://
   statspace.linkedwidgets.org/mapping/
   wb.ttl&refArea=DM&indicator=TM.VAL.MMTL.ZS.UN&cache=no
48 http://statspace.linkedwidgets.org/rml?rmlsource=http://
   statspace.linkedwidgets.org/mapping/wb.ttl&refArea=AR&
   indicator=per_allsp.adq_pop_tot&cache=no
49 http://statspace.linkedwidgets.org/rml?rmlsource=http://
   statspace.linkedwidgets.org/mapping/
   wb.ttl&refArea=PA&indicator=SM.POP.REFG.OR&cache=no
50 http://statspace.linkedwidgets.org/rml?rmlsource=http://
   statspace.linkedwidgets.org/mapping/
   wb.ttl&refArea=PL&indicator=DT.DIS.IDAG.CD&cache=no
51 http://statspace.linkedwidgets.org/rml?rmlsource=http://
   statspace.linkedwidgets.org/mapping/uk0.ttl&cache=no
52 http://statspace.linkedwidgets.org/rml?rmlsource=http://
   statspace.linkedwidgets.org/mapping/uk1.ttl&cache=no
53 http://statspace.linkedwidgets.org/rml?rmlsource=http://
   statspace.linkedwidgets.org/mapping/uk2.ttl&cache=no
54 http://statspace.linkedwidgets.org/rml?rmlsource=http://
   statspace.linkedwidgets.org/mapping/uk3.ttl&cache=no
55 http://statspace.linkedwidgets.org/rml?rmlsource=http://
   statspace.linkedwidgets.org/mapping/uk4.ttl&cache=no
56 http://statspace.linkedwidgets.org/rml?rmlsource=http://
   statspace.linkedwidgets.org/mapping/uk5.ttl&cache=no
```

57	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/uk6.ttl&cache=no
58	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/uk7.ttl&cache=no

Table D.7: RML Queries used to transform data of one country into RDF

D.2 RML Queries used to transform data of all countries into RDF

No.	Query
1	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=all&indicator=SE.TER.ENRL.TC.ZS&cache=no
2	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=all&indicator=DC.DAC.DEUL.CD&cache=no
3	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=all&indicator=DT.TDS.PNGB.CD&cache=no
4	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=all&indicator=IC.ELC.OUTG&cache=no
5	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=all&indicator=SL.FAM.0714.MA.ZS&cache=no
6	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=all&indicator=DT.DOD.OFFT.CD&cache=no
7	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=all&indicator=SL.TLF.0714.SW.MA.TM&cache=no
8	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=all&indicator=DT.DXR.DPPG.CD&cache=no
9	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=all&indicator=DT.MAT.PRVT&cache=no
10	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=all&indicator=EN.ATM.CO2E.SF.ZS&cache=no
11	http://statspace.linkedwidgets.org/rml?rmlsource=http://statspace.linkedwidgets.org/mapping/wb.ttl&refArea=all&indicator=FM.LBL.BMNY.GD.ZS&cache=no

```
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
12  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=DT.NFL.UNRW.CD&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
13  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=DT.INT.PCBK.CD&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
14  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=NY.GDP.FCST.CN&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
15  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=FS.AST.DOMS.GD.ZS&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
16  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=SH.STA.OWGH.ZS&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
17  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=BM.KLT.DINV.CD.WD&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
18  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=NE.TRD.GNFS.ZS&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
19  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=DT.DOD.MLTC.CD&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
20  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=SP.UWT.TFRT&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
21  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=SH.DYN.MORT.MA&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
22  statspace.linkedwidgets.org/mapping/wb.ttl&refArea=all&
    indicator=SL.TLF.CACT.FM.NE.ZS&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
23  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=BM.GSR.GNFS.CD&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
24  statspace.linkedwidgets.org/mapping/wb.ttl&refArea=all&
    indicator=EN.ATM.METH.EG.KT.CE&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
25  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=NE.EXP.GNFS.KN&cache=no
```

```
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
26  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=SE.XPD.CPRM.ZS&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
27  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=SP.DYN.IMRT.FE.IN&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
28  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=TM.VAL.MRCH.R3.ZS&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
29  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=BN.CAB.XOKA.CD&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
30  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=DC.DAC.LUXL.CD&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
31  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=SE.SEC.ENRR.FE&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
32  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=IC.TAX.METG&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
33  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=SI.DST.03RD.20&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
34  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=SH.STA.ACSN.UR&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
35  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=SL.TLF.ACTI.MA.ZS&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
36  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=SE.PRM.NENR.MA&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
37  statspace.linkedwidgets.org/mapping/wb.ttl&refArea=all&
    indicator=TM.TAX.MANF.SM.FN.ZS&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
