Comparative Analysis of Open Linked Fiscal Data

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

von
Fathoni Arief Musyaffa
aus
Kediri, Indonesien

Bonn, 08.09.2020

Dieser Forschungsbericht wurde als Dissertation von der
Mathematisch-Naturwissenschaftlichen Fakultät der Universität Bonn angenommen und ist auf dem Publikationsserver der ULB Bonn
https://nbn-resolving.org/urn:nbn:de:hbz:5-62298 elektronisch publiziert.

1. Gutachter:	Prof. Dr. Sören Auer
2. Gutachter:	Prof. Dr. Jens Lehmann
Tag der Promotion:	11.03.2021
Erscheinungsjahr:	2021

Abstract

The open data movement within public administrations has provided data regarding governance publicly. As public administrators and governments produce data and release the data as open data, the volume of the data is highly increasing. One of these datasets is budget and spending data, which has been gaining interest to the extent that several working groups and CSO/NGOs started working on this particular open data domain. The majority of these datasets are part of the open budget and spending datasets, which laid out data regarding how public administrations plan, revise, allocate, and expense their governance funding. The disclosure of public administration budget and spending data is expected to improve governance transparency, accountability, law enforcement, and political participation.

Unfortunately, the analysis of budget and spending datasets is not a trivial task to do for several reasons. *First*, the quality of open fiscal data varies. Standards and recommendations for publishing open data are available, however, these standards are often not met and no framework specifically addresses fiscal data quality measurements. *Second*, the datasets are heterogeneous, since it is produced by different public administrations with different business process, accounting practice, requirements, and language. This lead to a challenging task in data integration across public budget and spending data. The structural and linguistic heterogeneity of open budget and spending data makes comparative analysis across datasets difficult to perform. *Third*, datasets within the budget and spending domain are complicated. To be able to comprehend such data, expertise is needed both from the public accounting/budgeting domain, as well as the technical domain to digest the datasets properly. *Fourth*, a platform to transform, store, analyze, and visualize datasets is necessary, especially those that make the utilization of semantic analysis is possible. *Fifth*, there is no conceptual association between datasets, which can be used as a comparison point to analyze fiscal records between compared public administrations. *Lastly*, there is a lack of methodology to consume and compare linked open fiscal data records across different public administrations.

Our focus in this thesis is hence to perform research to help the community gain a better understanding of open fiscal data, provide analysis of their quality, suggest a way to publish open fiscal data in an improved manner, analyze the open fiscal data heterogeneity while also laying out lessons learned regarding their current state and supporting data formats that are capable for open fiscal data integration. Consequently, a platform to digest, analyze and visualize these datasets is devised, continued with performing experiments on multilingual fiscal data concept mapping and wrapped up with a proof-of-concept description of comparative analysis over linked open fiscal data.

Acknowledgements

My path had been predominantly shaped by very different socio-cultural values and norms. As I moved to Germany for doing my Ph.D., I began to both see different societal perspectives and faced challenges that I would have never encountered had I decide not to pursue a Ph.D. in Bonn. While pursuing a Ph.D. can be mentally tasking, the process of encountering new perspectives and facing its challenges has stretched one's growth as a person. For that, I am feeling blessed and grateful. I am thankful to Prof. Sören Auer for giving me the chance to join the Enterprise Information Systems (EIS) as well as providing me with guidance on starting research and persuading me to write academic papers. I am also grateful to Prof. Jens Lehmann in Smart Data Analytics (SDA) for providing a nice environment and facilitating my research activities further.

Diving into research requires some effort and patience in mentorship. I am lucky and grateful to have been mentored by Dr. Fabrizio Orlandi at EIS for his friendly, nurturing advice and supervision. From him, I learned the logical process behind research and how it can be presented. When I had doubts about my initial start, he convinces me that I will be able to eventually make it to the end of this Ph.D. journey. I am also thankful to Dr. Hajira Jabeen for her analytical, caring mentorship and kindness in SDA. She keeps assuring me that I am bigger than I think I am. As I progress with research, I was also helped by numerous students and researchers at EIS and SDA to work with my scientific papers. For that, I would like to thank Christiane, Lavdim, Yakun, and Prof. Maria-Esther Vidal. It has been a delight to work within EIS and SDA for its diverse and friendly working environments, whose members I cannot mention here one by one. My appreciation goes to many people in this group and I would like to mention especially Kemele, Mohamed, and Elisa. I am also grateful for the generous material and non-material support provided by the DAAD during my doctoral study.

I would like to also express my gratitude to Tonggie, Zakia, and Carlos for all the supports. Living in a new city can be tough, but having friends like them makes Bonn a far homier city than I thought it ever would. I am also grateful for Cipta as a comrade who shares this journey from the very beginning and I am grateful for the friendships and inputs. Looking forward to seeing where life brings each of us once this adventure ends and a new chapter begins.

To my family, I would like to assert my gratitude for not only trusting me but also encouraging me to pursue education from the very beginning even though it most often means that I will be far away, again, for a long time. The calls have been another source of joy and encouragement. To my brothers and sisters, it has been a delight to see what you've built at home and hopefully, I'll soon be part of it even from afar.

This thesis is dedicated to my late father.

Contents

Part I

Preliminaries

CHAPTER 1

Introduction

Public governments and international bodies have increasingly published open government data. Open Government Partnership (OGP)[1] establishes an open data working group to develop open data plans and actions across its member countries. By January 2020, there are 78 member countries of OGP. Open Knowledge Foundation (OKF) indexed 122 countries in the Global Open Data Index (GODI) [1], indicating that these countries have published open data in various domains. Beyond the country level, publication of open data is also done on supra-national levels, such as United Nations,[2] European Union,[3] and World Bank.[4]

One of the high-value domains in open data is public finance [2] or fiscal data. It is also the most frequent type of open data being released [3]. The term *fiscal* refers to any activities related to government expenditures, revenues, and debt [4]. Open fiscal datasets include, but are not limited to, budget, spending, contract, and procurement data. Budget datasets determine the various income and expenditure allocations within a certain period. Spending datasets provide details regarding the amount of money paid for specific items.

Open fiscal data, especially budget data, is mentioned to be one of the most important [5] and most published data [6]. Budget data has ranked in the top three domains of open data published in the years 2013-2015. OpenSpending[5] states that they host more than 3,200 fiscal (budget and spending) datasets openly as of July 2019, which is comprised of more than 132 million fiscal records.

The open publishing of such datasets has a number of motivating benefits, such as: increasing transparency and compliance [7, 8], preventing corruption [9], raising democratic control and participation in politics [8, 10], encouraging innovation in services and products [7, 10, 11], enabling comparative analysis [8], enhancing law enforcement [10], adding business value [8], improving efficiency and effectiveness [8], as well as generally

[1] http://www.opengovpartnership.org/
[2] http://data.un.org//
[3] https://data.europa.eu/euodp/en/data
[4] http://data.worldbank.org/
[5] https://openspending.org/

reducing the barrier between government and citizens [8, 10].

Publishing public fiscal data improves public administration transparency and accountability. Governance *transparency* is concerned with the capability of finding information about what happened in the government [12]. *Accountability* exists when tasks done by a particular individual or an organization can be requested, overseen, and directed by others [13]. *Transparent public administration* improves public trust which engages more political participation from their citizens. Open data implementation in Brazil has successfully uncovered corruption scandals, as reported by [9]. A summary of open data values and impacts (projected market value, number of open data jobs created, economic benefit, etc.) is provided by [14].

The growing government budget and spending data make it possible to perform cross-administration budget analysis. An analysis can be done on different fiscal datasets that have similar properties. For example, comparing budget allocations from different municipalities with a similar population, area size, and/or GDP. This analysis requires the datasets to be consistent and to contain common classifications, which make the datasets comparable. When such detailed fiscal practices can be publicly scrutinized, the chance of public officials conducting fiscal malpractice is lower, as illustrated in the report by [9]. Open fiscal data allows for measuring the effectiveness and efficiency of an executed funding program, which can be assessed by the outcome of the funding.

The benefits of having an open fiscal data published do not come without challenges. Many open datasets have quality issues; an analysis by Computer Weekly on UK's Cabinet Office open spending data showed that the released data have inconsistent computer encoding and therefore an advanced programming skill is required to scrutinize the data systematically [15]. In addition, open datasets generally have not been designed to be interoperable [16], for example, datasets are provided in various formats, structure, classification schemes, and languages that prevent the datasets from being effectively integrated and analyzed [16]. Due to the decentralized nature of the data publication and creation, the published budget and spending data are often disparate; they are published in different structures, formats, languages, metrics (e.g., feet/meters), granularity (e.g., years/months), and possess different forms of heterogeneity [17]. Moreover, the data are normally found to be incomplete and of low-quality [5]. Furthermore, additional information about domain-specific concepts also accompanies most of the released open fiscal data in the form of non-standardized classification terms.

1.1 Motivation

The challenges of ingesting open budget and spending data are illustrated in the mind map in Figure 1.1. These challenges come from the nature of independently published open budget and spending data, which is often messy, heterogeneous, hard to integrate, hard to link, and hard to analyze. Each of these challenging characteristics is a consequence of different factors, for example, heterogeneity is caused by the absence of a unified open budget and spending data publication standard regarding which data model, format, and structure that should be followed by public administrators. This is illustrated in Figure 1.2,

in which Municipality A and Municipality B publish each of their datasets differently. Municipality A separates between income and expenditure budget dataset. They also provide a detailed description of their classification codes within the dataset itself. On the other hand, Municipality B does not separate between income and expenditure budget dataset. Instead, they use positive and negative values in their budget lines to indicate income and expenditure. Also, Municipality B does not describe the classifications in their datasets. In addition, within the dataset published by Municipality B, different classification codes are appended to form one compound classification code, thus, making it more complicated to analyze.

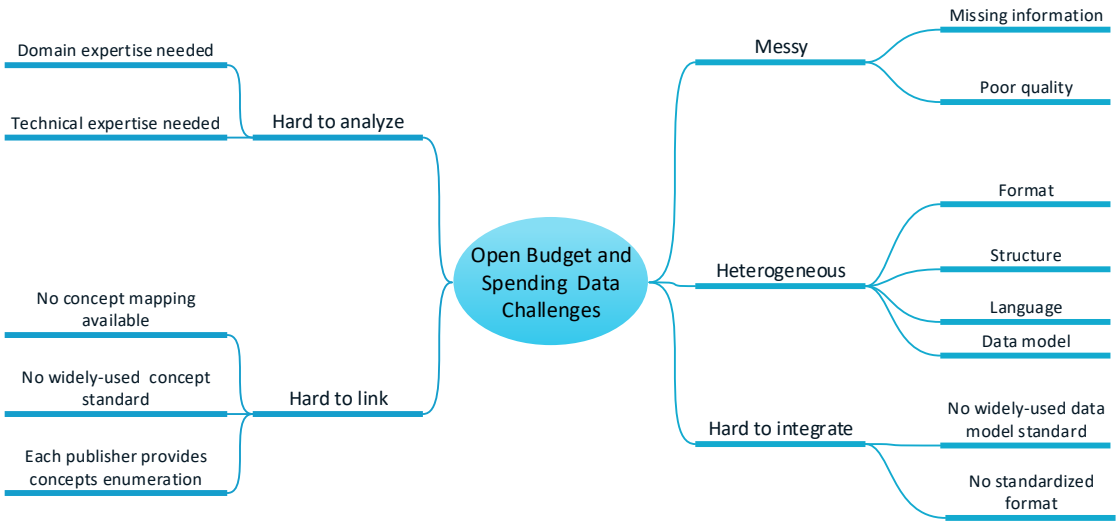

Figure 1.1: Open budget and spending data: motivating challenges.

The overall goals and objectives of this thesis is depicted in Figure 1.3. *Public administrators* typically publish their datasets in a decentralized manner. There are no enforced standards on how the budget and spending datasets are published. Hence, these datasets are not consistent among each other. Problems arise when the datasets need to be integrated for further processing and analysis. The integration and ingestion of these datasets into a unified platform is important for further analytical tools and visualization. Datasets that can not be digested into any integration platform hinders data comprehension by *consumers* (citizen, journalists, and other stakeholders). In the end, data that cannot be comprehended by its consumers defy the purpose of transparent and accountable administrations that motivates sharing open budget and spending data in the first place.

The work laid out in this thesis is intended to tackle the challenges surrounding the open budget and spending data publishing and analytics. Since the main component of open fiscal data published by most public administration is budget and spending data, this work will refer to open budget and spending data simply as *open fiscal data*.

In general, this thesis is comprised of three main layers. From the *datasets analysis* layer, an *assessment* needs to be done on the quality of the current state of datasets. Prior

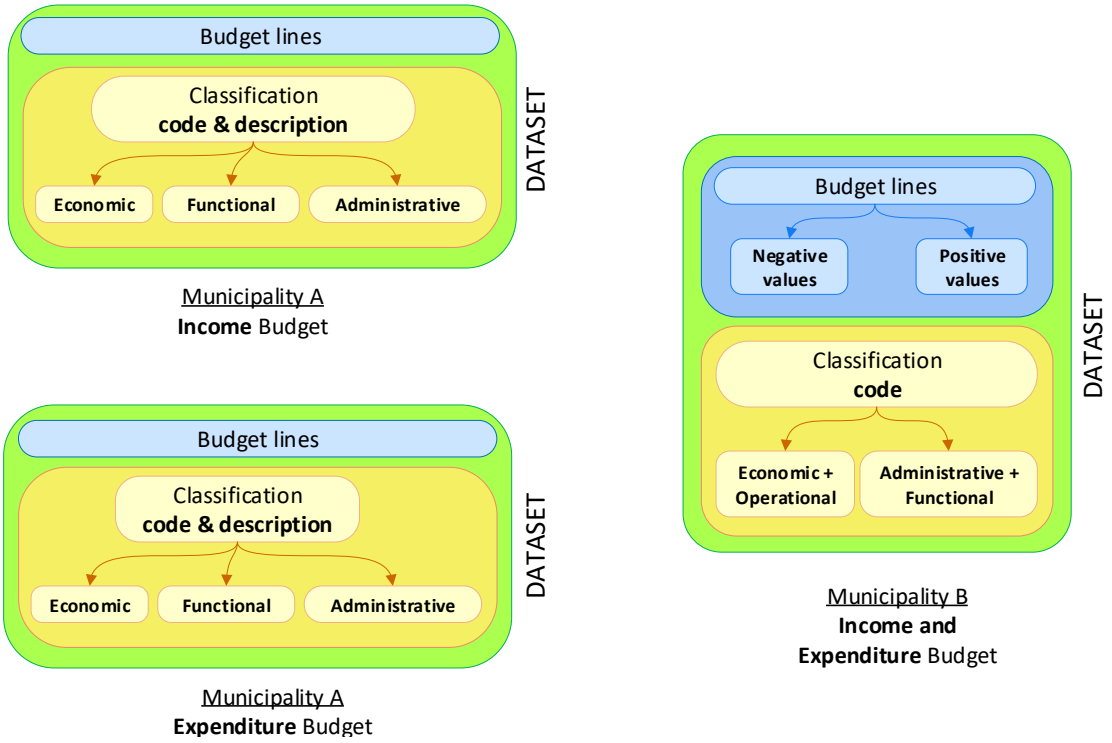

Figure 1.2: Motivating example: structural differences across datasets published by two different municipalities.

to integrating the data, we need to check the challenges introduced by the *heterogeneity* of these datasets, and recommends what type of data model is suitable for integrating these datasets. Additionally, a *standard* is necessary to publish a good quality open fiscal data in a more *regulated* manner.

On *datasets processing* layer, works need to be done include 1) performing *semantic enrichment* on the data, 2) performing datasets *integration*, and 3) building a *semantic platform for public fiscal data*. Enriching datasets with semantics allows machines to understand more detailed information regarding the data being published. For example, once the data is annotated with a specific city name, the whole properties of the information available on that city could be queried from online knowledge bases, allowing data mash-up with the latest and up-to-date data in the open knowledge bases. Once the data has been enriched, datasets are integrated in a uniform manner with other datasets. This could be made easier when a platform is made available for those specific datasets with recommended data models, allowing effective data enrichment, storage, retrieval, and analytics.

The next layer, *datasets linking and analytics*, involves creating *links* on similar concepts that are represented in different languages. The link is necessary to perform *comparative analysis* on different datasets containing similar concepts. This is because, with the increasing number of datasets across public administration, the chance of similar concepts

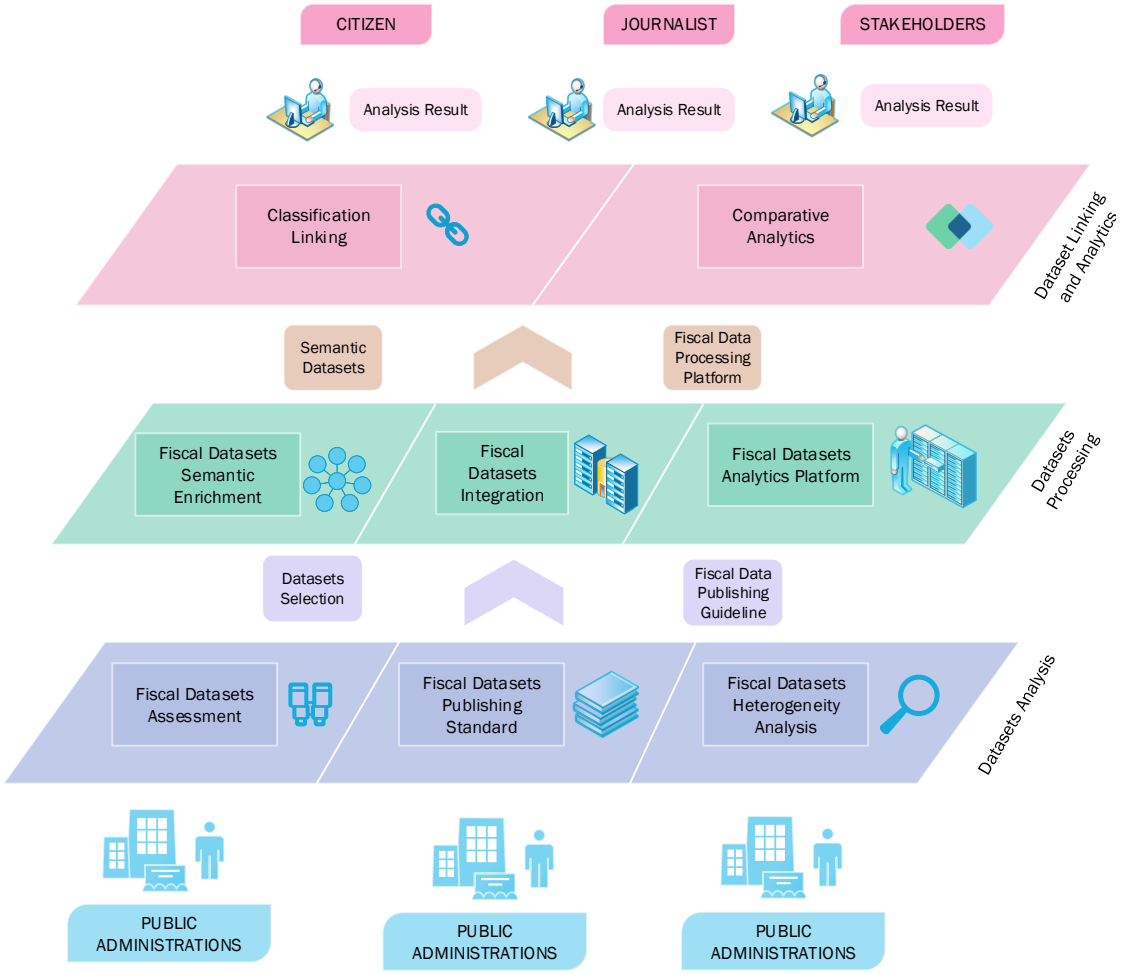

Figure 1.3: Goals and objectives for the work laid out in this thesis. There are three main layers tackled, mainly public fiscal data analysis, processing, and linking.

from the published datasets being compared is higher. In addition, semantic knowledge bases consist of open and free information regarding the city/state/country a dataset comes from, enabling each information item (e.g., GDP, population, HDI, etc) to be used as a pivot to compare the budget items across different datasets. At last, a pipeline to enable comparative analysis is provided. This process requires the integration of most works from different stages laid out in this thesis. It makes use of the datasets transformed into compatible, domain-specific, public fiscal ontology, as well as external open knowledge bases to obtain additional information regarding the characteristics of each city/state/country to be used as a comparison point.