38  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=EN.HPT.THRD.NO&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
39  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=DT.INT.DLXF.CD&cache=no
```

```
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
40  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=SE.PRM.TENR.FE&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
41  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=DT.AMT.DPNG.CD&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
42  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=EG.GDP.PUSE.KO.PP&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
43  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=NY.GDP.NGAS.RT.ZS&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
44  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=NY.ADJ.DMIN.CD&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
45  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=NE.GDI.FTOT.KN&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
46  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=SH.STA.OWGH.ZS&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
47  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=TM.VAL.MMTL.ZS.UN&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
48  statspace.linkedwidgets.org/mapping/wb.ttl&refArea=all&
    indicator=per_allsp.adq_pop_tot&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
49  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=SM.POP.REFG.OR&cache=no
    http://statspace.linkedwidgets.org/rml?rmlsource=http://
50  statspace.linkedwidgets.org/mapping/
    wb.ttl&refArea=all&indicator=DT.DIS.IDAG.CD&cache=no
```

Table D.8: RML Queries used to transform data of all countries into RDF

E Runtime of RML mapping service in Evaluation

E.1 Time Consumption for Data Transformation of one Country

Query	1st	2nd	3rd
1	1701	2064	1666

2	1089	1317	1088
3	1079	1374	1103
4	1075	1253	1060
5	1198	1370	1175
6	1057	1208	1059
7	1251	1351	1195
8	1023	1333	1037
9	1061	1525	1082
10	1050	1203	1033
11	1221	1227	1102
12	1348	1416	1166
13	1188	1171	1039
14	1014	1390	1017
15	1016	1242	1083
16	1166	1445	1114
17	1402	1377	1108
18	1244	1173	996
19	1063	1227	1062
20	1016	1192	1033
21	1044	1192	1053
22	1226	1574	1103
23	1287	1389	1048
24	1141	1283	1134
25	1066	1193	1049
26	1122	1301	1165
27	1079	1235	1063
28	1311	1499	1239
29	1139	1229	991
30	1186	1181	1049
31	1019	1231	1075
32	1002	1130	1017
33	1078	1212	1047
34	1000	1164	1044
35	1224	1355	1111
36	1202	1175	1017
37	1103	1428	1163
38	1016	1181	1024
39	1019	1226	1025
40	1041	1249	1026

41	1211	1271	1179
42	1032	1230	1021
43	1022	1306	1071
44	1001	1239	1021
45	1010	1207	1049
46	989	1145	1002
47	1378	1252	1040
48	1113	1293	1120
49	1046	1411	1054
50	1052	1196	1014
51	862	1531	1182
52	729	635	531
53	363	487	389
54	586	546	509
55	688	811	722
56	593	768	687
57	759	674	862
58	579	441	304

Table E.9: Time consumption for data transformation of one country into RDF

E.2 Time Consumption for Data Transformation of all Countries

Query	1st	2nd	3rd
1	5644	6137	6145
2	4495	5406	5043
3	5056	4412	5495
4	4283	4193	5671
5	4177	5249	5038
6	5684	4242	5329
7	4080	5354	5831
8	4192	6042	5115
9	4999	4881	4152
10	4219	6185	4234
11	4036	4385	6504
12	6466	4667	4251
13	4062	4714	5867
14	5483	5993	4562
15	4506	6038	4900
16	5820	5253	6168

17	4848	5267	6024
18	3203	6534	5611
19	3956	6399	4425
20	3274	7405	4357
21	4548	4399	6326
22	4798	4560	4481
23	4519	5226	6229
24	5886	6655	4844
25	4227	4982	5761
26	4068	5130	6115
27	4383	4869	5686
28	4384	5363	5475
29	3130	6545	5712
30	4109	4826	4624
31	5219	6211	4655
32	4035	4786	5653
33	5277	5914	4260
34	3359	7278	5936
35	4709	6127	6701
36	3424	4852	4539
37	6562	6197	5959
38	5899	5358	5840
39	3172	4454	5872
40	3171	4632	4454
41	5871	4600	4506
42	4670	5233	5940
43	6138	4465	4545
44	4259	6021	5574
45	5301	5808	5869
46	4825	3365	4436
47	5368	6608	5413
48	4713	6265	5619
49	4762	5031	4496
50	5037	6217	4531

Table E.10: Time consumption for data transformation of all countries into RDF

List of Figures

2.1 Overview of the QB [16] . . . 11
2.2 A representation according to the QB (excerpt) . . . 11
2.3 A Generic Data Integration Architecture [51] . . . 13
2.4 A Data Integration Architecture for Enterprise Information Systems [54] . . . 14
2.5 A Data Integration Architecture for Multimedia Sources [55] . . . 15
2.6 An Architecture for Exploring Statistical Data [57, 58] . . . 16
2.7 Communication between components in search scenario [57] . . . 16

3.1 Architecture overview . . . 18
3.2 Structure of metadata . . . 22
3.3 An example of metadata description (excerpt) . . . 23
3.4 Communication between components to answer an input SPARQL query . . . 26
3.5 Communication between components to satisfy users' access needs . . . 27

4.1 An example of spatial value mapping . . . 42
4.3 Query interface of the mediator . . . 48
4.4 An example of equivalent relationships used by the mediator . . . 48
4.5 The explorer interface: Search field (1); Filters (2); List of results (3) . . . 49
4.6 Visualization of a raw data set . . . 51
4.7 Visualization of an RDF data set . . . 51
4.8 Interface for visualization of a metadata description . . . 52
4.9 Visualization used in data integration . . . 52
4.10 Illustrations of requirements for data integration . . . 55
4.11 Result of the example query . . . 56
4.12 An example of out-of-date data of the EEA . . . 57
4.13 Relationship between Inflation and Unemployment indicators for Japan . . . 58
4.14 Exploration of a geographical area . . . 60
4.15 Statistical data comparison of different areas . . . 61

5.1 Elapsed time for raw data transformation in the first experiment . . . 64
5.2 Elapsed time for raw data transformation in the second experiment . . . 65
5.3 Elapsed time for raw data transformation in the third experiment . . . 65
5.4 The interface of the coreference resolution service . . . 68
5.5 The interface of application used to collect users' evaluations . . . 69

List of Tables

2.1 An example of data structure definition file (excerpt) [42] 9
2.2 Example data set of the life expectancy (excerpt) [16] 11