1.2 Problem Definition and Challenges

The gap in the domain-specific background should be minimized for open data fiscal data analytics on various stages, by contributing to solving underlying challenges of open fiscal data processing. These challenges are laid out in Figure 1.4.

Challenge 1: Quality issues of Open Fiscal Data (OFD).

Despite the increasing volume of open fiscal data and the availability of recommendations and standards available for general-domain open data, the quality of published fiscal data across different public administrations has not yet been assessed. Additionally, there are no specific standards especially designed for publishing high-quality open fiscal data.

Challenge 2: Complexity of fiscal data.

Open fiscal data requires public administration, finance, fiscal, and accounting background to analyze. Additionally, technical background and programming skills are also often needed. To analyze these data, most people need a substantial resource and time to understand the required domains. These requirements are often not met by the common citizen who is interested to analyze these data.

Challenge 3: Structural and linguistic heterogeneity of OFD.

Datasets are published by different public administrations from diverse geographical locations. There is no consensus on the structure of the datasets, and the language used by each public administrations are likely to differ. These conditions impose the challenge of structural and linguistic heterogeneity.

Challenge 4: Lack of conceptual association between datasets.

The availability of association or links across datasets enables further analysis, for example comparing budget and spending items between different cities. This could in theory be facilitated by the use of available classifications or enumerated terms provided by supranational organizations (e.g., United Nations, European Union, or IMF). However,

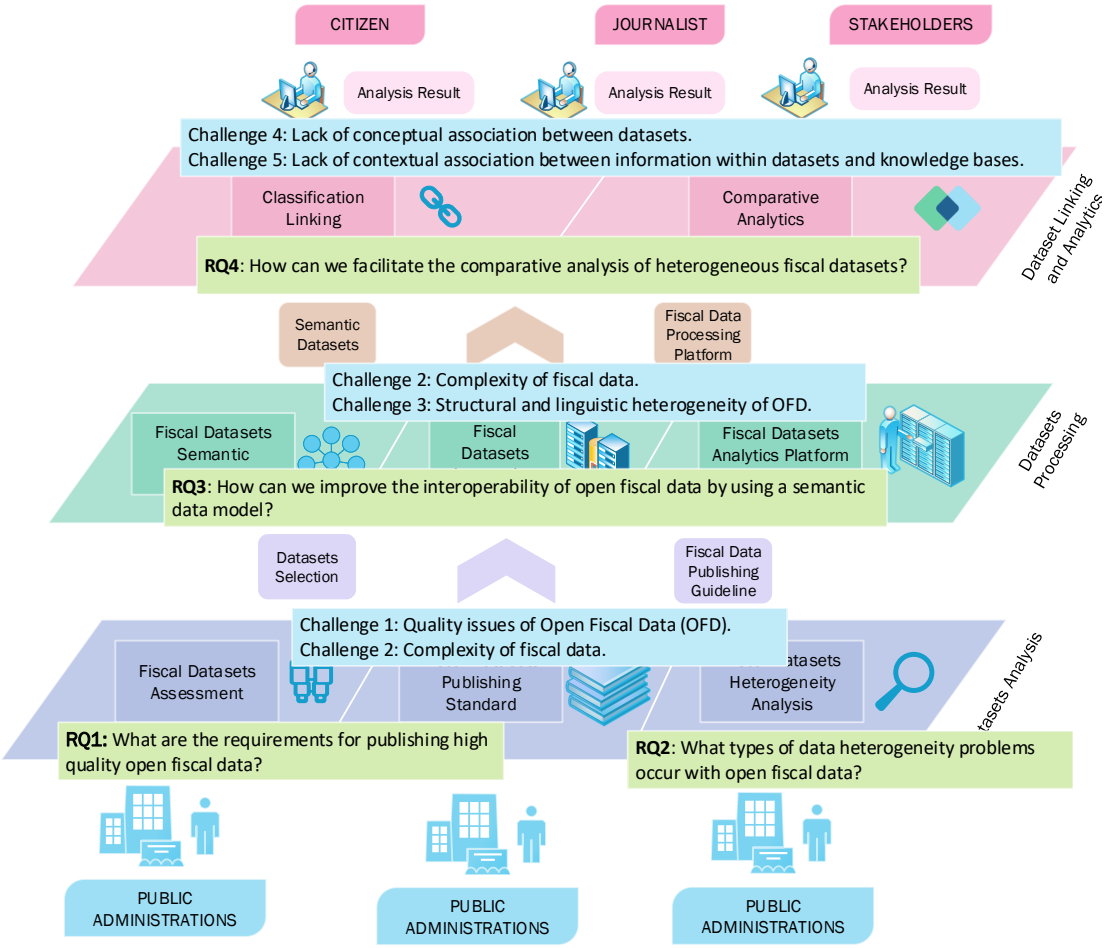

Figure 1.4: Challenges and research questions covered in this thesis.

using these terms requires adapting available terms originating from dataset publishers with term published by the supranational organizations, leading to a necessity in additional overhead. Pragmatically, the public administrations do not have a strong interest to address this overhead as long as there is neither regulated obligation to conform with supranationally standardized terms nor compelling use cases proving that publishing datasets using those standardized terms would create a greater value.

Challenge 5: Lack of contextual association between information within datasets and knowledge bases.

Open knowledge bases keep expanding. The information available in these knowledge bases has the potential to augment the analysis of published datasets, including open fiscal data. However, associating the datasets with information in the knowledge bases also requires an overhead that can not be carried out by most public administrations.

1.3 Research Questions

The following section describes the research questions discussed in this thesis. There are four main research questions, which are strongly tied to the motivation as illustrated in Figure 1.4.

RQ1: What are the requirements for publishing high quality open fiscal data?

In order to answer this research question, we analyze more than 70 open budget and spending datasets across different public administration levels, ranging from cities, states, countries, as well as supranational entities. To ensure quality factors for these data, we reuse quality factors from different standards, recommendations, and guidelines from different civic communities. We also provide additional quality factors specifically needed on open fiscal data. We prioritize the factors by conducting a questionnaire participated by people with different backgrounds. At last, we finally provide the result of this assessment, along with recommended quality factors that need to be considered upon publishing open fiscal data.

RQ2: What types of data heterogeneity problems occur with open fiscal data?

We continue by analyzing the result of RQ1 and see the heterogeneities we found on the dataset and then categorize, and classify these heterogeneities accordingly. Additionally, we provide a comparison of two state-of-the-art data models that are designed to represent these fiscal data, and then assert how compatible these data models with the heterogeneities that we find. A recommendation is provided for data model developer/maintainer and for the data publishers.

RQ3: How can we improve the interoperability of open fiscal data by using a semantic data model?

We use the semantic data model we have analyzed in RQ2, and transform raw datasets from its original format (e.g., XLSX, CSV) to semantic format. In this part, there are several processes involved, such as designing a platform to annotate, enrich, store, and analyze the fiscal datasets.

RQ4: How can we facilitate the comparative analysis of heterogeneous fiscal datasets?

There are no mappings between published open fiscal datasets. These mappings are required to enable comparative analysis across different public fiscal data. To answer this research question, we perform an experiment to link datasets from different languages and administrations. We use a combination of machine translation and various string similarity measures to detect possible similar concepts across datasets. Since the comparison of concepts within these datasets requires a large number of comparison operations and some of the string similarity measures are computationally complex, we use distributed

computing to optimize our experiment. Afterward, we build a prototype for a comparative analysis of open public fiscal data by utilizing additional information on the knowledge graph.

1.4 Thesis Overview

To provide the readers with a general overview of this thesis, we present our contributions, our scientific publications, and the structure of this thesis in this section.

1.4.1 Contributions

Figure 1.4 mentions the challenges and the research questions covered in this thesis, which implies the contribution overview of this work, consisting of surveys, analysis, platform, and tools that laid the ground for supporting open fiscal data analytics both from a methodological and technical perspective. The contributions of this work are summarized in the following points:

1. *A comprehensive analysis of open fiscal data followed by standards and recommendations for data publishers.*

   Contribution for RQ1. As an initial ground for the research in the open fiscal data domain, we perform a survey regarding the state of open fiscal data. This is done by studying various resources regarding how open data from the general domain is published. Subsequently, we perform a careful analysis of various open fiscal datasets published by different public administration on various administration levels (cities, states, countries, and supranational organizations). Afterward, we provide open data publishing recommendations with additional quality factors that also consider the public fiscal domain. As a result, we: 1) propose a comprehensive assessment framework for open fiscal data, 2) provide an assessment of the current state open fiscal data, 3) present a number of quality issues that were found, and 4) provide guidelines for publishing open fiscal data based on the assessment.

2. *Classification of open fiscal data heterogeneity and its conformity to current state-of-the-art open fiscal data model*

   Contribution for RQ2. Following the work for the initial contribution, we found that there are some patterns of heterogeneity in open fiscal data. We present and provide a hierarchical view of these heterogeneities. Additionally, there are data models that specifically created to represent a unified data model for open fiscal data. We analyze the compatibility of these data models against found heterogeneities. We also present lessons learned that could be aimed at datasets publishers and technical/scientific communities that involved in public open fiscal data.

3. *An integrated platform for semantic open fiscal data analytics.*

   Contribution for RQ3. Publishing fiscal datasets with semantic annotation provide the possibility of enhanced analytics. Platform and architecture for performing analysis of semantic open fiscal data are not yet available. For this reason, we provide an open data architecture, based on requirements that have been specifically collected for the open budget and spending data publishing/analytics life cycle. We proceed with instantiating this architecture and improve an existing open data platform (OpenSpending.org) for open fiscal data by adding support for semantic fiscal data and integrate additional tools for data analysis and citizen participation. Later, we evaluate the platform in terms of usability and applicability in real-world scenarios provided by three different municipalities, as well as analyzing how much of the requirements have been satisfied.

4. *A framework for linking multilingual fiscal concepts.*

   Contribution for RQ4. Semantics allows providing mapping across similar concepts. However, this mapping should be created first to make e.g., comparative analysis across datasets possible. Open fiscal data from different cities are rich in fiscal concepts, yet these concepts are published in their own language, making the mapping of similar concepts more challenging. To contribute to solving this multilingual mapping problem, we devise a framework, namely *Interlinking of Heterogeneous Multilingual Open Fiscal DaTA* (IOTA). IOTA uses fiscal data classifications in conjunction with machine translations to provide mappings for heterogeneous and cross-lingual data coming from different regions. Three language pairs (German-Spanish, German-French, and Spanish-French) have been tested with this approach. IOTA also provides a comparative analysis of 19 different string similarity measures for fiscal data linking. Since performing such mapping involves a computationally expensive task, IOTA uses the distributed scalable computing framework to enable complex string similarity assessment over large datasets.

5. *An ontology to make the OpenAPI-based API endpoint semantically discoverable.*

   Contribution for RQ3. Some open data portals provide API endpoints, and these endpoints can be described in open standard documentation, namely OpenAPI (or formerly known as Swagger). We propose a non-intrusive approach for the addition of semantic annotations (similar to RDFa and JSON-LD for HTML) to specific fields of the OpenAPI Specification. We created a lightweight vocabulary for describing RESTful web services using this specification. Furthermore, we practically demonstrate how OpenAPI objects can be enriched with semantic descriptions in a minimally invasive way by adding URIs in the values of chosen OpenAPI properties.

6. *A prototype for semantic-based comparative analysis of open fiscal data.*

   Contribution for RQ4. At last, we facilitate the comparative analysis of fiscal data by providing a prototype that uses the existing technologies and advancements for data interlinking and transformation. In addition, we demonstrate the applicability of the proposed proof-of-concept on real-world heterogeneous fiscal datasets.

1.4.2 List of Publications

This thesis is based on the following publications:

1. **Fathoni A. Musyaffa**, Lavdim Halilaj, Ronald Siebes, Fabrizio Orlandi, Sören Auer, *Minimally Invasive Semantification of Lightweight Service Descriptions*, Proceedings of the 23rd International Conference on Web Services 2016; San Francisco, CA, USA. Some open data initiatives publish their open data in the form of API. OpenAPI standards, formerly known as SwaggerAPI, is a standard in publishing API's metadata, which can potentially be utilized to assemble open datasets with similar characteristics that are published via APIs. This paper is a joint work with Lavdim Halilaj, a former Ph.D. student at the University of Bonn. In this article, I am the main contributor and taking role in analyzing the OpenAPI standard as well as designing an ontology that able to represent the OpenAPI specification in the RDF format.

2. Jindřich Mynarz, Jakub Klímek, Marek Dudáš, Christiane Engels, **Fathoni A. Musyaffa**, and Vojtech Svátek, *Reusable transformations of Data Cube Vocabulary datasets from the fiscal domain*. Semstats 2016 in ISWC. Kobe, Japan. This paper is a collaboration with colleagues from the University of Economics, Prague. In this paper, I contribute to providing a use case of reusable transformation pipelines for open fiscal data.

3. **Fathoni A. Musyaffa**, Christiane Engels, Maria-Esther Vidal, Fabrizio Orlandi, Sören Auer, *Experience: Open Fiscal Datasets, Common Issues, and Recommendations*. ACM Journal of Data and Information Quality, 2018. This paper is a joint work with Christiane Engels, a Ph.D. student at the University of Bonn. There are several contributions I have done for this paper. I collect eligible budget and spending datasets from a wide range of public administrators, survey quality factors for both generic and budget-/-spending-specific datasets, conduct a questionnaire regarding the importance of those quality factors, analyze and score collected datasets according to the quality factors and questionnaire result, rank the datasets according to the score, and finally provide a recommendation on best practices of publishing budget and spending datasets.

4. **Fathoni A. Musyaffa**, Fabrizio Orlandi, Tiansi Dong, Lavdim Halilaj, *Open-Budgets.eu: A Distributed Open-Platform for Managing Heterogeneous Budget Data*, SEMANTiCS 2017. Poster Paper.

5. **Fathoni A. Musyaffa**, Fabrizio Orlandi, Maria-Esther Vidal, Hajira Jabeen. *Classifying Data Heterogeneity within Budget and Spending Open Data*. International Conference on Theory and Practice of Electronic Governance (ICEGOV) 2018. Galway, Ireland. Integrating budget and spending datasets is difficult due to the decentralized nature of its publication. In this paper, I contribute to 1) analyzing every heterogeneity factor we found on budget and spending datasets across different public administration levels and languages, 2) analyzing how compatible each

heterogeneity factor with state-of-the-art data models to represent open budget and spending data, and 3) recommending actions that should be taken by open budget and spending data publishers, civil communities, and academics that work on open fiscal data domain, as well as data model developer/maintainers.

6. **Fathoni A. Musyaffa**, Lavdim Halilaj, Yakun Li, Fabrizio Orlandi, Hajira Jabeen, Sören Auer, Maria-Esther Vidal, *OpenBudgets.eu: A Platform for Analyzing Semantic and Open Fiscal Data*. International Conference on Web Engineering (ICWE), 2018. Caceres, Spain. This is joint work with Yakun Li, Fabrizio Orlandi (a former Fraunhofer IAIS postdoctoral researcher), and Lavdim Halilaj, a former student of the University of Bonn. This paper describes a platform architecture to annotate, semantically transform, store, visualize, and analyze the open budget and spending data with a specific semantic fiscal data model.

7. **Fathoni A. Musyaffa**, Maria-Esther Vidal, Fabrizio Orlandi, Jens Lehmann, Hajira Jabeen. *IOTA: Interlinking of Heterogeneous Multilingual Open Fiscal DaTA*. Elsevier Journal of Expert Systems with Applications (ESWA). 2020. Integrating budget and spending datasets originating from different public administrations requires a mapping of similar concepts from different data sources. In this paper, I contribute in 1) designing an experiment to map similar concepts from different datasets, 2) experimenting with 19 string similarity measures, 3) implementing big data approach to handle a large number of string comparison, 4) creating a framework consisting of machine translation, string similarity measures, and optimization based on cluster computing, and 5) evaluating the experiment result.

8. **Fathoni A. Musyaffa**, Jens Lehmann, Hajira Jabeen. *Cross Administration Comparative Analysis of Open Fiscal Data*. International Conference on Theory and Practice of Electronic Governance (ICEGOV) 2020. Athens, Greece. In this paper, my contribution ranges from designing a pipeline to enable comparative analysis based on previous works, selecting appropriate datasets to be used as a proof of concept, implementing the pipeline by using approaches I have experimented in the past.

The entire list of publications completed during this Ph.D. study can be found in Appendix E.

1.5 Thesis Structure

This thesis is comprised of five parts. Each part has one or more chapters. Part I provides the introduction, preliminaries, and general related work. In the introduction chapter, the overall motivation, challenges, research questions, and contributions are described. This chapter is followed by a background chapter, explaining the terms used in the rest of the chapters. Subsequently, the general, relevant works are also provided in this first part. Part II focuses on fiscal datasets' current quality state, publication recommendation,

heterogeneities, and data models. Part III elaborates the semantic OpenAPI description, datasets semantic enrichment process, and open fiscal data platform design as well as its implementation. Cross-datasets mapping and a proof of concept for cross-lingual comparative analysis are provided in Part IV. Lastly, Part V concludes the thesis by revisiting research questions as well as the future direction for the research on the semantic open public fiscal data domain.

CHAPTER 2

Background

In this chapter, we define some commonly used terms in this thesis to lay the ground for the next subsequent chapters. First, we explain the basic terms we use within public administrations and open data in section 2.1 to familiarize the readers with the open data domain. This is especially relevant for Part II of this thesis as well as for some chapters in Part III. The background is continued with the explanation of semantic web in section 2.2, which is relevant for Part III and Part IV of this thesis. In the end, some computing terms and technologies related to data integration and processing are elaborated in section 2.3 to also familiarize the readers with the work provided in Part III as well Part IV.

2.1 Public Administrations and Open Data

The term *public administration* refers to the concern regarding how public programs are managed, which ranges from different levels (local, international, organizations, associations, interest groups) and different interests (e.g., human resources, financial resources, infrastructure construction, etc.) [18]. Kettl [19] states that public administration is an adaptation of politics into daily life as seen by the citizens. According to Rosenbloom [20], public administrators' roles are related to three views and functions: 1) *managerial approach*, which relates to government executive function (policy implementation), 2) *political approach*, which relates to government legislative function (policy-making through legislation enactment), and 3) *legal approach*, which relates to judicial function (law interpretation). Within the political context, the actions of public administrators are in the form of commitment to the ideals and practices of democracy. When compared to a business entity, public administrations are focused more on service delivery, as well as public interest-based individual/group behavior regulation. Additionally, compared to business, public administrations are more ambiguous (e.g., how objectives are specified and measured?), more plural in decision-making, and more visible operation [18]. The open data movement helps in making the governance service delivery more transparent.

Open Data

According to Open Data Handbook,[1] open data is defined as data in which *anyone* can use, reuse and redistribute the data. These abilities, however, imply that: 1) the data should be interoperable across different systems and organizations in the form that is convenient and can be modified, 2) data intermixing is permitted, 3) the data is available as a whole, with cost less than the cost of reproduction, and 4) there is no restriction (e.g., non-commercial, education-only, etc.) on the data [21]. Open data can be published by a different type of organization, but governmental public administrations are one of the most common organizations that typically publish open data, as one of the major differential characteristics for them is their visibility. There are different domains of open data that are published by these public administrations. Open Knowledge Foundation has conducted a survey in 2013-2015 regarding the availability of these domains.[2] This survey checks various domains of data published from each country, including government budget, national statistics, procurement, national laws, administrative boundaries, draft legislation, air quality, national maps, weather forecast, company register, election result, locations, water quality, government spending, and land ownership.

Fiscal, Budget, and Spending Data

According to a survey done by the Open Knowledge Foundation (OKF) on the Global Open Data Index (GODI) [6], among 94 observed countries in 2015, 88 countries have made their budget data publicly available, as can be seen in Figure 2.1 [1]. This makes budget data the most popular type of open data published by public administrations on the country level. The Open Data Barometer (ODB) [22] reports that budget and spending datasets are among the most important datasets, along with company registers, contracts, and land ownership, that are needed to restore public trust.

To provide a more detailed explanation of terms used around open fiscal data, we provide the definition of these terms that will be used within this thesis. These terms include:

- *Spending* defines the actual value that is spent on an item. In this paper, the *executed budget* is considered similar to spending.
- *Budget* contains a list of planned values to be spent in regard to specified dimensions and attributes (details on dimensions and attributes are explained in section 2.2, in particular, regarding Data Cube Vocabulary). Public budgets contain different budget phases, such as *draft* or *proposed* budget before it is approved by politicians, the *approved* budget after it is agreed upon by the politicians, *revised* or *adjusted* budget for a budget that has been changed with regard to the approved budget, and *executed* budget for the actual value paid after the budget of that particular item has been spent.

[1] https://opendatahandbook.org/guide/en/what-is-open-data/
[2] https://index.okfn.org/dataset/

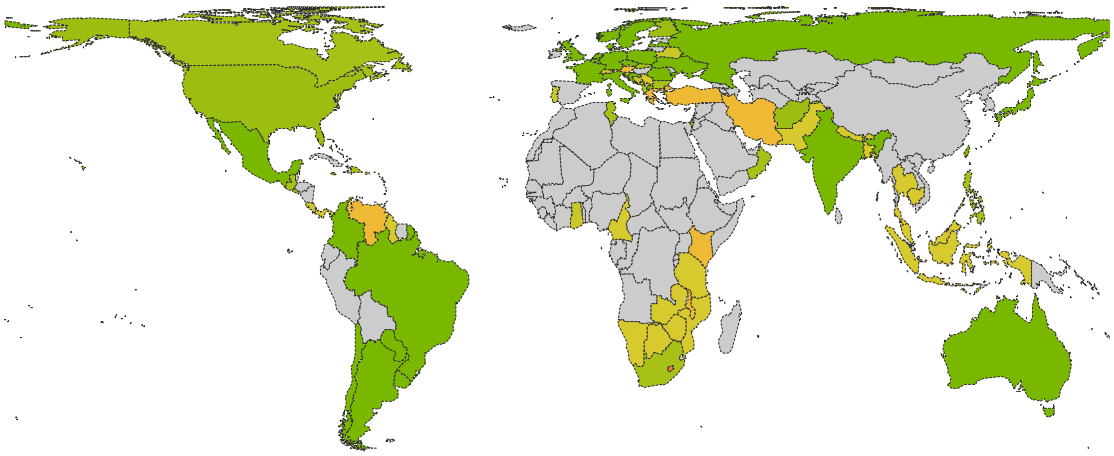

Figure 2.1: Countries surveyed by OKF that provides government budget datasets as mentioned in the GODI report [1]. Colorized country indicates the availability of budget datasets, with the intensity of the color expresses the degree of satisfaction regarding certain quality factors (open license, in open and machine readable format, downloadable at once, up-to-date, publicly available, and available free of charge).

- *Expenditure* refers to the amount of money budgeted to be spent on an item. To be consistent in this thesis, while *expenditure* refers to the budget that may have been or has not been spent, *spending* refers specifically to actual budgeted money that has *already* been spent.

- *Income* refers to the amount of budgeted money that would flow in as revenue for the corresponding public administration.

Classifications

Budget and spending data typically contain the temporal information (i.e., year, month, or date), the amount of money being received or spent, and labels that indicate the explanation of the amount being received or spent. These labels are normally a set of controlled terms/vocabulary which is most likely independently enumerated across public administrations, organized as a specific type of *classifications*. In this thesis, we refer to *classification* as a set of controlled terms published by respective official bodies to categorize budget/spending items, consisting of concise textual labels. These classifications can also be referred to as *vocabulary* or *code list*. The common classifications that are published by open fiscal data publishers include:

- *Functional classification* describes the expense usage (e.g., `Vivienda y Urbanismo` or, translated into English, `Housing and Urbanism` concept as found in the 2013 Aragon Government Budget[3]).

[3] https://opendata.aragon.es/catalogo/presupuesto-gobierno-aragon-2013

- *Administrative classification* states which administrative office is responsible for a particular budget line (e.g., `Secretaría General Técnica de Obras Públicas, Urbanismo, Vivienda y Transportes`, or in English, `General Technical Secretariat of Public Works, Urban Planning, Housing, and Transportation` found in the same Aragon datasets).
- Other classifications, include: *economic classification* (e.g., `capital transfers, real investments`), *procurement items* (e.g., `Agricultural products`, as well as `Electricity and heating`), and so on.

In practice, there are a lot more classification types and these types have their own characteristics. For example, 1) some datasets are published with or without unique keys, 2) datasets are published in different languages, 3) some datasets are published with or without hierarchy, and so on. The use of standardized vocabulary increases open fiscal data reusability and enables the comparative analysis of fiscal datasets.

Some classifications are standardized by international bodies. For example, *Classification of the Functions of Government*/COFOG [23], a functional classification developed by the United Nations and *Common Procurement Vocabulary*/CPV [24], a procurement item classification by the European Union. In reality, however, very few datasets use standardized classifications. In such cases, non-standard classifications published by different public administrations can potentially be mapped for similar concepts, and then can be exploited further to improve data reusability and data comprehension.

2.2 Semantic Web Stack

By the end of the 90s, an idea emerged to make the World Wide Web understandable by machines.[4] This idea, coined as Semantic Web,[5]is implemented through several subsequently-published data model, specifications, standards, and recommendations published by the World Wide Web Consortium (W3C). The main property of the semantic web is the distribution of the data across multiple sources. The machine-understandability aspect of the idea is achieved by adding metadata within published information. This metadata is provided by a specific vocabulary and controlled terms. A semantic web cake or semantic web layer is commonly illustrated in Figure 2.2, proposed its mainly top layers are still evolving. However, its lower layers have been mature for some time, including the concepts of URI, XML, RDF, RDF-S, Ontology and SPARQL. Most of these concepts are elaborated in this section.

Resource Description Framework (RDF)

Resource Description Framework (RDF) is a specification by the W3C to represent and exchange data over the World Wide Web [26]. Data items in RDF are represented as a Uniform Resource Identifier (URI) if it represents specific things or objects, or the data

[4] http://www.dblab.ntua.gr/~bikakis/XMLSemanticWebW3CTimeline1.2.pdf
[5] https://www.w3.org/standards/semanticweb/

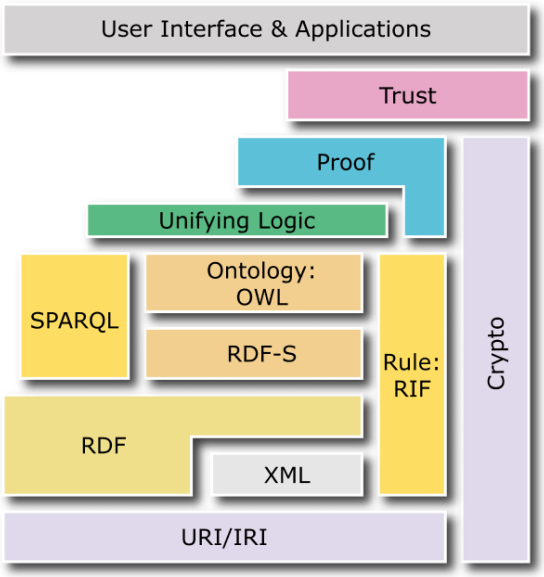

Figure 2.2: The intended semantic web technology stack [25], which keeps evolving. Image copyright 2007 W3C (MIT, ERCIM, Keio, Beihang).

items can also be represented as a literal especially for representing values (e.g., amount of spending by government for a specific budget). The relationship between data items is represented using a Subject-Predicate-Object (SPO) pattern coined as a *triple* in RDF specification.

Prefixes are defined at the beginning of RDF documents to ease notation. The prefix part enlists the abbreviated terms of the URIs used. URI consists of a protocol (e.g., https), server name, path on the server as well as the fragment identifier of the object being presented. The use of URIs provides unique identifiers for the resources, and it can be dereferenced.

For example, the RDF data snippet in Listing 2.1 contains three prefixes, `obeu-dimension`, `obeu-ds`, and `dbr`, pointing to respective URIs. It has `obeu-ds:budget-thessaloniki-expenditure-2017` as a subject, `obeu-dimension:organization` as a predicate, and `dbr:Thessaloniki` or `http://dbpedia.org/resource/Thessaloniki` as an object.

```
@prefix obeu-dimension: <http://data.openbudgets.eu/ontology/dsd/dimension/>
@prefix obeu-ds: <http://data.openbudgets.eu/resource/dataset/>
@prefix dbr: <http://dbpedia.org/resource//>

obeu-ds:budget-thessaloniki-expenditure-2017 obeu-dimension:organization
↪  dbr:Thessaloniki
```

Listing 2.1: A snippet information represented in RDF data model using Turtle serialization.

The triple describes that the dataset of `budget-thessaloniki-expenditure-2017` has an `organization` (as in, associated with) the city of `Thessaloniki`. All the triple components: the subject (Thessaloniki budget expenditure 2017 dataset), the predicate (organization), and the object (Thessaloniki) are represented as URIs in the triple. The details and further information of Thessaloniki are published publicly, and extra information on it (e.g., area size, population size, and other information) can be traced by following the link to `http://dbpedia.org/resource/Thessaloniki` that provide further information regarding the object.

RDF comes with its base terms to lay the foundation for representing information on an abstract level.[6] The IRI namespace for RDF is `http://www.w3.org/1999/02/22-rdf-syntax-ns#` and `rdf` is frequently used as the prefix. The terms include properties such as `rdf:about`, `rdf:value`. and `rdf:type` as well as class such as `rdf:Property`. In a more advanced manner, RDF also facilitates the use of *blank nodes*. Blank nodes are useful when a name for an exiting resource is not needed, as well as when certain information needs to be grouped together.

By representing data in RDF, the relationship between each granular item in the datasets could be made explicit and referenced. Any data item in the triple that is represented as a URI can then be referenced and linked with other related data [7]. To publish the data as RDF, several design issues need to be considered, such as [27]:

1. Using URIs to name things,
2. Using HTTP so that the URI can be looked up,
3. Using RDF/SPARQL standard to provide useful information when the URI is looked up, and
4. Including links to other URIs to make the data more discoverable.

Ontologies

The metadata within the semantic web stack is represented formally using an *ontology*. An ontology typically defines classes, individuals, attributes, and relations of a certain domain. A *class* declares a concept or a category, while *individual* is an instantiation of a class. An *attribute* defines the properties of objects (classes or individuals). A *relation* formally describes the relationship between classes and individuals. An ontology that defines the abstract concepts, their properties and relations is known as *upper ontology*. An upper ontology provides the generic terms of concepts so that it could be used to define more-specific ontologies [28]. An example of an upper ontology is Dublin Core, an ontology to model digital resources.[7] In contrast, there are ontologies that are developed to model very specific domains. This specific ontology is known as *domain ontology*.

[6] https://www.w3.org/TR/rdf-syntax-grammar/
[7] https://dublincore.org/specifications/dublin-core/

OpenBudgets.eu (OBEU) [29, 30] ontology is an example of domain ontology to represent public fiscal data (see chapter 7).

As a good practice in the ontology engineering field, an ontology should be reused to promote data linking. Figure 2.3 illustrates the organization ontology [31], an ontology to represent information regarding information about organizations. It can a how other ontologies are used to define these ontologies, such as FOAF[8] (Friend of a Friend). FOAF is designed to describe persons, their relation to other people and objects, as well as their activities. In the case of the organization ontology, some terms in FOAF are used to represent `Agent` and `Person`.

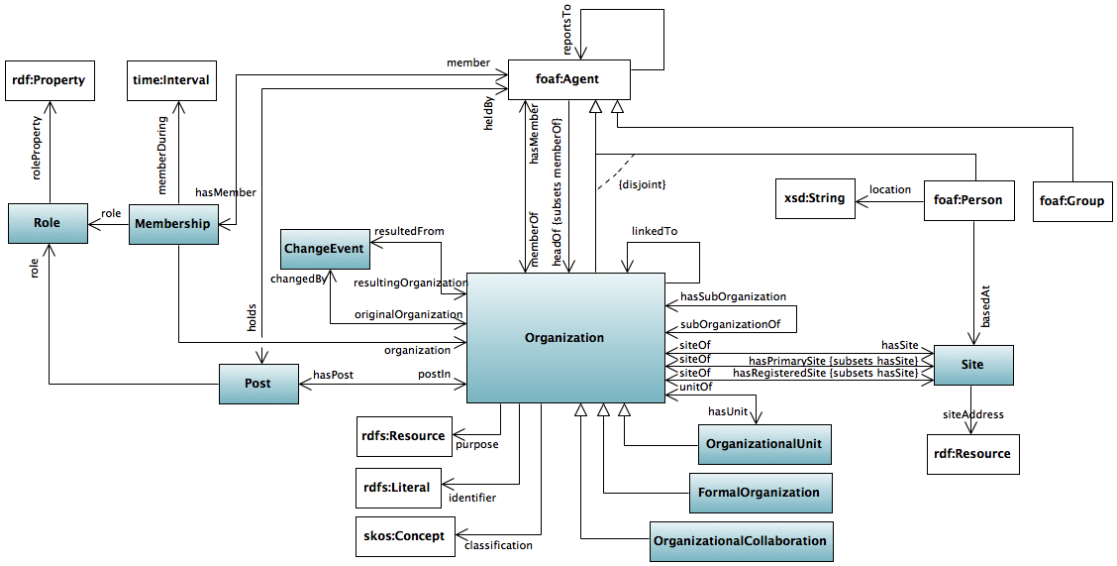

Figure 2.3: The organization ontology concepts and its relations [31]. Image copyright 2012-2014 W3C (MIT, ERCIM, Keio, Beihang).

Linked Open Data (LOD)

Datasets published in RDF acts as a building block for *Linked Open Data*. A guide on publishing Linked Data is summarized by Bauer and Kaltenböck [32]. Publishing data in RDF enables datasets from different sources to act as a global database [32]. This differs from the older paradigm of accessing a conservative database and silos, in which access to the data inside those datasets is private and locked up in a certain application [32]. By publishing datasets in RDF, data from different sources can be combined to enrich the context of information being analyzed.

In the past few years, the number of datasets provided in RDF as Linked Open Data has increased. Linked Open Data can be used to enrich open fiscal datasets for further analysis. For example, DBpedia [33] provides huge information extracted from Wikipedia.