3.1 Summary of components and code lists reused 21

4.1 An example spreadsheet data set uses the first layout 31
4.2 An example of spreadsheet data set uses the second layout 31
4.3 Shared URIs used to represent dimensions 33
4.4 Shared URI used to represent measure 33
4.5 Shared URI used to represent attribute 34
4.6 Characteristics of code lists 34
4.9 SPARQL endpoints used for testing algorithms 37
4.10 Keywords used for the recognition of components 39
4.11 UK population data in spreadsheet format (excerpt) 53

5.1 Characteristics of queries 64
5.2 A comparison of size of code lists in data sources 66
5.3 Sources and numbers of data sets covered 66

6.1 Key characteristics of mapping languages
✓: Yes; ✓*: Possible; ✗: No 73
6.2 A comparison between approaches of link discovery (based on the survey of link discovery frameworks [99]). ✓: Yes; ✗: No; –: Undefined 75
6.3 A comparison between coreference resolution services
✓: Yes; ✗: No 75
6.4 A comparison between approaches of statistical data integration
✓: Yes; ✗: No; –: Undefined 79
6.5 Key characteristics of applications of statistical data exploration
✓: Yes; ✗: No 82

A.1 Patterns designed for code list of temporal dimension 88
A.2 Patterns for the code list of age dimension 88
A.3 URIs in the code list of education level dimension 89
A.4 URIs in the code list of civil status dimension 91

A.5 URIs in the code list of frequency dimension 92
A.6 URIs in the code list of sex dimension 92
D.7 RML Queries used to transform data of one country into RDF 110
D.8 RML Queries used to transform data of all countries into RDF 113
E.9 Time consumption for data transformation of one country into RDF 115
E.10 Time consumption for data transformation of all countries into RDF 116

List of Listings

2.1 Reference exchange rate of the ECB (excerpt) [42] 10
4.1 An example input data set 30
4.2 Excerpt from RML mapping[1] 30
4.3 An example query sent to the RML mapping service 30
4.4 An example of iteration declaration used for the first layout 31
4.5 An example of iteration declaration used for the second layout 32
4.6 An RML mapping contains variables (excerpt) 32
4.7 Queries used to identify statistical data sets 37
4.8 An example concerning a geocoding request 40
4.9 An example about geocoding response (excerpt) 41
4.10 Query all data sets in the metadata repository 47
4.11 Query information concerning a specific metadata description 47
4.12 World Bank population data for the UK in XML format (excerpt) 53
4.13 Example input query for cross-data set population comparison 55
4.14 EEA data set query generated by the mediator 56
4.15 World Bank data set query generated by the mediator 56
4.16 UK data set query generated by the mediator 56
B.1 RML mapping for the World Bank data source 95
C.2 RML mapping for the first data set collected 98
C.3 RML mapping for the second data set collected 101

List of Algorithms

1 Geographical Area Mapping 44
2 Temporal Value Mapping 45

List of Abbreviations

APIs	Application Programming Interfaces.
CSO	Central Statistics Office of Ireland.
CSV	Comma Separated Values.
EC	European Commission.
ECB	European Central Bank.
EEA	European Environment Agency.
ETL	Extract Transform Load.
EUODP	European Union Open Data Portal.
IMF	International Monetary Fund.
IRIs	Internationalized Resource Identifiers.
JSON	JavaScript Object Notation.
MDX	Multidimensional Expression Language.
NUTS	Nomenclature of territorial units for statistics.
ODC	Open Data Communities.
OLAP	Online Analytical Processing.
ONS	Office for National Statistics.
OWL	Web Ontology Language.
QB	RDF Data Cube Vocabulary.

RML	RDF Mapping Language.
ScotStat	Scottish Statistics.
SDMX	Statistical Data and Metadata eXchange.
SPARQL	SPARQL Protocol and RDF Query Language.
SQL	Structured Query Language.
UK	United Kingdom.
URIs	Uniform Resource Identifiers.
VOGD	Vienna Open Government Data.
WB	Work Bank.
XLS	Microsoft Excel Spreadsheet.
XML	Extensible Markup Language.

Bibliography

[1] Ramine Tinati, Les Carr, Susan Halford, and Catherine Pope. Exploring the impact of adopting open data in the UK government. *Digital Futures*, 2012.

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

[3] Evangelos Kalampokis, Efthimios Tambouris, and Konstantinos Tarabanis. Linked Open Government Data Analytics. In Maria A. Wimmer, Marijn Janssen, and Hans J. Scholl, editors, *Electronic Government: 12th IFIP WG 8.5 International Conference, EGOV 2013*, volume 8074, pages 99–110. Springer Berlin Heidelberg, 2013.

[4] Barbara Ubaldi. Open Government Data. OECD Working Papers on Public Governance 22, OECD, 2013.

[5] Rob Kitchin. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. SAGE, August 2014.

[6] Evangelos Kalampokis, Andriy Nikolov, Peter Haase, Richard Cyganiak, Arkadiusz Stasiewicz, Areti Karamanou, Maria Zotou, Dimitris Zeginis, Efthimios Tambouris, and Konstantinos Tarabanis. Exploiting linked data cubes with opencube toolkit. In *ISWC: Posters & Demonstrations Track*, pages 137–140. CEUR-WS. org, 2014.

[7] Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. Adoption of the Linked Data Best Practices in Different Topical Domains. In Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandčić, Paul Groth, Natasha Noy, Krzysztof Janowicz, and Carole Goble, editors, *The Semantic Web – ISWC*, pages 245–260. Springer International Publishing, 2014.

[8] Evangelos Kalampokis, Bill Roberts, Areti Karamanou, Efthimios Tambouris, and Konstantinos Tarabanis. Challenges on Developing Tools for Exploiting Linked Open Data Cubes. In *Proceedings of the 3rd International Workshop on Semantic Statistics*. CEUR-WS.org, 2015.

[9] Alain Berro, Imen Megdiche, and Olivier Teste. A Content-Driven ETL Processes for Open Data. In Nick Bassiliades, Mirjana Ivanovic, Margita Kon-Popovska, Yannis Manolopoulos, Themis Palpanas, Goce Trajcevski, and Athena Vakali, editors, *New Trends in Database and Information Systems II: Selected Papers of the 18th East European Conference on Advances in Databases and Information Systems and Associated Satellite Events*, pages 29–40. Springer International Publishing, 2015.