[8] http://xmlns.com/foaf/spec/

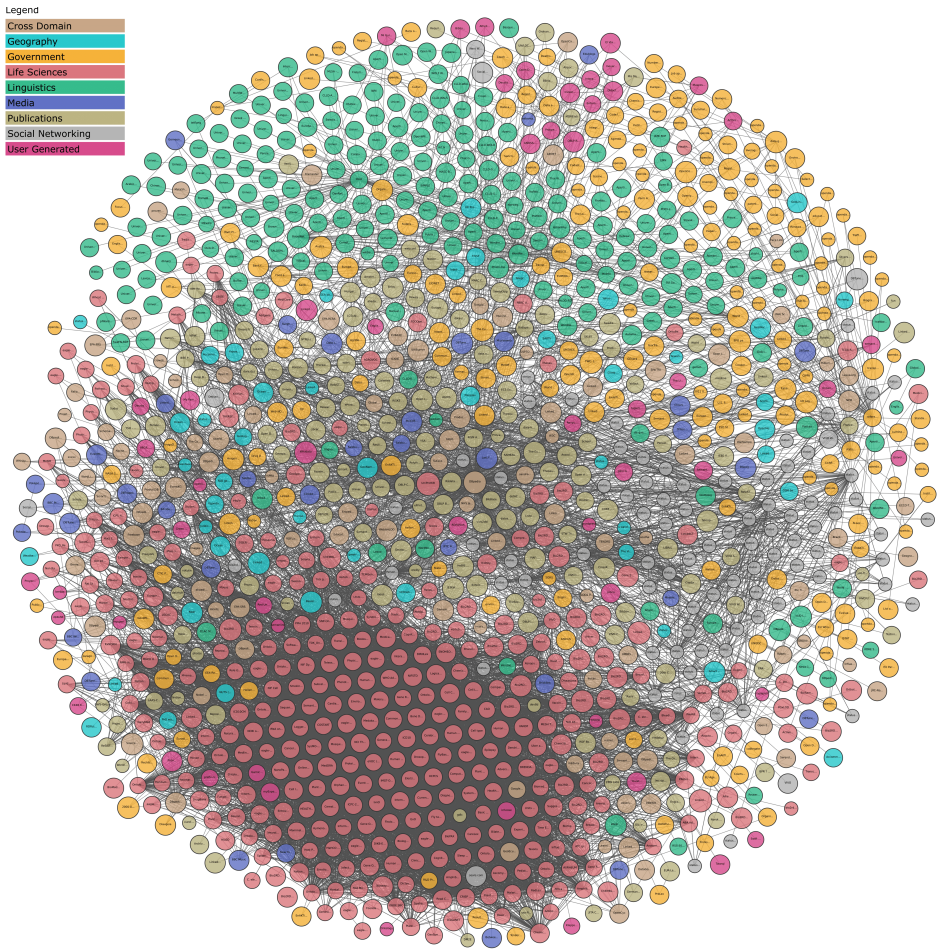

Figure 2.4: A current state of Linked Open Data cloud registry, with each color represent specific major domain (derived from `lod-cloud.net`) [35].

The English version of DBpedia (version 2016-04) contains 1.3 billion triples. A sister project of Wikimedia Foundation, Wikidata [34], provides a knowledge base in RDF that is collaboratively edited in a more fine-grained manner ensuring higher quality control over the information provided, although the amount of information is not as much yet when compared to the information automatically extracted in DBpedia.

LOD cloud registry[9] provides an overview of linked open data which are interconnected publicly available over the web, as illustrated in Figure 2.4. The registry only accepts submission of data that satisfy open linked data requirements: can be resolved by HTTP URIs, resolve to RDF data in popular RDF formats, contain minimum 1,000 triples, connected to other datasets in the LOD cloud registry, and can be accessed by either RDF crawling, RDF dump or a SPARQL endpoint [35]. As of May 2020, there are 1,255 datasets with 16,174 links registered to the LOD cloud registry [35].

[9] `https://lod-cloud.net/`

RDF Serializations

Data formatted in RDF can be saved or transmitted using different syntax. The process of transforming data into a representation that can be saved or transmitted is called *serialization*. Certain types of serialization are aimed at slightly different purposes (human reader vs. machine readability). The following are some example of each RDF serialization:

1. RDF/XML format is based on XML which uses XML tags to provide metadata. Since the W3C introduces the semantic web using this format, there has been some confusion to refer to RDF (data model) as RDF/XML. RDF/XML serialization tends to be not intuitive for humans to read.

2. JSON-LD (JavaScript Object Notation for Linked Data) is a serialization of RDF using JSON format, intended to ease developers in transforming their JSON data into semantic, linked data format.

3. Turtle (Terse RDF Triple Language) is a simplification of RDF serialization and has become a W3C recommendation due to its increasing popularity and its ease for humans to read. An example of information serialized in Turtle is provided in Listing 2.2, which also illustrate the use of SKOS ontology.

4. Other RDF Syntaxes. Other serialization includes N-triples, Notation 3 (N3), and Microformats.

Resource Description Framework Schema (RDFS)

RDF is limited in terms of describing further typing abilities, therefore, RDFS is proposed by the W3C to describe related resources as well as their relationships. RDFS specification is formally defined in RDFS Vocabulary.[10] The IRI namespace for RDFS is `http://www.w3.org/2000/01/rdf-schema#` and `rdfs` is commonly used as the prefix. RDFS defines class (`rdfs:Class`) and subclass (`rdfs:subClassOf`) to state the relationship between resources. Using RDFS, the type restriction of the valid domain (subject) using `rdfs:domain` and range (object) using `rdfs:range` can also be specified. The sub-properties relation between properties can also be defined using RDFS using `rdfs:subPropertyOf`. To increase human-readability for the described concepts, comments and label can be provided using `rdfs:comment` and `rdfs:label`.

Web Ontology Language (OWL)

In addition to RDFS, OWL[11] is an ontology language that expands the expressiveness of RDFS, allowing the formal definition of resources with various axioms within the domain in a wider manner. The IRI namespace for OWL is `http://www.w3.org/2002/07/owl#` and `owl` is frequently used as the prefix. OWL allows, among others:

[10] `https://www.w3.org/TR/rdf-schema/`
[11] `https://www.w3.org/TR/owl2-overview/`

1. Enumeration (`owl:oneOf`) that restricts only certain individuals/instances can be allowed to be a member of a specific class;

2. Property restrictions through value constraints (`owl:allValuesFrom` or `owl:someValuesFrom`), as well as cardinality (`owl:cardinality`) constraints;

3. Classes description using different description logic terms (`owl:complementOf`, `owl:intersectionOf`, `owl:unionOf`), as well as (in OWL 2) disjoint (`owl:disjointWith`) to state that there is no shared instances between subject class and specified object class;

4. Describing the similarity between individuals using `owl:sameAs` or `owl:differentFrom` properties.

Simple Knowledge Organization System (SKOS)

Knowledge Organization System (KOS) is a set of tools to manage extensive collections of objects which have been a long practice in information and library science [36]. These objects are, for example, books and museum artifacts. From the practice of managing these objects, certain knowledge organization systems appear, including taxonomies, classification schemes, subject heading systems, and thesauri. SKOS[12] is a common data model that is intended to provide a bridge between semantic web and KOS, allowing the precise and systematic description of large-scale information resources. The IRI namespace for SKOS is `http://www.w3.org/2004/02/skos/core#` and `skos` is frequently used as the prefix. In SKOS, *concepts* (`skos:Concept`) are organized into a concept scheme (`skos:ConceptScheme`). Similar to other resources in semantic web, concept and concept schemes are determined using URIs. Concepts can be: 1) labeled (`skos:prefLabel` and `skos:altLabel`) using Unicode strings with language tag, 2) designated with notations (`skos:notation`) for unique identification within the scope of its concept scheme, 3) annotated with different types of notes, 4) associated with other SKOS concepts from other concept schemes using hierarchical (`skos:narrower` and `skos:broader`), associative, close equivalent, or exact equivalent properties, and 4) extended using optional extension.

Listing 2.2 illustrates a sample of Aragon's public functional classification concepts[13] represented using terms defined in SKOS. The whole functional classification structure is represented as a concept scheme, which consists of several concepts. The underlying concepts may consist of hierarchical structure, for example, *"Culture"* and *"Education"* are sub-terms of *"Production of public goods of Social character"*.

[12] https://www.w3.org/TR/swbp-skos-core-spec/
[13] https://opendata.aragon.es/

```
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix skos: <http://www.w3.org/2004/02/skos/core#>.
@prefix dc: <http://purl.org/dc/terms/>.
@prefix dcat: <http://www.w3.org/ns/dcat#>.
@prefix arfunc:
↪ <http://data.openbudgets.eu/resource/codelist/estructura_funcional_aragon_2014/>.
<http://data.openbudgets.eu/resource/codelist/estructura_funcional_aragon_2014>
↪ dc:description "The codelist of Aragon (ES) functional classification in 2014."@en ;
    dc:modified "2017-01-27"^^xsd:date ;
    dc:publisher <http://openbudgets.eu/> ;
    dc:title "Aragon functional classification codelist 2014"@en ;
    a dcat:Dataset , skos:ConceptScheme ;
    dcat:keyword "codelist, budget, aragon, 2014, functional classification"@en ;
    rdfs:label "Estructura Funcional Aragon 2014"@en.
arfunc:4 a skos:Concept ;
    skos:altLabel "Producción de Bienes Públicos de Carácter Social"@es , "Production of
↪ public goods of Social character"@en ;
    skos:notation "4" ;
    skos:prefLabel "Prod.Bienes Púb. c.social"@es , "Prod.Bienes Pub. c.social"@en ;
    skos:inScheme
↪ <http://data.openbudgets.eu/resource/codelist/estructura_funcional_aragon_2014>.
arfunc:42 a skos:Concept ;
    skos:altLabel "Educación"@es , "Education"@en ;
    skos:broader arfunc:4 ;
    skos:notation "42" ;
    skos:prefLabel "Educación"@es , "Education"@en ;
    skos:inScheme
↪ <http://data.openbudgets.eu/resource/codelist/estructura_funcional_aragon_2014>.
arfunc:45 a skos:Concept ;
    skos:altLabel "Cultura"@es , "Culture"@en ;
    skos:broader arfunc:4 ;
    skos:notation "45" ;
    skos:prefLabel "Cultura"@es , "Culture"@en ;
    skos:inScheme
↪ <http://data.openbudgets.eu/resource/codelist/estructura_funcional_aragon_2014>.
```

Listing 2.2: Examples of concepts represented using SKOS terms, serialized using RDF Turtle serialization format.

Data Cube Vocabulary (DCV)

Several international organization (UN, OECD, IMF, Eurostat, ECB and BIS) worked together in 2001 to make statistical practice more efficient. The initiative, named SDMX,[14] resulted in the widely-adopted SDMX technical specification (ISO:TS 17369) and the SDMX Content-Oriented Guidelines (COG). COG allows terminology sharing across SDMX adopters by providing a collection of categories, code lists, and concepts from multiple domains, thus, backing interoperability and comparability among datasets.

[14] www.sdmx.org

Data Cube Vocabulary (DCV) specification[15][37] allows the adoption of SDMX in linked data, allowing the publication of multidimensional data in RDF. DCV has `http://purl.org/linked-data/cube#` as the namespace IRI, with `qb` as its frequently-used prefix. The summary of terms in the DCV along with the relationship between each term is illustrated in Figure 2.5. Copyright © 2012-2014 W3C® (MIT, ERCIM, Keio, Beihang)

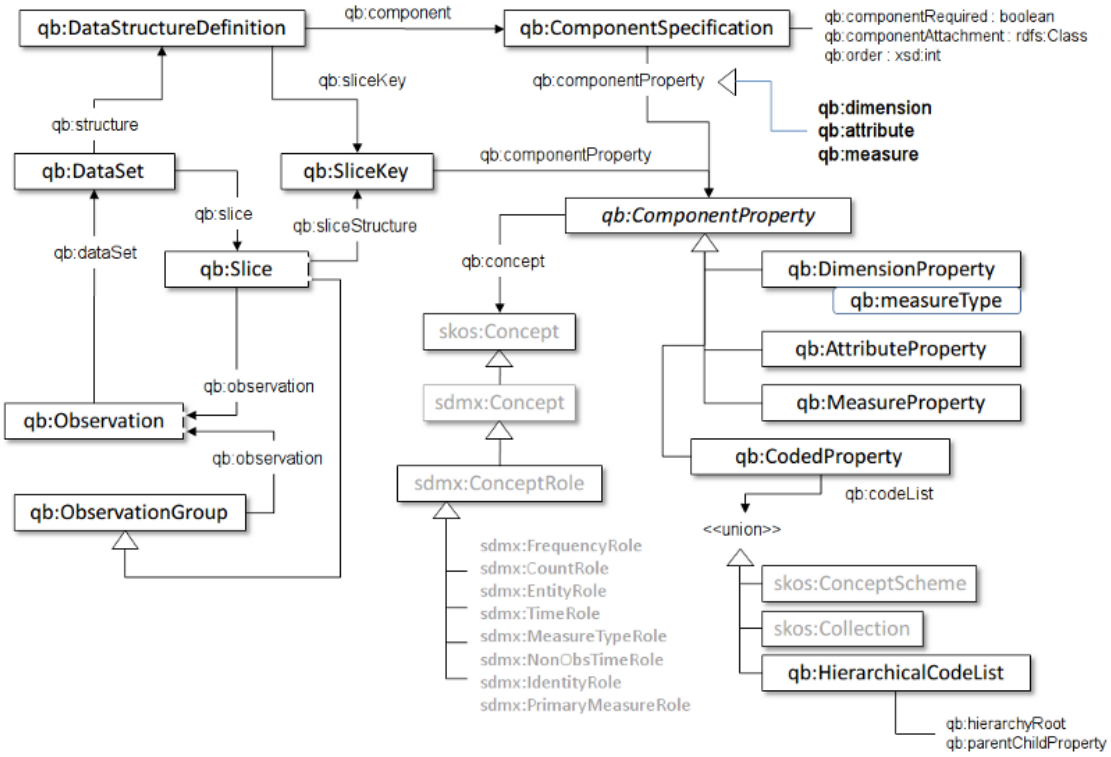

Figure 2.5: Summary of terms in DCV, as illustrated in [37]. Image copyright 2012-2014 W3C (MIT, ERCIM, Keio, Beihang).

Statistical data are often thought of as multidimensional space, also known as hypercube, but often referred to simply as cube even though the number of dimensions is not necessarily three. A cube comprised of three basic components: dimensions, measures, and attributes. The following definitions of terms are related to DCV, which are relevant through many parts of this thesis.

- *Dataset* (`qb:DataSet`) in terms of data cube refers to any set of statistical data which has the following properties: 1) observations - a statistical table which has measured values in table cells, 2) organizational structure - the value of each dimension can be determined when the observation is known, 3) structural metadata to determine the structure e.g., the datasets have a certain unit of measurement, datasets have

[15] https://www.w3.org/TR/vocab-data-cube/

normal value or a series break, or whether the value in the datasets are estimated or measured, 4) reference metadata which provides a whole description of metadata (e.g., datasets categorization, publisher, accessible SPARQL endpoint).

- *Observation* (`qb:Observation`). A single phenomenon being observed which contains the measure, dimensions, and attributes.
- *Measure* (`qb:MeasureProperty`, needs to be instantiated). A component of a cube that shows the value of the aspect being observed. For example, how much money is spent on high school by the office of education on a certain fiscal year.
- *Dimension* (`qb:DimensionProperty`, needs to be instantiated). A component of a cube that identifies the observations. A combination of dimensions makes the observation unique. For example, the functional usage and administrative office on observation within a budget dataset, or geographic region on a certain observation.
- *Attribute* (`qb:AttributeProperty`, needs to be instantiated). A component of a cube that qualifies and interprets the values of an observation. For example, the currency of spending being used on an observation.
- *Slices* (`qb:Slice`). A subset of observations that have been grouped by e.g., making all the other dimensions fixed with an exception that a single dimension can vary.
- *Dimension, classification,* and *code list. A dimension* defines the qualitative element of a budgeting line [38]. The term dimension corresponds to the definition within Data Cube Vocabulary (DCV). One particular type of dimension is a *classification.* The catalog that enlists the possible values of classification is coined as a *code list.*

Publishing statistical data as linked data in RDF allows several advantages, such as 1) making observations or a collection of observations addressable from the web, 2) allowing different datasets to be combined, 3) letting previously-static, published datasets to be machine-readable, 4) allowing the reuse of standardized components and tools [37]. DCV utilizes available ontologies and vocabularies such as `skos`, `scovo`, `dc`, `void`, `foaf`, and `org`. The URIs for these namespaces can be found on the prefix.cc portal,[16] and the terms can also be searched via The Linked Open Vocabulary website.[17]

SPARQL

SPARQL, abbreviated recursively from *SPARQL Protocol and RDF Query Language*, is a query language designed to query semantic databases that are stored in the RDF data model [39]. These semantic databases are also known as a *triple store*, due to the fact that the information stored in these databases is provided in a triple format as previously described in the RDF section. SPARQL's syntax is adapted partially from the popular database query language, SQL (Structured Query Language), to increase its

[16] http://prefix.cc/
[17] http://lov.linkeddata.es/dataset/lov

adoption rate. There are four types of SPARQL queries, `SELECT`, `CONSTRUCT`, `ASK`, and `DESCRIBE`. `SELECT` is used when certain data needs to be returned based on the provided matching pattern, by considering the constraints provided in the subsequent `WHERE` clause. `CONSTRUCT` is utilized to formulate a valid RDF graph based on the template provided in the query. `ASK` is used to get a binary answer (true/false) regarding queried statement. Finally, `DESCRIBE` is performed when all the information describing a resource needs to be retrieved. Using SPARQL, data from multiple graph sources can be retrieved. This is known as a *federated query*. Some examples of these queries are provided in the later chapters (e.g., Listing 8.1 in chapter 8 for `SELECT` and `CONSTRUCT` query, and Listing 11.1 in chapter 11 for the federated query).

2.3 Data Processing and Integration

In this section, we briefly go through some concepts in data processing and integration. API calls can be utilized to fetch the relevant data from open data portals that provide API endpoints. The discussion regarding the utilization of semantic API is available in Part III, and hence an introduction about API is provided in this section. To process the concept mapping from translated fiscal concepts, big data and cluster computing tools are used later on in Part IV. For this reason, we provide the brief introduction to big data and Apache Spark here as well.

Application Programming Interface (API) and OpenAPI

Application Programming Interface (API) is an interface in which various software intermediaries interact with each other. For the interface to work, a clear documentation is needed regarding what is expected from the API, the type of call/requests that can be made and how it can be done, what are the prerequisites prior to making these calls, and what type of response and data model/format will be returned from the API. *Web APIs* are mechanisms to enable exposure of the data and operations on these data to third-party clients through web protocols (e.g. HTTP and HTTPS), commonly uses GET and POST parameters as input, and usually returns data in a standard format such as CSV, XML, JSON, TSV or HTML.

Traditionally, the allowed and expected communication flows between a client and the API often is expressed in Web Service description standards like *WADL*[18] or *WSDL*.[19] Recently, we can add a new standard referred as the OpenAPI standard to the client-API communication flow category, although it was originally not intended to be processed by machines except for generating human-readable documentation. *OpenAPI Specification*[20] is a specification developed to describe RESTful APIs in a standardized way. An interactive visualization of OpenAPI terns is provided by [40] and accessible online, with some of its terms is can be seen in Figure 2.6.

[18] https://www.w3.org/Submission/wadl/
[19] https://www.w3.org/TR/wsdl
[20] https://openapis.org/

Figure 2.6: A partially-expanded visualization of terms defined in OpenAPI specification, generated by [40].

Swagger framework[21] is a set of tools used to make RESTful services documented by the OpenAPI Specification. This framework consists of a code editor, a GUI rendering tool (Swagger UI), and a code generator (Swagger Codegen). The code editor is an IDE that provides code-editing features (e.g syntax highlighting, code validation, auto-complete) for YAML[22] file that used to describe an API using OpenAPI specification. YAML is a serialization format that aimed to make data serialization a more human-readable in contrast to the XML serialization format which is more complicated for humans to read. Swagger Codegen generates a stub code containing annotations that will be further implemented by the API programmers. The Swagger UI generates the API documentation in HTML format.

[21] https://swagger.io/
[22] https://yaml.org/

Big Data

The widely adopted definition of big data considers any data with volume, velocity, and/or variety that are challenging to be processed with conventional methods and systems. Due to its complexity, it requires the development of a novel approach to answer previously inaccessible questions [41]. Occasionally, in addition to the previous big data properties, veracity also characterizes big data. All these properties are often referred to as *the four Vs* of big data. *Volume* characterizes that the data cannot be stored on a single machine for processing, therefore, it has to be distributed across cluster computer. *Velocity* refers to the fact that the data are produced at a speed that cannot be managed by the current methods. Variety characterizes different structures and formats of the data. *Veracity* considers the accuracy and the noise of captured, processed, and stored data [42].

Cluster Computing

A set of either desktop or server computers that are connected within a local area network and operates as a single large computer is considered as a *cluster* [43]. *Cluster computing* can be defined as a collection of computers that jointly perform a given task in a distributed manner. Cluster computing software framework, such as Apache Spark[23] [44], provides an implementation of computing task distribution across computers. To perform more efficient computation and task distribution, several features are designed and implemented within Apache Spark, such as the *Resilient Distributed Datasets* (RDD) data structure. RDD allows computations to be performed in-memory within large clusters in a fault-tolerant manner [45, 46]. In case the computed RDD does not fit in the host memory, Apache Spark automatically performs *spill to disk* operation which moves the RDD from host RAM to host disk. Apache Spark offers internal optimizations, such as an optimized operation for Cartesian join. Since Apache Spark requires *cluster manager* as well as *distributed file system*, Apache Spark provides *Spark Standalone* as a cluster manager and *Hadoop Distributed File System* (HDFS) as a distributed file system, among others. HDFS is a file system that is known for its scalability, portability, fault-tolerant, and distributed manner with a master/slave architecture.

[23] https://spark.apache.org/

CHAPTER 3

Related Work

In this chapter, we provide related works regarding open budget and spending data analytics. We begin with the state of fiscal data publication and budget participation, continued by heterogeneity challenges and efforts to achieve interoperability within open fiscal data, and finished with different platforms that aim to analyze open data or open fiscal data.

3.1 Fiscal Data and Budget Participation

Around the world, we saw dissatisfaction with the government in different forms. Dissatisfaction on governance can be, to some extent, indicated by the increasing popularity of nationalist politicians that have polarizing democratic views. Increasing problems with wealth inequality also contribute to the rise of anti-establishment revolts and populism in 2016-2017. Unfortunately, many governments respond with tighter control on civil society which serves a disservice to their governance [47]. The way political institutions and representations are structured might not meet the citizen's expectations, leading to anti-corruption protests. This especially happens in countries with middle-income [48].

On the other hand, many countries around the world have been publishing open budget and spending data, which can be a bridge between government/public administration and their citizens as well as involved stakeholders. Open Knowledge Foundation (OKF) conducted a survey during the year 2013-2015 [49], showing that the majority of surveyed countries are consistently publishing open data. The survey also indicates that during the transitions of those years, open budget datasets are more likely to be published, moving up from the third position to the first position as the open data domain that is most frequently published.

Having budget and spending data published openly can be used as an indicator to enable the analysis of whether political institutions and representations are prioritized according to citizens' expectations. Several questions can be answered when the budget data is made available openly, for example, [48]:

1. How much does the government spend on a certain purpose? This requires functional classifications (see section 2.1) of budget and spending data to be published.

2. Is the budget implementation has been in accordance with the legislative approval? This requires different stages of budget phases to be published.

3. Which achievement is being targeted by the government through the raised and spent money? This requires policy goals and targets to be published.

The budget transparency can be used as a bridge to facilitate a more trustworthy government, due to the fact that the citizen can be made aware of how financial resources are being collected and spent. This is especially true when budget-related decision making can involve citizens through a budget-related mechanism.

A comparative survey conducted by International Budget Partnership (IBP) in 2017 [48] assesses 115 countries regarding citizen budget participation using a budget participation score.[1] This score is based on a methodology considering transparency, participation, and accountability principles proposed by the Global Initiative for Fiscal Transparency (GIFT).[2] The principles include inclusiveness, timeliness, openness, and sustainability on different participation mechanisms. These participation mechanisms range from several engagement practices [48]:

1. The executive branch mechanism engages the public as the budget is formulated and engages the public as the budget is executed later.

2. The legislative mechanism engages the public before the budget approval through public hearings.

3. The auditor mechanism allows the citizen to submit reports during the auditing process and to track the auditing progress.

As a result, the 2017 IBP survey categorizes the surveyed countries into three different groups: countries with *low* budget participation (resulting in 47 countries), countries with *limited* participation (resulting in 42 countries), and countries with *sufficient* budget participation (resulting in 26 countries). This can be seen in Figure 3.1. The budget transparency report is updated in the second quarter of 2020 [50], with the budget-transparent countries illustrated in Figure 3.2 [50]. From these figures, it can be seen that more countries are moving towards sufficient and extensive transparency, with the trend of transparency score average for the reporting period 2017-2019 rising as a highest-ever average score since the IBP reports have been periodically released. This comparative table for the transparency average score can seen in Table 3.1.

The publication of public fiscal data enables the citizen to engage further with budget participation activities. Also, there are several participation use cases examples for budget data, for example [48],

[1] These materials were developed by the International Budget Partnership. IBP has given us permission to use the materials solely for noncommercial, educational purposes. See https://www.internationalbudget.org/library/copyright/ for more details.

[2] http://www.fiscaltransparency.net/

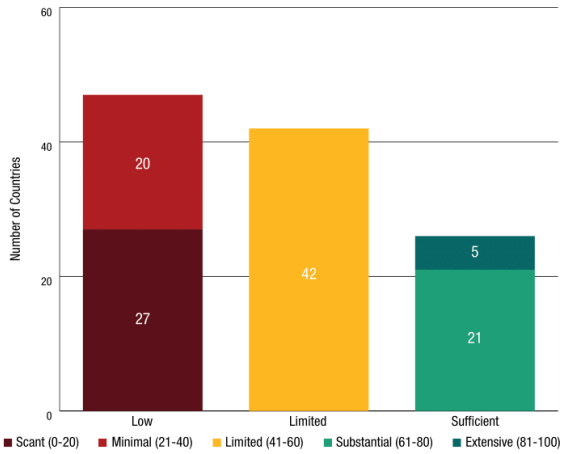

Figure 3.1: The distribution of budget-transparent countries in the period of 2015-2017, as reported in the IBP 2017 Survey [48].

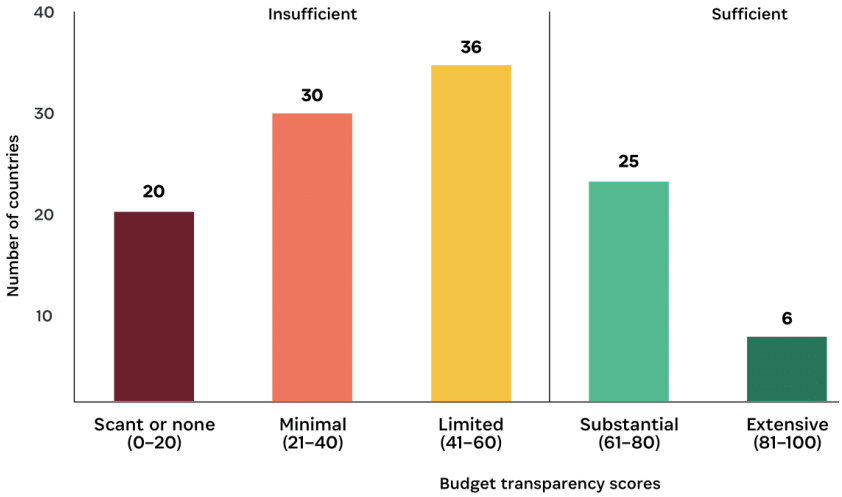

Figure 3.2: The distribution of budget-transparent countries in the period of 2017-2019, as reported in the IBP 2019 Survey [50].

- Publishing fiscal data publicly enable Civil Society Organizations (CSO)[3] to evaluate whether governments are taking into account vulnerable groups' perspective.
- In the Philippines, Budget Partnership Agreements (BPA) allows CSO to be formally involved in budgeting decision. CSOs (both invited and uninvited) can also attend regional/national level public hearings. CSOs monitor programs. Some CSOs also organize consultations, assess national programs, and provide summary accordingly [48]. About 80% of Filipinos are affiliated with CSOs [52].

[3] Civil Society Organizations (CSO) is an organization that is formed by people and is distinct from the state and business. CSO is also a non-profit in nature. CSO can be community-based, but can also be an NGO [51].

OBS Assessments	Number of Comparable Countries	Global Average Score Change
OBS 2008 to 2010	77	3
OBS 2010 to 2012	93	2
OBS 2012 to 2015	100	3
OBS 2015 to 2017	102	-2
OBS 2017 to 2019	115	3

Table 3.1: The global average score changes of budget transparency from different TBS reporting period, as reported in the IBP 2019 Survey [50].

- In Brazil, Public Policy Management Council (PPMC) normally consists of elected officials (50%), elected citizens (25%), and a mix between policy experts, service providers, and union representatives (25%). PPMCs work at different public administration levels: municipal, state, and national levels. The role of PPMC ranges from approving the annual budget, monitoring budget implementation, approving budget item changes during a fiscal year, as well as holding meetings that are open to the public. Financial transfers can be withheld by the federal government to the respective governmental level if the budget is not approved by PPMC [48].

- In South Korea, citizen can register an allegation of government resources waste using a website,[4] a hotline or through reporting centers. This allocation will be investigated and responded within 30 days in the form of a report issued to the person that reports the problem. If confirmed, the person that registers the allegation is awarded 200.000 KRW (USD 175 in 2017). The amount can be bigger (from USD 175 to USD 2.600) if the case considered as the best case and even bigger (up to USD 50.000) when the report saves a large number of government resources. These reports, however, are different from corruption prevention, which deals with using public resources for private gain. The website of the machine-translated reporting portal can be seen in Figure 3.3, enabling citizens to report negative administration, corruption and public interest, abrupt damage, budget waste, and administrative judge request.

- FINA,[5] the standing committee of Finance within Canada's House of Commons, holds pre-budget consultations annually to understand the social-economic changes in Canada. Canadian organizations and citizens can submit short reports which answer specific framing questions from the governments. From the submitted reports, witnesses are chosen and invited for public hearings. Most people submitting these reports are from trade associations, professional NGOs, and lobbies. In the end, the House of Commons publishes reports with recommendations that consider issues from the public hearing [48].

  With the importance of the comprehension regarding fiscal datasets' role for the

[4] as of March 2020, this can be accessed from https://www.epeople.go.kr/index.npaid.
[5] https://www.ourcommons.ca/Committees/en/FINA/About

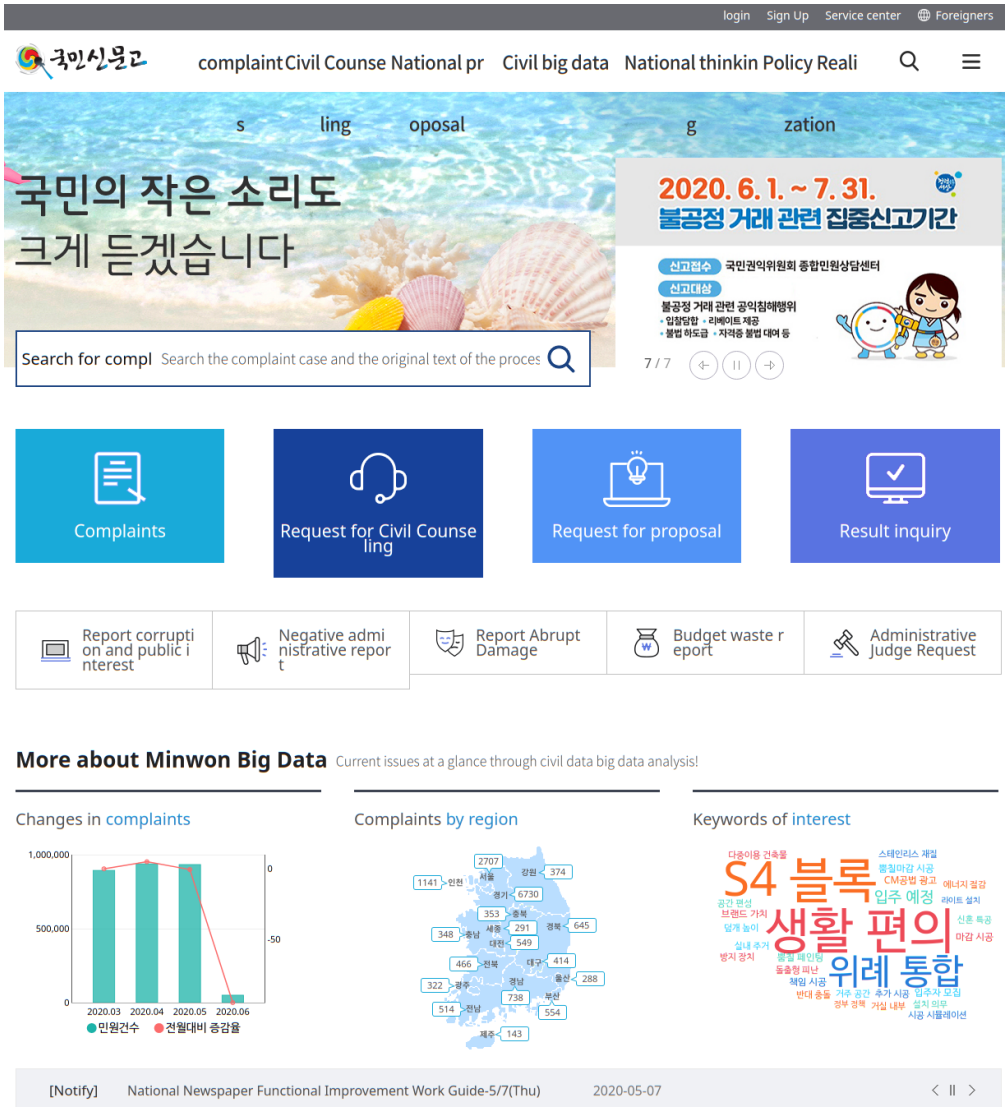

Figure 3.3: The website of machine-translated e-people reporting portal of South Korea (derived from `epeople.go.kr`).

citizen as mentioned above, performing research on the open fiscal data domain is deemed important. The work of GODI [49] provides an overview of how open data are published in general, while the work of IBP [48] asserts how governments worldwide have been, in general, making more efforts to be more transparent in their budget transparency. While the volume of open fiscal datasets is increasing, open fiscal data analysis still faces several challenges, including the quality of fiscal datasets and its heterogeneous nature. Our contributions to these challenges are described in Part II, with related work for the data quality and heterogeneity challenge elaborated in the next subsection.

3.2 Data Quality and Heterogeneity

A major challenge that hinders the cross-datasets analytics on the open fiscal data domain is the quality of the published fiscal data. Several works provide best practices for publishing open data and this is summarized by the work of [53]. As the best practice from summarized recommendations, open data published should be machine readable [54–58], legally open [54–58], bulk downloadable [54, 57, 58], persistent (should remain online with version tracking and archiving feature provided accordingly) [57], free of charge [57], comprehensive [55, 57, 58], in an open format [54, 55, 57], timely [55, 57, 58], available in transaction level [54, 58], available with historical data [57, 58], protecting sensitive information [57, 58], available with associated documents [57, 58], compliant with relevant data standards [23, 54, 59], and provided as initiative to make information accessible for citizens [58]. On a more specific, fiscal-related factors, it should be available with Financial Management Information System (FMIS)[6] description [58] as well as available with off budget fiscal data [57, 58].

A survey of published open fiscal data state has not yet been researched. Therefore, in chapter 4, we provide a survey of important quality factors for publishing open fiscal data, including how far these quality factors are satisfied within surveyed datasets.

Open fiscal data are increasingly being published by different public administrations. With this increasing number of data, comes the question, how is the data published and if there are differences regarding how the data is published. Since there is no binding standard followed by public administrations regarding fiscal data publishing, datasets can be very heterogeneous. In general, the classification of data heterogeneity on relational databases has been done by Kim and Seo [60]. Their work classified and enumerated general structural heterogeneity of relational databases, including schema and data conflicts. There are also heterogeneities in terms of an accounting standard. The attempt of accounting standardization across different public administrations have been made through several initiatives, such as International Public Sector Accounting Standards Board (IPSAS)[7] and European Public Sector Accounting Standard (EPSAS).[8] Since there is no study regarding heterogeneities of open fiscal data, in chapter 5, we provide a thorough analysis of heterogeneities available specifically on open fiscal data.

One of the solutions to ensure data interoperability among heterogeneous datasets is by using a specific data model and format. Open Knowledge Foundation (OKF) works on the *OpenSpending* project.[9] By April 2020, OpenSpending has collected 3.393 datasets from 83 countries with more than 155 million fiscal records. OpenSpending provides an open-source technology stack to manage fiscal data, including Fiscal Data Package (FDP) data model,[10] which is currently being developed by fiscal and transparency communities

[6] https://www.worldbank.org/en/topic/governance/brief/financial-management-information-systems-fmis
[7] https://www.ipsasb.org/
[8] http://ec.europa.eu/eurostat/web/government-finance-statistics/government-accounting
[9] https://openspending.org/
[10] https://specs.frictionlessdata.io/fiscal-data-package/

to model budget and spending datasets. A dataset in FDP consists of CSV and JSON files, with the CSV file as the core fiscal dataset and the JSON file as the dataset metadata. The JSON file also contains dataset column mapping information into a logical model that has been defined by the FDP specification. Once the datasets have been successfully packaged, the datasets can be visualized using the OpenSpending Viewer tool. Further detail regarding the properties of the FDP data model is provided in chapter 7.