[10] Efthimios Tambouris, Evangelos Kalampokis, and Konstantinos Tarabanis. Processing Linked Open Data Cubes. In Efthimios Tambouris, Marijn Janssen, Hans Jochen Scholl, Maria A. Wimmer, Konstantinos Tarabanis, Mila Gascó, Bram Klievink, Ida Lindgren, and Peter Parycek, editors, *Electronic Government: 14th IFIP WG 8.5 International Conference, EGOV 2015*, pages 130–143. Springer International Publishing, 2015.

[11] Christian Philipp Geiger and Jörn von Lucke. Open government and (linked)(open)(government)(data). *JeDEM-eJournal of eDemocracy and Open Government*, 4(2):265–278, 2012.

[12] Nigel Bowles, James T. Hamilton, and David Levy. *Transparency in Politics and the Media: Accountability and Open Government*. I.B.Tauris, November 2013.

[13] Irene Petrou, Marios Meimaris, and George Papastefanatos. Towards a methodology for publishing Linked Open Statistical Data. *JeDEM-eJournal of eDemocracy and Open Government*, 6(1):97–105, 2014.

[14] Keith Andrews, Thomas Traunmüller, Thomas Wolkinger, Eva Goldgruber, Robert Gutounig, and Julian Ausserhofer. Styrian Diversity Visualisation: Visualising Statistical Open Data with a Lean Web App and Data Server. In *Eurographics Conference on Visualization (EuroVis), Posters Track*, 2016.

[15] Antoine Isaac and Bernhard Haslhofer. Europeana linked open data–data. europeana. eu. *Semantic Web*, 4(3):291–297, 2013.

[16] Richard Cyganiak, Dave Reynolds, and Jeni Tennison. The RDF Data Cube Vocabulary, 2014.

[17] G. Antoniou and Frank Van Harmelen. *A Semantic Web Primer*. Cooperative information systems. MIT Press, 2nd edition, 2008.

[18] John Hebeler, Matthew Fisher, Ryan Blace, Andrew Perez-Lopez, and Mike Dean. *Semantic Web Programming*. Wiley, 2009.

[19] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):205–227, 2009.

[20] Peter Haase, Jeen Broekstra, Andreas Eberhart, and Raphael Volz. A Comparison of RDF Query Languages. In Sheila A. McIlraith, Dimitris Plexousakis, and Frank van Harmelen, editors, *The Semantic Web – ISWC 2004*, pages 502–517. Springer Berlin Heidelberg, 2004.

[21] Stefan Bischof, Stefan Decker, Thomas Krennwallner, Nuno Lopes, and Axel Polleres. Mapping between RDF and XML with XSPARQL. *Journal on Data Semantics*, 1(3):147–185, 2012.

[22] Hugh Glaser, Afraz Jaffri, and Ian Millard. Managing co-reference on the semantic web. In *Proceedings of Workshop on Linked Data on the Web*. CEUR-WS.org, 2009.

[23] Kai Schlegel, Florian Stegmaier, Sebastian Bayerl, Michael Granitzer, and Harald Kosch. Balloon fusion: SPARQL rewriting based on unified co-reference information. In *Proceedings of the IEEE 30th International Conference on Data Engineering Workshops*, pages 254–259. IEEE, 2014.

[24] Benedikt Kämpgen and Andreas Harth. Transforming Statistical Linked Data for Use in OLAP Systems. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 33–40. ACM, 2011.

[25] Benedikt Kämpgen, Steffen Stadtmüller, and Andreas Harth. Querying the Global Cube: Integration of Multidimensional Datasets from the Web. In Krzysztof Janowicz, Stefan Schlobach, Patrick Lambrix, and Eero Hyvönen, editors, *Proceedings of the 19th International Conference on Knowledge Engineering and Knowledge Management*, volume 8876, pages 250–265. Springer International Publishing, 2014.

[26] Percy E. Rivera Salas, Michael Martin, Fernando Maia Da Mota, Soren Auer, Karin Breitman, and Marco A. Casanova. Publishing Statistical Data on the Web. In *Proceedings of the Sixth International Conference on Semantic Computing*, volume 6, pages 285–292. IEEE, 2012.

[27] Fadi Maali, Gofran Shukair, and Nikolaos Loutas. A dynamic faceted browser for data cube statistical data. In *Proceedings of Workshop on Using Open Data*, 2012.

[28] Marta Sabou, Irem Arsal, and Adrian MP Braşoveanu. Tourmislod: A tourism linked data set. *Semantic Web*, 4(3):271–276, 2013.

[29] Patrick Hoefler, Michael Granitzer, Eduardo E. Veas, and Christin Seifert. Linked Data Query Wizard: A Novel Interface for Accessing SPARQL Endpoints. In *Proceedings of Workshop on Linked Data on the Web*. CEUR-WS.org, 2014.

[30] Jiří Helmich, Jakub Klímek, and Martin Nečaský. Visualizing RDF Data Cubes Using the Linked Data Visualization Model. In Valentina Presutti, Eva Blomqvist, Raphael Troncy, Harald Sack, Ioannis Papadakis, and Anna Tordai, editors, *The Semantic Web: ESWC 2014 Satellite Events*, volume 8798, pages 368–373. Springer International Publishing, 2014.

[31] Sarven Capadisli, Sören Auer, and Axel-Cyrille Ngonga Ngomo. Linked SDMX Data: Path to high fidelity Statistical Linked Data. *Semantic Web*, 6(2):105–112, 2015.

[32] Ken Peffers, Tuure Tuunanen, Marcus A Rothenberger, and Samir Chatterjee. A design science research methodology for information systems research. *Journal of management information systems*, 24(3):45–77, 2007.