FDP does not support semantics. FDP provides a specification to model open fiscal datasets, but only supports a certain structure of open fiscal datasets (see chapter 7). Since there are previously no works that analyze how far FDP supports the heterogeneous characteristics of fiscal data, we provide the analysis of heterogeneity characteristics supported by FDP data model specification in the later chapters, comparing it with a semantic-based data model using OpenBudgets.eu data model/ontology in chapter 7. Additionally, in chapter 8 we describe how we transform the open fiscal datasets into RDF format using OpenBudgets.eu data model.

3.3 Open Data and Open Fiscal Data Platforms

The importance of open data platform leads to works done by researchers and developers. A conceptual architecture for open data architecture is proposed by *DIGO* [61]. DIGO presents a semantic open data architecture based on five layers: knowledge base layer, syntactic data layer, semantic data layer, fusion data layer, and information layer. DIGO elaborates a high-level overview of open data architecture in the general domain of open data but provides neither a concrete architecture nor implementation of the architecture proposed.

OpenSpending[11] (OS) is a platform to analyze open budget and spending datasets. The users can upload and annotate their CSV datasets on the OS platform. The whole platform consists of mainly a data store, API, platform utilities (e.g. conductor, status and incident notifications, command-line interface, authorization client, monitoring tool), data packager, data viewer, data explorer, as well as Where-Does-My-Money-Go app (an app for analyzing and visualizing tax allocation per taxpayer).[12] There is no linked data support in the OS platform and data representation and hence the analytics do not provide the advantages of semantic data integration.

LinkedSpending [62] transforms the datasets available in the OS platform into the semantic format by following DCV specification.[13] There are several components collected, developed and integrated by the LinkedSpending platform, such as ontology, datasets transformation application, data store, error handler, web-based datasets browser and LinkedSpending - OS data synchronization tool. After conversion, the datasets can be browsed using faceted search, visualized using CubeViz[14] or queried using SPARQL. While LinkedSpending has provided the semantic layer and added necessary synchronizer

[11] https://openspending.org/
[12] https://docs.openspending.org/en/latest/developers
[13] https://www.w3.org/TR/vocab-data-cube/
[14] http://cubeviz.aksw.org/

to fetch and convert budget and spending datasets from OS, some other requirements for semantic budgets and spending platforms have not been met, such as the semantic transformation of datasets with unsupported-structure (i.e., non-compatible with OS).

In addition, there are several commercial open data platforms, such as Socrata, Junar, OpenGov, and WikiBudgets. *Socrata*[15] provides an Open Data Portal platform intended specifically for the government on different levels (city, country, state, and federal-state organizations) which includes several services: DataSpace (data storage, indexing, and retrieval), Data Publishing, Data Discovery, and Visualization, and Open Data API. Data supported in the Socrata platform ranges from digital content (e.g., video), operational, geospatial, financial, and performance data. *Junar*[16] is an SaaS platform to publish Open Data in general Open Data domain. Junar offers Open Data collection, enhancement (through tables, charts, and maps), publishing (including API), sharing, and analysis. *OpenGov*[17] offers an open data solution for public administrations, consisting of cloud-based open data publishing, visualization, financial tracking, and collaborative budget builder. *WikiBudgets*[18] is an interactive visualization tool for open budget data. These platforms are mostly commercial and some of them can be either generic whole-solution in terms of the domain (e.g., Socrata, Junar, OpenGov) or very visualization-specific for budget data (e.g., WikiBudgets) without supports for RDF semantics.

Implementation of a semantic, general-domain open data platform in a public administration is done by the open data program for Zaragoza [63]. The city has a long term vision to open up their data as a knowledge graph.[19] The administration provides a large set of open data, and envisioned the release of their open data in a semantic format and whenever possible, in an agreed-upon vocabulary. The knowledge generation flow is illustrated in Figure 3.4, as provided by [63]. The datasets are available ranging in different domains, such as grants, equipment, administrative boards, data catalog, traffic accidents, pollen information, streets in Zaragoza, fuel stations, public parking lot, accommodations, monuments, air quality, city council regulations, job offer, contractor profile, neighborhoods in Zaragoza, public roads, parkland, historic buildings, procedures and services, public services providers, bicycle parking, parking for people with disabilities, city council organization, regulated parking zones, taxi stations, agenda of Zaragoza, restaurants, and motorcycle parking. This build in total, 10 graphs, >300 concepts, almost 600 properties, and >28 million triples. To keep the semantic data updated, there is an automatic update mechanism over the triple store when a change is performed on the data source. Data can then be accessed via APIs. As a result, there are 48 apps registered by the city council as of April 2020.[20] Zaragoza budget execution datasets are not yet made available in their portal, but it is on their road map.

The mentioned platforms come with different features and constraints, as we pointed

[15] https://socrata.com/

[16] http://www.junar.com/index9ed2.html?lang=en

[17] https://opengov.com/

[18] https://www.wikibudgets.org/

[19] While there is no single official definition of knowledge graph (KG), KG according to this paper is a data graph that is intended for knowledge composition.

[20] http://www.zaragoza.es/sede/servicio/aplicacion

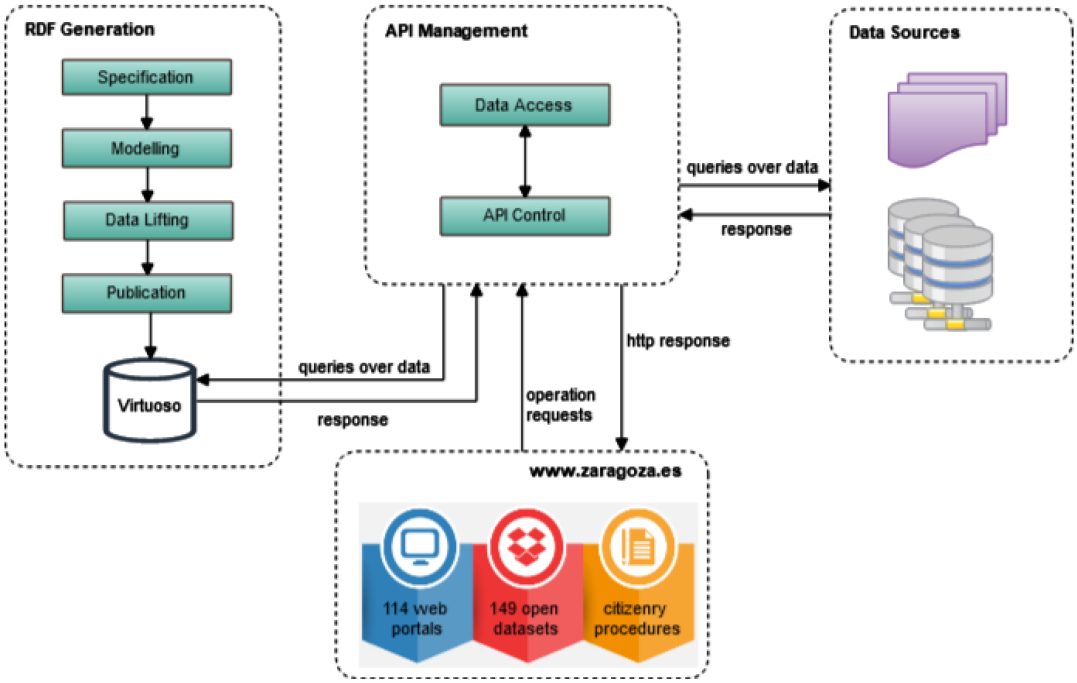

Figure 3.4: The knowledge graph generation flow of semantic open data in the Zaragoza municipality, as provided in [63].

in this platform section. This is a motivation for chapter 9, in which we contribute by proposing a concrete platform architecture for the pipeline of semantic open fiscal data.

Data from different public administrations across different countries are very often published in different languages. The challenges of dealing with multilingualism in linked data are detailed by [64]. This involves how ontology can be localized, how cross-lingual mappings (in conceptual, instance, or linguistic level) can be done, how multilingual lexical information can be represented, and how cross-lingual linked data can be accessed and queried. They also suggest a general architecture for multilingual linked data, by appending additional services (multilingual linked data generation, translation and ontology localization, cross-lingual linking, and cross-lingual access) as well as additional multilingual mappings and linguistic information in addition to the general linked data architecture. Addressing the multilingualism challenge, we design and evaluate a framework to map similar concepts from the fiscal domain in chapter 10. The result of this mapping, along with the previous research contributions, is then used for providing a proof of concept for federated comparative analysis of open fiscal data as described in chapter 11.

Part II

The Current Open Fiscal Data Ecosystem

CHAPTER 4

Current State of Open Fiscal Data in Public Administrations

Due to the decentralized nature of datasets publication, the quality of the datasets is varied. There have been unofficial recommendations and standards on publishing open data in general, provided by civic societies, NGOs, and open data enthusiasts. It is, however, unclear to what extent do the data publishers follow these standards, and whether the open fiscal data domain can benefit from adding additional quality factors. This chapter clarifies these issues, in addition to providing an expanded collection of open fiscal data quality factors.

This chapter is based on the following publication:

- **Fathoni A. Musyaffa**, Christiane Engels, Maria-Esther Vidal, Fabrizio Orlandi, Sören Auer, *Experience: Open Fiscal Datasets, Common Issues, and Recommendations*. ACM Journal of Data and Information Quality, 2018.

4.1 Existing Standards

The challenge of data interoperability across different dataset publishers is not particularly new. From the company and business perspective, XBRL (eXtensible Business Reporting Language) format has been used as a standard for business information exchange. XBRL allows the representation of standardized accounting processes. Authorities play an important role in the XBRL adoption, as it became mandatory in 2009 for the top 500 U.S. companies to report in XBRL [65]. Within the open fiscal data domain, such matured standard for datasets publishing is not developed due to lack of a binding order from authorizing bodies, since the standards of open fiscal publication normally come from the grassroots communities and NGOs. Moreover, differences in open fiscal data accounting processes and classification hierarchies across different public administrations complicate datasets standardization. Such differences limit the usefulness of financial disclosures [66, 67].

There are few works that assess open fiscal datasets based on common quality factors that should be present in open fiscal. The Open Data Monitor project[1] reports open fiscal implementation across Europe. An assessment in the general domain was reported by GODI [49] and ODB [68]. GODI provides country rankings based on nine GODI factors and the availability of 13 different open fiscal domains in each country, including budget and spending. ODB provides open fiscal analysis and ranking based on open fiscal initiative readiness, program implementation, and impact on business, civil society, and politics. Peters et al. [69] asses open fiscal data and portal quality specifically for ESIF funding in EU countries. Currently, several open fiscal publishing guidelines for the general open fiscal domain exist, including [70], [71], and [57]. The Open Data Handbook [72] provides a guide for the legal, social, and technical aspects of open fiscal. The 5-star data schema [27] is well-known among Linked Open Data communities. It is also worth mentioning a survey by [73] for Linked Data quality assessment. A specific guide to publishing open fiscal as Linked Data is provided by [32]. [7] mention lessons learned from data.gov.uk implementation. Data Management Maturity (DMM) Model provides *Capability and Maturity Levels* [74], which has six different data management process areas: data management strategy, data governance, data quality, platform and architecture, data operations, and supporting processes. Data quality process area is directly affecting data management strategy, data governance, platform and architecture, and data operations, which is why the data quality process area is important. Data quality process area is composed further of data quality strategy, data profiling, data quality assessment, and data cleansing. The OFDP Framework provides a more elaborate framework of *data quality* for fiscal data within the sub-process area of data quality strategy, data profiling, and data assessment.

Fung et al. [75] mention that a sustainable transparency system improves on three important dimensions over time: expanding information scope, increasing information accuracy and quality, and increasing the use of information. Compared to other works in data quality assessment, we propose a framework, OFDP, which aims to improve these dimensions within fiscal data. Our assessment of surveyed fiscal datasets exemplifies the heterogeneity issues as mentioned by [66]. In comparison with related work regarding datasets assessment, we aim to assess fiscal datasets on multiple public administration levels and provide guidelines accordingly. We believe that the dimensions mentioned by [75] will be improved if the dataset publishers comply with our guidelines, as can be seen in section 4.5.

4.2 Methodology

Our methodology to analyze open fiscal data using the OFDP Framework is summarized in Figure 4.1. We gather links to the fiscal datasets from the OpenSpending community,[2] which is an active community which aims to track and analyze global public financial

[1] http://project.opendatamonitor.eu/
[2] https://openspending.org/

Figure 4.1: Methodology for obtaining the proposed OFDP framework and guidelines.

information. The community members submit extensive links via Github.[3] These links are then used to obtain the actual datasets for our assessment. In addition, we also explore and add additional links outside the submitted links. As for the main framework, we study the literature to acquire common motivations for publishing open fiscal. Later, we gather quality factors that support these motivations and then measured the weights of each quality factor. To achieve a more objective weighing of the factors, we collect fiscal communities' views through a questionnaire[4] (also provided in Appendix A), which is distributed in several fiscal communities (OpenSpending, Follow the Money, OpenBudgets.eu, IODC 2016) and government officials. We collect 24 responses from this questionnaire and use the median as the weight for our identified quality factors. The collected and assessed data are then ranked using three methods: OFDP, ODB, and GODI. We evaluated the ranking results using Spearman's coefficient. Finally, we highlight the deviation between the OFDP framework and assessment results in the form of a guideline for fiscal data publishers.

4.3 The OFDP Framework

We identify a set of comprehensive quality factors which are presented in previous works [27, 49, 57, 68, 73, 76, 77]. We also present additional quality factors from our experience in processing open fiscal data.

The readers are referred to the documents by GODI [1] and ODB [68] for the explanation of factors that are originating from the respective documents (see Figure 4.2), as well as the open data guide [72, 78]. Some of the quality factors are ambiguous or not self-explanatory, such as the availability of semantics, data being sustainable, timely, permanence,, and open format. The *availability of semantics*, i.e., data availability in a semantic format such as RDF, facilitates data integration and concept linking between different datasets. A dataset publication is *sustainable* whenever it is hosted on a government open fiscal portal, an official website, or a preservable public platform (e.g., Github). *Timeliness* relates to how soon the data are published by government officials after the data have been collected. This is especially relevant for time-sensitive data. *Permanence* is concerned with getting information over time, which ideally provides an archival feature and *version tracking*. *Open Format* refers to any file format that is published publicly, free of charge, and without reuse limitations, so anyone can read and implement the format without intellectual property constraints [79] (e.g., CSV format). A *code list* is a set of enumerated concepts that restricts the possible values of a field, e.g., currency or country code.

[3] https://github.com/os-data/registry
[4] http://bit.ly/open-fiscal-data-survey

Table 4.1: Our OFDP quality factors and their weights according to the survey result.

FACTOR	WEIGHT	FACTOR	WEIGHT	FACTOR	WEIGHT
Data Existence*	5	Authoritative	4	Regular Update	4
Easily Available*	5	Complete Code List	4	Search Mechanism	4
Documentation	5	Contact Point	4	Sustainable Publication	4
Free of charge	5	Dataset Filtering	4	Up to date	4
In Digital Form	5	English Info Available	4	Version Tracking	4
Mentioned License	5	In Bulk	4	Mentioned Contributors	3
Online	5	Metadata	4	RDF Availability	3
Public	5	Open format	4	Visualization	3
Structured data	5	Open License	4	Dereferenceable LD URI*	3
API Availability	4	Persistent URI	4		

Some studied quality factors are excluded from OFDP because it is non-trivial to measure these factors in a dataset. We are constrained by several quality factors that are not included in our assessment, such as *granularity*, *accuracy*, and *completeness*. Each of those factors requires a fine-grained definition regarding the level of granularity/accuracy/completeness to make the assessment of the datasets within these factors objective. In practice, the granularity levels of these factors are very diverse across different budgets and spending datasets publishers. The quality factor *primacy* [57] is implicitly provided by a composition of three quality factors: *authoritative*, *mentioned contributors*, and *version tracking*.

Three of the quality factors are excluded from the questionnaire (*easily available* and *data existence* due to their obvious importance, and *dereferenceable linked data URI* since it is overly technical for people outside the linked data community). For these factors, we assign the weight manually. The term *easily available* refers to how easy it is to obtain the full datasets that contain all of the complementary information without investing a significant amount of time. Being *easily available* and having the *data exist* are very important. Easily available determines how the data can be found for further consumption by interested entities.

Overall, we collect 29 quality factors (see Figure 4.2). Quality factors in OFDP subsume all quality factors in GODI (up to May 2017) and ODB (2013). Subsequently, we weight the quality factors for the OFDP framework according to the questionnaire result (described in section 4.2). The quality factors and their weights are provided in Table 4.1.

4.4 Evaluation

The detailed analysis of the datasets is available publicly in an online spreadsheet.[5] The overview table of the surveyed datasets is summarized in Table B.1 of Appendix B, and due to the lack of space in this thesis, we recommend the readers to also see the comprehensive online spreadsheet. This spreadsheet includes links, full assessment, total assessment score for each dataset, and additional contexts (e.g., geographical area, data model, coverage, domain, granularity, and comments). We outlined the analysis in Figure 4.3, Figure 4.4, and Figure 4.5.

Figure 4.3 provides an overview of the star-data categorization rating [27] of the datasets. Due to restricted license or license unclarity in many of the datasets, a major percentage (72.7%) of the dataset is listed as zero-star. This means that at least one of the necessary permissions required in Open License (access, use, modify, and redistribute) is not clearly mentioned. There are 2.6% of assessed datasets that were categorized as one-star data, none as two-star data, 20.8% as three-star data, 3.9% as four-star data, and none as five-star data. As a side note, two-star data requires the data to be published on the web with an open license, in a structured but proprietary format. In our analysis, several datasets are published in Excel format, which previously was a proprietary format. However, Microsoft has published the Excel file format specification openly so that this format can be implemented by anyone.

The percentage of each quality factor's presence in the datasets is shown in Figure 4.4. Figure 4.5 plots the resulting score for each dataset using GODI, ODB, and OFDP methodologies. The scores of these methodologies are normalized and therefore range from 1-100. We rank the datasets according to these scores. Based on the ranking results, Spearman correlation values are computed. Value of 0.86 between ODB-GODI shows that both rankings are correlated. The values between ODB-OFDP (0.78) and GODI-OFDP (0.75) show a lower correlation as our newly developed OFDP takes more comprehensive quality factors into consideration (see Figure 4.2). The detailed correlation calculation is available in the dataset analysis spreadsheet.