[33] Anastasia Dimou, Miel Vander Sande, Pieter Colpaert, Ruben Verborgh, Erik Mannens, and Rik Van de Walle. RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data. In *Proceedings of Workshop on Linked Data on the Web*. CEUR-WS.org, 2014.

[34] Ba-Lam Do, Peter Wetz, Elmar Kiesling, Peb Ruswono Aryan, Tuan-Dat Trinh, and A. Min Tjoa. StatSpace: A Unified Platform for Statistical Data Exploration. In Christophe Debruyne, Hervé Panetto, Robert Meersman, Tharam Dillon, eva Kühn, Declan O'Sullivan, and Claudio Agostino Ardagna, editors, *On the Move to Meaningful Internet Systems: OTM 2016 Conferences: Confederated International Conferences: CoopIS, C&TC, and ODBASE 2016*, volume 10033, pages 792–809. Springer International Publishing, 2016.

[35] Ba-Lam Do, Peb Ruswono Aryan, Tuan-Dat Trinh, Peter Wetz, Elmar Kiesling, and A. Min Tjoa. Toward a framework for statistical data integration. In *Proceedings of the 3rd International Workshop on Semantic Statistics*. CEUR-WS.org, 2015.

[36] Ba-Lam Do, Tuan-Dat Trinh, Peb Ruswono Aryan, Peter Wetz, Elmar Kiesling, and A Min Tjoa. Toward a Statistical Data Integration Environment: The Role of Semantic Metadata. In *Proceedings of the 11th International Conference on Semantic Systems*, pages 25–32. ACM, 2015.

[37] Ba-Lam Do, Tuan-Dat Trinh, Peter Wetz, Elmar Kiesling, Amin Anjomshoaa, and A Min Tjoa. Multiscale Exploration of Spatial Statistical Datasets: A Linked Data Mashup Approach. In *Proceedings of the 2nd International Workshop on Semantic Statistics*. CEUR-WS.org, 2014.

[38] Ba-Lam Do, Tuan-Dat Trinh, Peter Wetz, Amin Anjomshoaa, Elmar Kiesling, and A. Min Tjoa. Widget-based Exploration of Linked Statistical Data Spaces. In *Proceedings of 3rd International Conference on Data Management Technologies and Application*, pages 282–290. SCITEPRESS, 2014.

[39] Gary Marchionini, Carol Hert, Ben Shneiderman, and Liz Liddy. E-Tables: Non-Specialist Use and Understanding of Statistical Data. *The School of Information Studies: Faculty Scholarship*, 2001.

[40] EuroStat and SDMX. Introduction to SDMX. SDMX self-learning package No. 1 Student book, 2010.

[41] United Nations Statistics Division. A brief introduction to SDMX Statistical Data and Metadata Exchange, 2016.

[42] ECB. SDMX tutorial: Using SDMX-ML to publish the euro foreign exchange reference rates on the ECB website, 2016.

[43] Maurizio Lenzerini. Data Integration Is Harder than You Thought. In Carlo Batini, Fausto Giunchiglia, Paolo Giorgini, and Massimo Mecella, editors, *International Conference on Cooperative Information Systems - CoopIS*, pages 22–26. Springer Berlin Heidelberg, 2001.

[44] Maurizio Lenzerini. Data Integration: A Theoretical Perspective. In *Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 233–246. ACM, 2002.

[45] Alon Halevy, Anand Rajaraman, and Joann Ordille. Data Integration: The Teenage Years. In *Proceedings of the 32nd International Conference on Very Large Data Bases*, VLDB '06, pages 9–16. VLDB Endowment, 2006.

[46] Saed Sayad. Data Exploration. `http://chem-eng.utoronto.ca/~datamining/Presentations/Data_Exploration.pdf`, 2010.

[47] Vijay Kotu and Bala Deshpande. *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. Morgan Kaufmann, 2014.

[48] Margaret Rouse. Data Exploration Definition. `http://searchbusinessanalytics.techtarget.com/definition/data-exploration`, 2015.

[49] Areti Karamanou, Evangelos Kalampokis, Efthimios Tambouris, and Konstantinos Tarabanis. Linked Data Cubes: Research results so far. In *Proceedings of the 4th International Workshop on Semantic Statistics*. CEUR-WS.org, 2016.

[50] Adrian MP Brasoveanua, Marta Sabou, Arno Scharla, Alexander Hubmann-Haidvogela, and Daniel Fischla. Visualizing statistical linked knowledge for decision support. *Semantic Web*, 2017.

[51] AnHai Doan, Alon Halevy, and Zachary Ives. *Principles of Data Integration*. Elsevier, 2012.

[52] Alon Y. Levy. The Information Manifold Approach to Data Integration. *IEEE Intelligent Systems*, 13:12–16, 1998.

[53] Hector Garcia-Molina, Jeffrey D. Ullman, and Jennifer Widom. *Database Systems: The Complete Book*. Prentice Hall Press, 2 edition, 2008.

[54] Soumaya El Kadiri, Bernard Grabot, Klaus-Dieter Thoben, Karl Hribernik, Christos Emmanouilidis, Gregor von Cieminski, and Dimitris Kiritsis. Current trends on ICT technologies for enterprise information systems. *Computers in Industry*, 79:14–33, June 2016.

[55] Domenico Beneventano, Claudio Gennaro, Sonia Bergamaschi, and Fausto Rabitti. A mediator-based approach for integrating heterogeneous multimedia sources. *Multimedia Tools and Applications*, 62(2):427–450, 2013.