4.5 OFDP Guidelines to Publish Fiscal Data

As a result of our assessment, we recommend that open fiscal data publishers follow quality factors listed in Table 4.1. A higher weight indicates a higher priority for the quality factor. We find that most analyzed datasets have performed well for being available online, free of charge, public, in digital form, easily available, in an open format, published sustainably, published with the contact point information, authoritative, and also published with search mechanism provided. However, many quality factors need more attention from fiscal data publishers, which we provide in the following section in alphabetical order.

API Availability. Publishing datasets as an API is useful if the data are large and frequently change, especially when only a small portion of the data is needed [80]. Only

[5] http://bit.ly/jdiq-datasheet-view

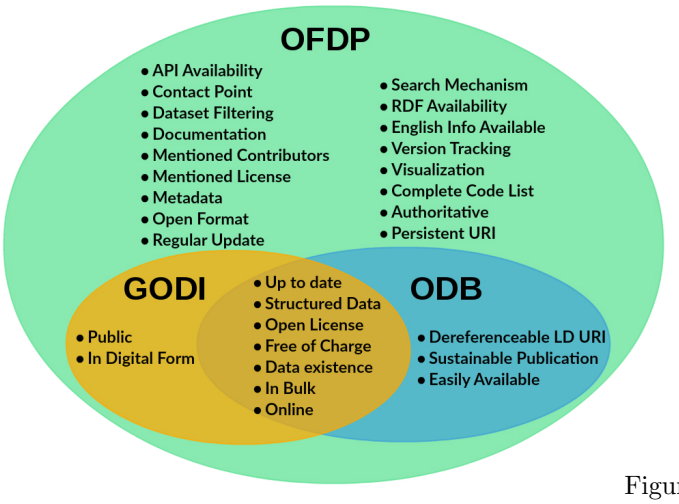

Figure 4.2: Quality factors considered in our OFDP framework which subsume factors from GODI and ODB.

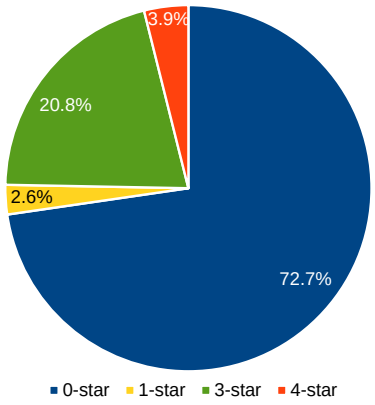

Figure 4.3: Datasets categorization according to the 5-stars data schema by Sir Tim Berners-Lee.

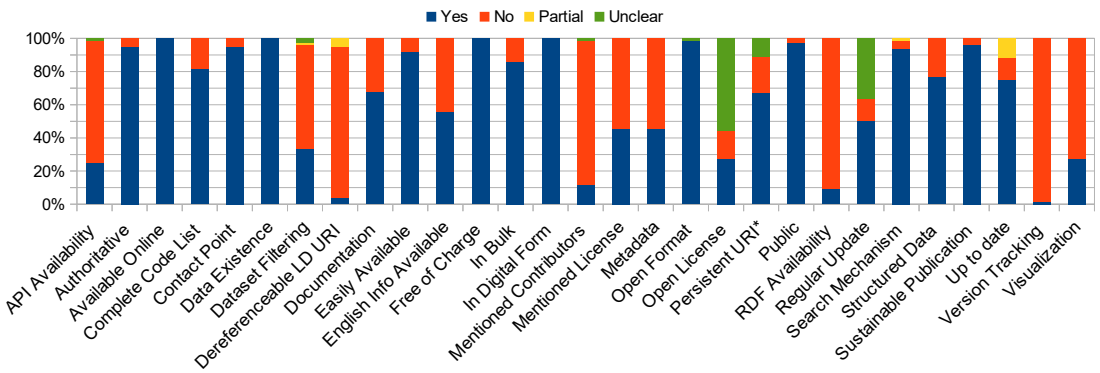

Figure 4.4: Open fiscal data quality factor presence in our sample of 77 open fiscal datasets.

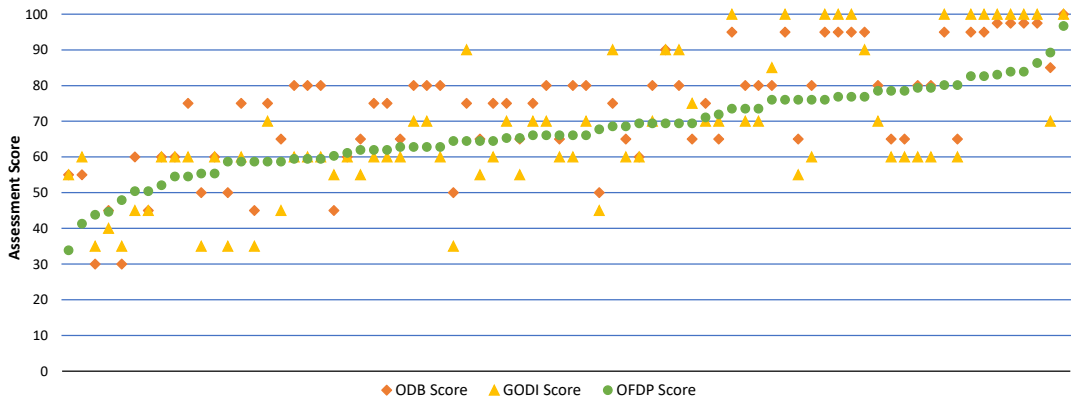

Figure 4.5: The proposed OFDP score compared to ODB and GODI.

24.7% of the datasets publishers from our analysis provide an API endpoint for their dataset. Some data publishing Content Management Systems (e.g., CKAN, DKAN) provide an API endpoint feature. Publishing datasets via an API endpoint should ideally be accompanied by publishing datasets for bulk downloads, too.

Complete Code list. Code lists can be used to link concepts among different datasets and enable comparative analysis between different fiscal dataset sources. However, 18.2% of the datasets are not published with complete code lists. We encourage open fiscal dataset publishers to provide full code lists in a structured format (instead of legal, textual documents). The most popular way is to publish the code lists within the main dataset itself. Based on our experience, the most efficient way of providing code lists is by including a list of the code and the description of each code in a separate file. Therefore, codes and information (e.g., label, descriptions) are covered without redundancy on the main fiscal data itself, and no manual code list extraction effort is necessary.

Dataset filtering. The dataset filtering feature is recommended as it eases the users if a particular selection over the dataset is required. This feature is essential for understanding and analyzing the data by giving a specific selection criterion. Among the analyzed datasets, 62.3% of data publishers do not provide this feature.

Documentation. For 32.5% of the analyzed datasets, no documentation can be found, which hinders the understanding of the datasets. Meaningful documentation should be provided and shall consist of at least the datasets content, datasets context, available classifications, and the definition of fields present in the datasets. The OpenCoesione[6] initiative provides a good example for documentation.

English Info Availability. In assessed datasets, 44.2% are published without English documentation. Machine translation is prone to errors especially for classifications and specific terms. We recommend the dataset maintainer to provide at least English documentation, especially for international communities who analyze the data.

In Bulk. Publishing open fiscal in bulk (e.g., CSV, instead of only as API endpoint) is important because it is a familiar format for non-programmers, easy to mirror, produce, host, and distribute [80]. Datasets not published in bulk are not *open* according to the open definition[7] as they are not provided as a whole. For technical practicality and openness reasons, the bulk availability should always be considered while publishing open fiscal datasets. Around 14.3% of the analyzed datasets lack this feature.

Mentioned license. The dataset's license should be mentioned explicitly, stating all basic permissions for public access, usage, modification, and sharing. *Restricted* or *open* license clarity is counted only on 45.4% of the surveyed dataset (as can be seen in Figure 4.4). We recommend the datasets' publications with a commonly known license type.

Metadata. Although metadata helps data acquisition and identification, 54.5% of the datasets are not published with metadata. There are three kinds of metadata: descriptive, structural, and administrative metadata [81]. Descriptive metadata explain datasets' discovery and identification, such as title, abstract, author, and keywords. Structural

[6] http://www.opencoesione.gov.it/opendata/#fs0713-title
[7] http://opendefinition.org/od/2.1/en/

metadata describes the arrangement of objects within the data, e.g., table of contents and chapters. Administrative metadata indicates resource management, e.g., technical information, how and when the data was created, intellectual property rights, and archival information. Whenever possible, we recommend providing all these types of metadata following, for example, the W3C recommendation DCAT [82], or it's adaption DCAT-AP by the EU Committee.

Open License. An open license allows data users to access, use, modify, and redistribute the data. This is essential to enable and foster data reuse for analysis purposes. Open definition enlists open conformant licenses which we recommend. In assessed datasets, 27.3% are openly-licensed, 16.9% are restricted, and 55.8% are unclear.

Persistent URI. Maintaining permanent links to datasets is recommended, and at least a redirection mechanism should be provided from the original link once the link has changed. Persistent URI is a relevant concern, as 21.8% of the analyzed datasets are no longer accessible under the previously-valid URI (see Figure 4.4). In addition, a human-readable URI is preferred for the datasets to improve search engine optimization.

Regular Update. The regular update provides an expectation of when interested stakeholders can find the latest dataset. Most of our questionnaire respondents agree that regular update is an important quality factor in publishing open fiscal data. We also recommend that the dataset's publisher publish their datasets regularly. Unfortunately, only 50.6% of the datasets provide regular updates, while 13.0% of the datasets do not provide regular updates and the other 36.4% are unclear.

Up to Date. We encourage the dataset publishers to provide the latest information so that the dataset's analysis process can be more interesting to do for the stakeholders and journalists. During our analysis (which was done in 2016), we categorize any budget datasets up to 2016 and spending datasets up to 2015 as up to date datasets. From the analyzed datasets, 11.7% are partially up to date, and 13% are not up to date.

Structured Data. Even though publishing structured data allows users to analyze the data easily and maximizes the technical access, 23.4% of the assessed datasets are published in a non-structured format (e.g., PDF). The importance of publishing structured data has been highlighted in previous works [70, 71, 78, 83]. Publishing the dataset in a non-structured format makes the transformation process more difficult as no specific pattern can be followed by tools performing datasets transformation. Hence, we highly recommend publishing datasets in a structured data format.

Version Tracking. Version tracking or version control for data supports distributed data contribution, collaboration, broader participation, provenance tracking, and incremental development [84]. We encourage the publishers to provide version tracking to see the changes made and the user who changed the datasets. In our analysis, 98.7% of dataset publishers do not provide a version tracking feature on their datasets' web page. CKAN features basic activity monitoring on published datasets, specifying modified data, and the user involved.

CHAPTER 5

Managing Heterogeneities of Open Fiscal Data

Open data has gained momentum during the past few years, but not much analytics has been performed over published open budget and spending datasets. Many challenges to consume open budget and spending data are still open. One of the challenges is the heterogeneity of these datasets. We analyze more than 75 different budgets and spending datasets released by different public administrations from various levels of administrations and locations. We select five datasets from those analyzed datasets to illustrate and represent several types of budget and spending heterogeneities found on the analyzed datasets.

This chapter is based on the following publication:

- **Fathoni A. Musyaffa**, Fabrizio Orlandi, Maria-Esther Vidal, Hajira Jabeen. *Classifying Data Heterogeneity within Budget and Spending Open Data.* International Conference on Theory and Practice of Electronic Governance (ICEGOV) 2018. Galway, Ireland.

5.1 Heterogeneities on Fiscal Data

Many public administrators have published budget and spending data as part of their open data program. A survey conducted by Open Knowledge Foundation shows that budget datasets topped the first rank as the most published open datasets, among other types of datasets (e.g., national statistics, procurement, national laws, administrative boundaries, draft legislation, air quality, national maps, weather forecast, company register, election results, locations, water quality, government spending, and land ownership) [3]. Having a flexible way to publish a dataset simplifies the work of dataset publishers. Unfortunately, this flexibility leads to datasets complexity, which makes the datasets difficult to consume and integrate. In addition, the published fiscal data requires highly technical skills to analyze [15].

Publishing open data in the domain of budget and spending is often accompanied by different types of *classifications*, as briefly elaborated in section 2.1. The structures of these classifications are also heterogeneous. The diversity ranges from the level of details (i.e., the availability of hierarchies available within the list) as well as how the classifications are normalized or attached (e.g., within the dataset or outside the dataset). Among the factors that contribute to these heterogeneities are the difference of business and budgeting process, the coverage level of the administration (e.g., supra-national vs. municipal) or how projects within the public administration are funded.

5.2 Motivating Example

Two datasets published by different public administrations from different coverage levels are provided in a different structure. Both datasets contain different coverage levels and details, along with different representations, which can be categorized by the *content*, *structure* and *syntax* perspective. Table 5.1 illustrates the heterogeneities between these two datasets.

Figure 5.1 (a) illustrates a sample row taken from the City of Madrid's income budget 2017 dataset. Figure 5.1 (b) provides an example of a row taken from the City of Bonn's budget 2017 dataset. Both datasets are published in their native languages (Spanish and German, respectively), and structured differently. The datasets from the city of Madrid include the description of each classification (*Descripcion Centro* describes *Centro, Descripcion Capitulo* describes *Capitulo,* and *Descripcion Economico* describes *Economico*) within the dataset itself. In contrast, the dataset from the City of Bonn does not directly provide the description of the classification (*Profitcenter, Konto, PSP element, Auftrag, Geschäftsbereich,* and *Version*). Additionally, Bonn datasets are not split into different operational character categories (e.g., *income budget* vs. *expenditure budget*), while the Madrid dataset split the datasets into different operational categories. The operational character category in Bonn dataset is provided implicitly via the code in the *Konto* classification as well as the sign in the amount of money indicated (minus sign for income, positive numbers for expenditure).

Despite the difference, some information between these datasets are relatable, as indicated in Figure 5.1 (c). For example, the amount of income is provided in the *PrCtrHw* column in Bonn datasets and in the *Importe* column for the Madrid dataset. *Konto* in Bonn dataset consists of *operational character* classification and *economic classification*. In Madrid dataset, *economic classification* is provided as *Economico*. *Profitcenter* in Bonn dataset merges *administrative classification* and *functional classification*. In Madrid dataset, the *administrative classification* and *functional classification* are provided as *Centro* and *Capitulo*, respectively.

We have conducted a comprehensive analysis of 77 heterogeneous budgets and spending datasets. The spreadsheet of the detailed analysis is available online.[1] These datasets come from different levels (supranational, national, regional, and municipalities). Among

[1] http://bit.ly/jdiq-datasheet-view

Table 5.1: Illustrations of heterogeneities between two datasets.

No	Heterogeneity	Dataset A	Dataset B
1	CONTENT		
1.1	Measure		
1.1.1	Observation granularity	Transaction	Aggregation
1.1.2	Funding source	Single-source funding	Multiple source funding
1.1.3	Numerical representation	Only positive values	Positive and negative values
1.1.4	Currency	EUR	EUR, GBP
1.1.5	Time granularity	Date	Year
1,2	Classifications		
1.2.1	Insignificant Classification Hierarchies	Unavailable	Available
1.2.2	Number of Available classifications	Functional, administrative	Functional, economic
1.2.3	Classification structure	Non-hierarchical	Hierarchical
1.2.4	Publication interval	Annual	Once, with occasional updates
1.2.5	Harmonizing / Standardizing office	Municipally harmonized	Nationally- and EU-harmonized
1.2.6	Classification Necessity	Only mandatory classifications available	Mandatory and optional classifications available
1,3	Availability		
1.3.1	Budget phases	Drafted, Proposed, Approved	Approved, Executed
1.3.2	Observation description	Available	Unavailable
1.3.3	Metadata availability	Unavailable	Available
1.3.4	Budget direction / operation character	Income and expenditure	Expenditure
2	STRUCTURE		
2,1	Table Normalization		
2.1.1	Budget phase attachment	Within the dataset	Different dataset
2.1.2	Operation attachment	In similar dataset	In different dataset
2.1.3	Classification attachment	Within the dataset	Different dataset
2,2	Classification structure		
2.2.1	Classification notation	Plain label	Encoded
2.2.2	Abbreviated Classification	Abbreviated	Non-abbreviated
3	SYNTAX		
3,1	File format	CSV	Excel
3,2	Character encoding	ISO-8859-3	UTF-8
3,3	Metadata	-	DCAT [7]

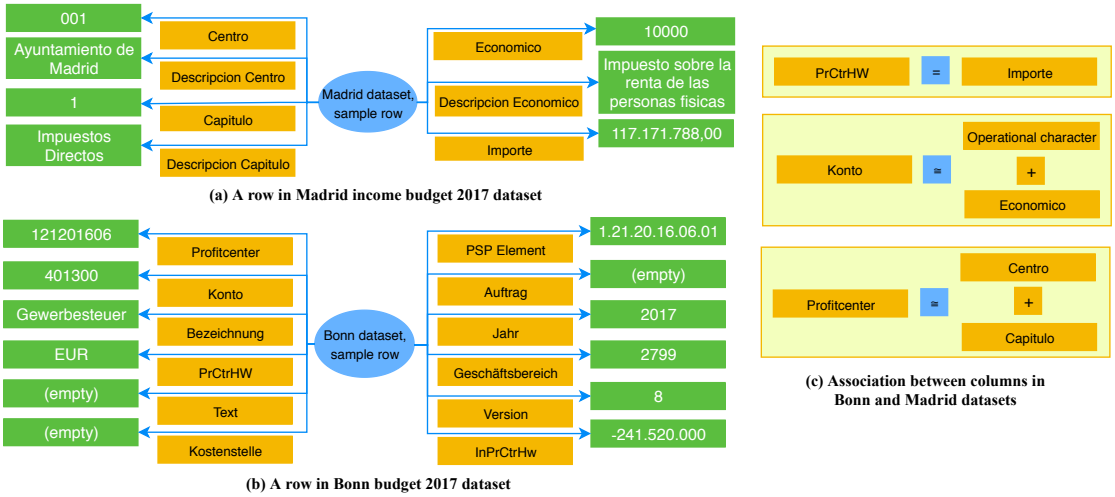

Figure 5.1: (a). Madrid datasets consists of seven columns including code description. (b). Bonn datasets consists of 11 columns with code not directly described. (c) Mapping across related columns between Bonn and Madrid dataset.

those analyzed datasets, we picked the following five datasets, which represent a good sample of possible heterogeneities on open fiscal datasets within budget and spending domain. These datasets are:

- *Bonn budget datasets* (from a private repository).[2] The Bonn datasets are currently obtained privately but licensed as Public Domain. These datasets contain budget data from 2008 - 2024, along with several classifications that published once yet valid for years, with occasional updates. Bonn budget datasets have likely similar structure with most of the budget datasets from the cities within German state North Rhine Westphalia.