[56] Giuseppe Amato, Claudio Gennaro, Fausto Rabitti, and Pasquale Savino. Milos: A Multimedia Content Management System for Digital Library Applications. In Rachel Heery and Liz Lyon, editors, *Research and Advanced Technology for Digital Libraries: 8th European Conference, ECDL*, pages 14–25. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[57] Marcia Lucas Pesce. RdXel: A toolkit for RDF statistical data manipulation through spreadsheets. Master thesis, 2012.

[58] Lívia Ruback, Marcia Pesce, Sofia Manso, Sérgio Ortiga, Percy E. Rivera Salas, and Marco A. Casanova. A mediator for statistical linked data. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 339–341. ACM, 2013.

[59] Surajit Chaudhuri and Umeshwar Dayal. An overview of data warehousing and OLAP technology. *ACM Sigmod record*, 26(1):65–74, 1997.

[60] Souripriya Das, Seema Sundara, and Richard Cyganiak. R2RML, 2012.

[61] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A Nucleus for a Web of Open Data. In Karl Aberer, Key-Sun Choi, Natasha Noy, Dean Allemang, Kyung-Il Lee, Lyndon Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *The Semantic Web: ISWC*, pages 722–735. Springer Berlin Heidelberg, 2007.

[62] Max Schmachtenberg, Christian Bizer, Anja Jentzsch, and Richard Cyganiak. Linking Open Data cloud diagram. `http://lod-cloud.net/`, 2014.

[63] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. DBpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8. ACM, 2011.

[64] Peter Christoffersen and Peter Doyle. From Inflation to Growth. *Economics of Transition*, 8(2):421–451, July 2000.

[65] Mohsin S. Khan and Abdelhak S. Ssnhadji. Threshold Effects in the Relationship between Inflation and Growth. *IMF Staff Papers*, 48(1):1–21, 2001.

[66] Le Thanh Tung and Pham Tien Thanh. Threshold in the Relationship between Inflation and Economic Growth: Empirical Evidence in Vietnam. *Asian Social Science*, 11(10), April 2015.

[67] Cinzia Cappiello, Chiara Francalanci, and Barbara Pernici. Data Quality Assessment from the User's Perspective. In *Proceedings of the International Workshop on Information Quality in Information Systems*, pages 68–73. ACM, 2004.

[68] Ken Orr. Data Quality and Systems Theory. *Commun. ACM*, 41(2):66–71, 1998.

[69] Richard Y. Wang. A Product Perspective on Total Data Quality Management. *Commun. ACM*, 41(2):58–65, 1998.

[70] Arkady Maydanchik. *Data Quality Assessment*. Technics Publications, 2007.

[71] Jiawei Han, Jian Pei, and Micheline Kamber. *Data Mining: Concepts and Techniques*. Elsevier, 2011.

[72] Md. Samiullah, Chowdhury Farhan Ahmed, Manziba Akanda Nishi, Anna Fariha, S. M. Abdullah, and Md. Rafiqul Islam. Correlation Mining in Graph Databases with a New Measure. In Yoshiharu Ishikawa, Jianzhong Li, Wei Wang, Rui Zhang, and Wenjie Zhang, editors, *Web Technologies and Applications: 15th Asia-Pacific Web Conference, APWeb*, pages 88–95. Springer Berlin Heidelberg, 2013.

[73] Ryan Shaun Baker and Paul Salvador Inventado. Educational Data Mining and Learning Analytics. In Johann Ari Larusson and Brandon White, editors, *Learning Analytics*, pages 61–75. Springer New York, 2014.

[74] Heiko Paulheim. Generating Possible Interpretations for Statistics from Linked Open Data. In Elena Simperl, Philipp Cimiano, Axel Polleres, Oscar Corcho, and Valentina Presutti, editors, *The Semantic Web: ESWC*, pages 560–574. Springer Berlin Heidelberg, 2012.

[75] Petar Ristoski and Heiko Paulheim. Analyzing Statistics with Background Knowledge from Linked Open Data.pdf. In *Proceedings of the 1st International Workshop on Semantic Statistics*. CEUR-WS.org, 2013.

[76] A. W. Phillips. The Relation between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861-1957. *Economica*, 25(100):283–299, 1958.

[77] Tuan-Dat Trinh, Ba-Lam Do, Peter Wetz, Amin Anjomshoaa, and A Min Tjoa. Linked Widgets: An Approach to Exploit Open Government Data. In *Proceedings of International Conference on Information Integration and Web-Based Applications & Services*, pages 438–442. ACM, 2013.

[78] Tuan-Dat Trinh, Peter Wetz, Ba-Lam Do, Amin Anjomshoaa, Elmar Kiesling, and A. Min Tjoa. Open Linked Widgets Mashup Platform. In *Proceedings of the AI Mashup Challenge*. CEUR-WS.org, 2014.

[79] Tuan-Dat Trinh, Peter Wetz, Ba-Lam Do, Amin Anjomshoaa, Elmar Kiesling, and A. Min Tjoa. A Web-based Platform for Dynamic Integration of Heterogeneous Data. In *Proceedings of the International Conference on Information Integration and Web-Based Applications & Services*, pages 253–261. ACM, 2014.

[80] Axel-Cyrille Ngonga Ngomo and Sören Auer. LIMES: A Time-efficient Approach for Large-scale Link Discovery on the Web of Data. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, pages 2312–2317. AAAI Press, 2011.

[81] Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. Silk-A Link Discovery Framework for the Web of Data. In *Proceedings of Workshop on Linked Data on the Web*. CEUR-WS.org, 2009.

[82] Afraz Jaffri, Hugh Glaser, and Ian Millard. Managing URI synonymity to enable consistent reference on the Semantic Web. In *Proceedings of International Workshop on Identity and Reference on the Semantic Web*. CEUR-WS.org, 2008.

[83] Christian YA Brenninkmeijer, Ian Dunlop, Carole A. Goble, Alasdair JG Gray, Steve Pettifer, and Robert Stevens. Computing Identity Co-Reference Across Drug Discovery Datasets. In *SWAT4LS*, 2013.

[84] Andreas Langegger and Wolfram Wöß. XLWrap – Querying and Integrating Arbitrary Spreadsheets with SPARQL. In Abraham Bernstein, David R. Karger, Tom Heath, Lee Feigenbaum, Diana Maynard, Enrico Motta, and Krishnaprasad Thirunarayan, editors, *The Semantic Web - ISWC*, pages 359–374. Springer Berlin Heidelberg, 2009.

[85] Martin J. O'Connor, Christian Halaschek-Wiener, and Mark A. Musen. Mapping Master: A Flexible Approach for Mapping Spreadsheets to OWL. 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*, pages 194–208. Springer Berlin Heidelberg, 2010.

[86] Matthew Horridge and Peter F. Patel-Schneider. OWL 2 Web Ontology Language, Manchester Syntax. `https://www.w3.org/TR/owl2-manchester-syntax/`, 2012.

[87] Christian Bizer. D2r map-a database to rdf mapping language. In *WWW - Poster Track*. CEUR-WS.org, 2003.

[88] Saleh Ghasemi, Wo-Shun Luk, and Norah Alrayes. M2RML: Multidimensional to RDF Mapping Language. In *Proceedings of the 25th International Workshop on Database and Expert Systems Applications*, pages 263–267, 2014.

[89] Saleh Ghasemi. M2RML: Mapping Multidimensional Data to RDF, Master thesis, 2014.

[90] Anastasia Dimou, Miel Vander Sande, Pieter Colpaert, Erik Mannens, and Rik Van de Walle. Extending R2RML to a source-independent mapping language for RDF. In *Proceedings of ISWC - Posters & Demonstrations Track*, pages 237–240. CEUR-WS.org, 2013.

[91] Jonathan H. Clark and José P. González-Brenes. Coreference resolution: Current trends and future directions. *Language and Statistics II Literature Review*, pages 1–14, 2008.

[92] Stanford Natuaral Language Processing Group. Coreference Resolution. `http://nlp.stanford.edu/projects/coref.shtml`, 2016.

[93] Andriy Nikolov, Victoria Uren, and Enrico Motta. KnoFuss: A comprehensive architecture for knowledge fusion. In *Proceedings of the 4th International Conference on Knowledge Capture*, pages 185–186. ACM, 2007.

[94] Andriy Nikolov, Victoria Uren, Enrico Motta, and Anne de Roeck. Integration of Semantically Annotated Data by the KnoFuss Architecture. In Aldo Gangemi and Jérôme Euzenat, editors, *Knowledge Engineering: Practice and Patterns, EKAW*, pages 265–274. Springer Berlin Heidelberg, 2008.

[95] Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. Discovering and maintaining links on the web of data. In *ISWC*, pages 650–665. Springer, 2009.

[96] William Cohen, Pradeep Ravikumar, and Stephen Fienberg. A comparison of string metrics for matching names and records. In *KDD Workshop on Data Cleaning and Object Consolidation*, volume 3, pages 73–78, 2003.

[97] Esko Ukkonen. Approximate string-matching with q-grams and maximal matches. *Theoretical computer science*, 92(1):191–211, 1992.

[98] Mohamed A. Khamsi and William A. Kirk. *An Introduction to Metric Spaces and Fixed Point Theory*. John Wiley & Sons, March 2001.

[99] Markus Nentwig, Michael Hartung, Axel-Cyrille Ngonga Ngomo, and Erhard Rahm. A survey of current Link Discovery frameworks. *Semantic Web*, pages 1–18, 2015.

[100] Afraz Jaffri, Hugh Glaser, and Ian Millard. URI Identity Management for Semantic Web Data Integration and Linkage. In Robert Meersman, Zahir Tari, and Pilar Herrero, editors, *On the Move to Meaningful Internet Systems: OTM Workshops*, pages 1125–1134. Springer Berlin Heidelberg, 2007.

[101] David Booth. URIs and the myth of resource identity. In *Proceedings of Identity, Reference, and the Web Workshop at the WWW Conference*, 2006.

[102] Hugh Glaser, Ian Millard, Afraz Jaffri, Tim Lewy, and Ben Dowling. On Coreference and The Semantic Web. In *ESWC*. Springer, 2008.

[103] Hugh Glaser, Ian Millard, T. Anderson, and B. Randell. ReSIST: Resilience for Survivability in IST, 2007.

[104] Sarven Capadisli, Sören Auer, and Reinhard Riedl. Towards Linked Statistical Data Analysis. In *Proceedings of the 1st International Workshop on Semantic Statistics*. CEUR-WS.org, 2013.

[105] Marta Sabou, Adrian M. P. Braşoveanu, and Irem Arsal. Supporting Tourism Decision Making with Linked Data. In *Proceedings of the 8th International Conference on Semantic Systems*, pages 201–204. ACM, 2012.

[106] Marta Sabou, Adrian M. P. Braşoveanu, and Irem Önder. Linked Data for Cross-Domain Decision-Making in Tourism. In Iis Tussyadiah and Alessandro Inversini, editors, *Proceedings of the International Conference on Information and Communication Technologies in Tourism*, pages 197–210. Springer International Publishing, 2015.

[107] Karl W Wöber. Information supply in tourism management by marketing decision support systems. *Tourism Management*, 24(3):241–255, 2003.

[108] Benedikt Kämpgen. *Flexible Integration and Effcient Analysis of Multidimensional Datasets from the Web*. PhD thesis, Karlsruhe Institute of Technology, 2015.

[109] Benedikt Kämpgen, Seán O'Riain, and Andreas Harth. Interacting with Statistical Linked Data via OLAP Operations. In Elena Simperl, Barry Norton, Dunja Mladenic, Emanuele Della Valle, Irini Fundulaki, Alexandre Passant, and Raphaël Troncy, editors, *The Semantic Web: ESWC 2012 Satellite Events*, pages 87–101. Springer Berlin Heidelberg, 2012.

[110] Jesús Pardillo, Jose-Norberto Mazón, and Juan Trujillo. Bridging the Semantic Gap in OLAP Models: Platform-independent Queries. In *Proceedings of the 11th International Workshop on Data Warehousing and OLAP*, pages 89–96. ACM, 2008.

[111] Sarven Capadisli, Albert Meroño-Peñuela, Sören Auer, and Reinhard Riedl. Semantic similarity and correlation of linked statistical data analysis. In *Proceedings of the 2nd International Workshop on Semantic Statistics*. CEUR-WS.org, 2014.

[112] Sebastian Bayerl and Michael Granitzer. Bacon: Linked Data Integration based on the RDF Data Cube Vocabulary. In *Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics*, pages 1–6. ACM, 2015.

[113] Marios Meimaris and George Papastefanatos. Containment and Complementarity Relationships in Multidimensional Linked Open Data. In *Proceedings of the 2nd International Workshop on Semantic Statistics*. CEUR-WS.org, 2014.

[114] Marios Meimaris, George Papastefanatos, Panos Vassiliadis, and Ioannis Anagnostopoulos. Efficient Computation of Containment and Complementarity in RDF Data Cubes. In *Proceedings of the 19th International Conference on Extending Database Technology (EDBT)*. OpenProceedings, 2016.

[115] Ivan Ermilov, Michael Martin, Jens Lehmann, and Sören Auer. Linked Open Data Statistics: Collection and Exploitation. In Pavel Klinov and Dmitry Mouromtsev, editors, *Knowledge Engineering and the Semantic Web: Proceedings of the 4th International Conference on Knowledge Engineering and Semantic Web, KESW*, volume 394, pages 242–249. Springer Berlin Heidelberg, 2013.

[116] Michael Martin, Konrad Abicht, Claus Stadler, Axel-Cyrille Ngonga Ngomo, Tommaso Soru, and Sören Auer. CubeViz: Exploration and Visualization of Statistical Linked Data. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 219–222. ACM, 2015.

[117] Norman Heino, Sebastian Dietzold, Michael Martin, and Sören Auer. Developing Semantic Web Applications with the OntoWiki Framework. In Tassilo Pellegrini, Sóren Auer, Klaus Tochtermann, and Sebastian Schaffert, editors, *Networked Knowledge - Networked Media*, pages 61–77. Springer Berlin Heidelberg, 2009.

[118] Patrick Hoefler, Michael Granitzer, Vedran Sabol, and Stefanie Lindstaedt. Linked Data Query Wizard: A Tabular Interface for the Semantic Web. In Philipp Cimiano, Miriam Fernández, Vanessa Lopez, Stefan Schlobach, and Johanna Völker, editors, *The Semantic Web: ESWC 2013 Satellite Events*, pages 173–177. Springer Berlin Heidelberg, 2013.

[119] Vedran Sabol, Gerwald Tschinkel, Eduardo Veas, Patrick Hoefler, Belgin Mutlu, and Michael Granitzer. Discovery and Visual Analysis of Linked Data for Humans. In Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandečić, Paul Groth, Natasha Noy, Krzysztof Janowicz, and Carole Goble, editors, *The Semantic Web – ISWC 2014, Part I*, volume 8796, pages 309–324. Springer International Publishing, 2014.

[120] Belgin Mutlu, Patrick Hoefler, Vedran Sabol, Gerwald Tschinkel, and Michael Granitzer. Automated Visualization Support for Linked Research Data. In *Proceedings of the I-SEMANTICS 2013 Posters & Demonstrations Track*, pages 40–44. CEUR-WS.org, 2013.

[121] Gerwald Tschinkel, Eduardo Veas, Belgin Mutlu, and Vedran Sabol. Using semantics for interactive visual analysis of linked open data. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track*, pages 133–136. CEUR-WS.org, 2014.

[122] Evangelos Kalampokis, Areti Karamanou, Andriy Nikolov, Peter Haase, Richard Cyganiak, Bill Roberts, Paul Hermans, Efthimios Tambouris, and Konstantinos

Tarabanis. Creating and utilizing linked open statistical data for the development of advanced analytics services. In *Proceedings of the 2nd International Workshop on Semantic Statistics*. CEUR-WS.org, 2014.

[123] Jakub Klímek, Jiří Helmich, and Martin Nečaský. Payola: Collaborative Linked Data Analysis and Visualization Framework. In Philipp Cimiano, Miriam Fernández, Vanessa Lopez, Stefan Schlobach, and Johanna Völker, editors, *The Semantic Web: ESWC 2013 Satellite Events*, pages 147–151. Springer Berlin Heidelberg, 2013.

[124] Josep Maria Brunetti, Sören Auer, Roberto García, Jakub Klímek, and Martin Nečaský. Formal Linked Data Visualization Model. In *Proceedings of International Conference on Information Integration and Web-Based Applications & Services*, pages 309–318. ACM, 2013.

[125] Albert Meroño-Peñuela. LSD Dimensions: Use and Reuse of Linked Statistical Data. In Patrick Lambrix, Eero Hyvönen, Eva Blomqvist, Valentina Presutti, Guilin Qi, Uli Sattler, Ying Ding, and Chiara Ghidini, editors, *Knowledge Engineering and Knowledge Management: EKAW 2014 Satellite Events*, pages 159–163. Springer International Publishing, 2015.