- *Aragon budget datasets.*[3] The Aragon budget datasets contain budget data of Aragon autonomous community from 2006 - 2017.

- *ESIF 2014 – 2022 financing plan datasets.*[4] This dataset contains financing plan for European Structural and Investment Funds (ESIF) which covers the financing details across EU member states for the year of 2014 - 2020.

- *Madrid 2017 budget datasets.*[5] This dataset covers the budget from the city of Madrid for the year 2017. The budget covers investment, spending, and income.

[2] https://goo.gl/BTxmNp
[3] https://opendata.aragon.es/datos/catalogo?texto=presupuestos
[4] https://bit.ly/esif-2014-2020
[5] https://bit.ly/madrid-budget-data

- *Swedish national project fund dataset.*[6] This dataset contains project funding in Sweden.

5.3 Heterogeneity Types

This subsection enumerates several types of heterogeneity illustrated with cases from datasets mentioned in section 5.2. Among these datasets, we enumerate several heterogeneities that also likely to occur over other datasets from different public administrations.

1. Content. The hierarchical content-related heterogeneities are summarized in Figure 5.2, which categorized within measure, classifications, and availability perspective.

   1.1. Measure

   1.1.1. *Observation Granularity.* Datasets that list paid beneficiaries are mostly granular/transactional. Datasets that are published based on the budget cycle are mostly aggregated. All the datasets listed in section 4.2 are aggregated.

   1.1.2. *Funding source*, or the availability of co-funding information. Some datasets contain co-funding information if the funding involves different administrations. For example, ESIF planned funding has several measure columns that separate the amount funded by the European Union or the amount funded by its own member state's administration.

   1.1.3. *Numerical representation* of the amount value. Some datasets provide negative and positive values for the amount measures, for example, Bonn datasets. In Bonn datasets, a negative sign interpreted as revenue, while a positive sign indicates expenditure. In case a dataset has both positive and negative signs, interpreting the meaning of these signs should be done carefully by consulting domain experts from the public administration which publishes the data, or by referring to datasets documentation if the documentation is available.

   1.1.4. Currency. The currency used on budget and spending datasets depends on the origin of the public administration. Some datasets are also provided with multiple currencies, such as Swedish EU Structural Fund projects.

   1.1.5. *Time* granularity. Most datasets are released annually, hence the information is granular per year (e.g., Bonn and Aragon datasets). Other public administration may release the budget information per budgeting period that is not annual. ESIF datasets, for example, is implemented based on a seven years' period of EU regional policy framework. Hence, the ESIF budget datasets are released for the budgeting period of 2014-2020.

[6] http://projektbank.tillvaxtverket.se/projektbanken

1.2. Classifications

1.2.1. *Dependent Classification or Insignificant Classification Hierarchies.* Bonn datasets, for example, have dependent classifications. There are at least five classifications within Bonn datasets, and four of them have dependency relations. The *internal orders* (or *Auftrag*), *project structure plan* (or *PSP-Element*), and *cost center* (or *Kostenstelle*) are all dependent on the *profit center* classification. The dependent classification may not be in the same classification type. This classification dependency comes from the public administration's requirement. Other datasets, such as Aragon budget datasets, do not have such dependent classifications.

1.2.2. Numbers *of available classifications*, as well as the types of the classification. The types of classifications from a dataset varies from one to another. For example, Aragon budget datasets contain income dataset which has four types of classifications: administrative, functional, economic, and financial classifications. Bonn datasets also contain administrative and functional classifications. However, there are more classifications provided in Bonn datasets, and these classifications are not necessarily relatable to classifications published by other datasets publishers, such as business area (*Geschäftsbereich*) and internal order (*Auftrag*) for accounting purposes.

1.2.3. *Classification structure.* Some items in the classifications on the datasets have a hierarchy. For example, the Aragon budget dataset's functional classification has a four-level hierarchy. On the other hand, the classification within the Swedish national project fund dataset does not have an explicit hierarchy.

1.2.4. *Classification publication interval.* Some public administrations publish their classifications once with occasional updates (such as Bonn datasets), while some other datasets publishers publish classifications each year, such as Aragon budget datasets.

1.2.5. *Harmonizing/standardizing office* of the classifications. Some classifications are provided in a distributed manner. Such a case is illustrated by ESIF 2014 – 2020 financing plan, in which *national priority* is created by different EU member states. However, no additional classifications document that explains each item within *national priority* can be obtained. In this case, a non-harmonized classification exists. In other datasets, such non-harmonized classifications are not found.

1.2.6. *Classification Necessity.* Optional classifications refer to an additional classification which unnecessarily available in each row (i.e., observation) in the datasets. Bonn datasets have several optional classifications, while other datasets above do not have optional classifications. The information regarding classification necessity could be important if the datasets are about to be transformed into a data cube-based data model, such as DCV (see chapter 2).

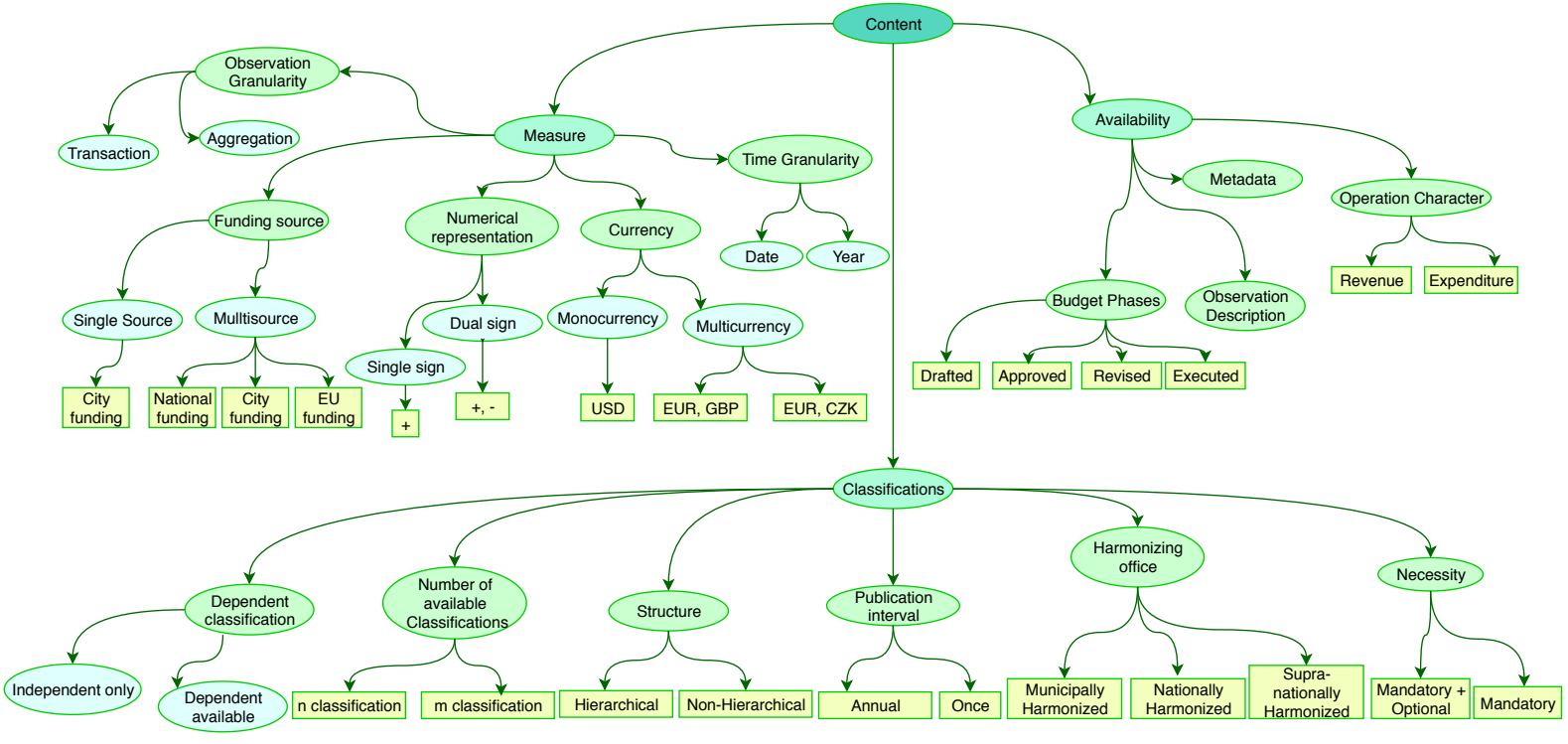

Figure 5.2: Budget and spending dataset heterogeneity hierarchy from the perspective of content.

1.3. Availability

1.3.1. *Budget phases*. There are at least four different budget phases: drafted, approved, revised, and executed. Not all these stages are usually provided, for example, Bonn currently provides drafted, approved, and executed budget phase while Aragon provides approved and executed budget phase.

1.3.2. *Observation description*. Aragon expenditure datasets and Bonn datasets are provided with a description within each row. Swedish and Madrid datasets do not provide such description for each observation.

1.3.3. *Budget direction* or operation character. Some datasets provide income and expenditure information, for example, the datasets of Aragon, Madrid, and Bonn.

1.3.4. *Metadata*. Some datasets are provided with metadata, such as ESIF and Aragon datasets. On the other hand, Bonn datasets, for example, is not provided with metadata.

2. Structure. The hierarchical structural heterogeneity is illustrated in Figure 5.3, namely:

2.1. Table *Normalization*

2.1.1. *Budget phase attachment*. Madrid executed budget datasets,[7] for example, provides drafted and approved amounts within the same file. Other datasets, such as Aragon and Bonn datasets, provided other versions of budgeting data in different files.

2.1.2. *Operation character attachment*. Income and expenditure data can be provided separately (e.g., Aragon and Madrid datasets) or in the same datasets (e.g., Bonn datasets).

2.1.3. *Classification attachment*. Some datasets provide the classifications labels within the same file, such as ESIF 2014 - 2022 financing plan datasets. Other datasets, such as Bonn budget datasets, provide the classification label outside the main dataset's file.

2.2. *Classification structure*

2.2.1. *Classification notation*. Some datasets encode the concepts within their classifications in unique notations, such as Bonn budget datasets. Others do not encode their classification labels into unique notations, such as the Swedish datasets.

2.2.2. *Abbreviated classification* label. Some datasets providers are limited by the systems they are using, which result in field-length limitation. Bonn datasets classifications have such limitations on their datasets. The abbreviation can be a problem if a further effort to analyze the datasets

[7] https://goo.gl/naqgv8

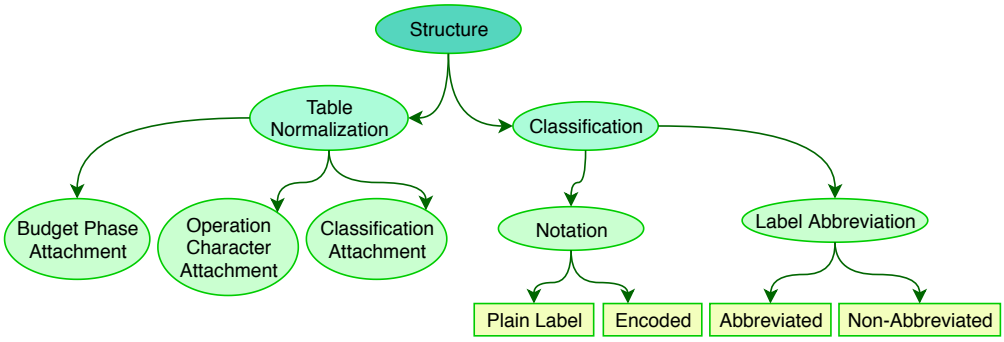

Figure 5.3: Budget and spending dataset heterogeneity hierarchy from the perspective of the structure.

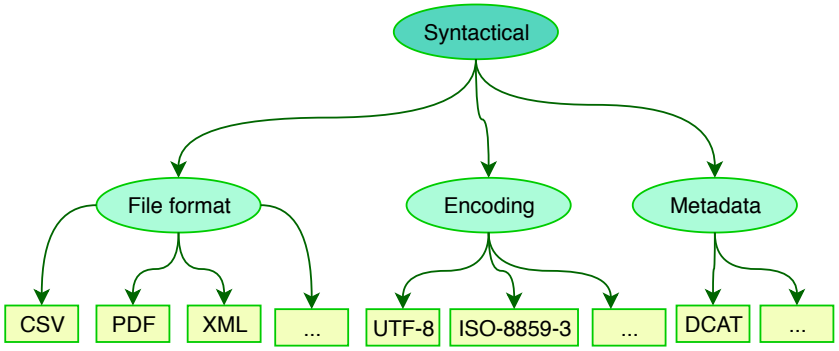

Figure 5.4: Budget and spending dataset heterogeneity hierarchy from the perspective of the syntax.

involves techniques such as word embedding, machine translation, or natural language processing. Fortunately, other datasets mentioned above do not contain abbreviated labels.

3. Syntax. The hierarchical syntax-related heterogeneity is illustrated in Figure 5.4, namely:

   3.1. *File Format.* The released file format can be different across public administrations. Most of the datasets are provided in tabular format (Excel, CSV, or both, such as Bonn, Aragon, and Madrid datasets). Some other releases datasets in another form, such as the HTML page for the Swedish dataset.

   3.2. *Character Encoding.* Even though datasets are published in the same file format, the character encoding may differ. The encoding information is often missing but can be guessed based on the originating geographical origin of the public administration by inferring to the ISO 8859 standard [85].

   3.3. *Metadata.* Different public administrations may provide a different type of metadata. For example, Aragon datasets are provided with DCAT metadata [82], while Bonn dataset is provided without metadata.

Concluding Remarks for Part II: The Current Open Fiscal Data Ecosystem

In this part of the thesis, we present the state of datasets. In particular, we focused on the following research question:

> **RQ1**: What are the requirements for publishing high quality open fiscal data?

The importance of open data quality leads to several open data publication recommendations and guidelines. However, even though such recommendations exist, the data published do not always satisfy the requirements in these recommendations. Moreover, as we performed the survey of the open fiscal datasets, we did not find any open data publishing guidelines specifically aimed for the fiscal data domain. Therefore, we cover this problem as an initial research question.

We began in chapter 4 in which we described our experience with open fiscal datasets and analyzed their quality under different aspects. In particular, we achieved the following goals: *(i)* identify several important factors impacting quality and reuse of open fiscal datasets, *(ii)* evaluate these factors' relevance, and *(iii)* assess the presence of these factors in recent open fiscal datasets. The assessment was performed on a representative number of datasets from different public administrations by thoroughly analyzing 77 datasets from different public administrations. In addition, we compared our assessment results with previous existing assessment frameworks. Our OFDP assessment framework considers a larger and more fine-grained set of quality factors, specifically targeted at fiscal datasets. Several qualitative issues of open fiscal datasets have been raised within our analysis. Hence, we highlighted these issues and provided guidelines for publishers of open fiscal data.

Further, we discussed the heterogeneities on open fiscal data which discussed in the second research question:

> **RQ2**: What types of data heterogeneity problems occur with open fiscal data?

Given the independent nature of open fiscal data publication, it often happens that the datasets are published in a heterogeneous nature. To understand better the heterogeneities on open fiscal datasets, we perform a follow up of our initial fiscal datasets quality analysis. In chapter 5, we presented a list of heterogeneities that appear in open fiscal datasets, thoroughly categorizing it in a hierarchical manner. The heterogeneities are collected after analyzing different datasets from different public administrations.

Part III

Data Management and Analytics for Open Fiscal Data

CHAPTER 6

OpenAPI Data Integration

Datasets, including open data, are often provided via APIs, and the RESTful web services have gained popularity over the past decade. Unification and automation of RESTful web services' documentation and descriptions are currently receiving increasing attention. The open-source *OpenAPI* Specification (formerly known as Swagger) has become the core of this effort and has been adopted by several major companies. It allows the description of RESTful web services using objects represented in JSON or YAML file formats. As a result, the created descriptions are human and machine-readable, but not machine-understandable. In this chapter, we propose a non-intrusive approach for semantically annotating the popular OpenAPI standard with an approach that is in many aspects similar to adding RDFa[1] "semantics" into HTML documents. For machines to understand the semantically enriched lightweight Web API descriptions, we present a comprehensive vocabulary for describing RESTful web services using the OpenAPI Specification. We demonstrate how the semantic descriptions can be added to the OpenAPI schema in a minimally invasive way by adding URIs in certain fields of the OpenAPI schema without breaking the standard.

This chapter is based on the following publication:

- **Fathoni A. Musyaffa**, Lavdim Halilaj, Ronald Siebes, Fabrizio Orlandi, Sören Auer, *Minimally Invasive Semantification of Lightweight Service Descriptions*, Proceedings of the 23rd International Conference on Web Services 2016; San Francisco, CA, USA.

6.1 Background

Extensive work has been conducted on XML-based Web Service descriptions as well as their integration with semantic annotations, widely known as Semantic Web Services [86]. However, XML-based Web Services are frequently perceived to be heavy-weight, complex[2]

[1] https://www.w3.org/TR/rdfa-primer/
[2] http://www.tbray.org/ongoing/When/200x/2004/09/21/WS-Research

and due to the overhead, slow [87]. The vast amount of research on making Web Services more 'Semantic' peaked about a decade ago and did not deliver a convincing approach outside its research community. This is mainly because previous efforts were based on complex XML-standards, verbose, and tedious to read by humans, as opposed to formats like JSON. In turn, previous standards (e.g. WSDL and WSMO) remained difficult to use and adapt to different use cases. Consequently, the amount of new work on this subject has faded over the last decade.

Swagger — recently named as OpenAPI[3] — has recently become the most popular schema to document a RESTful API [88], supported by a large community of active users and strong support for almost every modern programming language and deployment environment. It allows the description of the RESTful web services using objects represented in JSON or YAML file formats. As a result, the created descriptions are lightweight, human- and machine-readable, however, they lack support to allow machines to semantically 'understand' the functionality of the services and the data communicated. Having machine-understandable descriptions is crucial to support automation utilizing service discovery, composition, and choreography. We propose in this chapter to 'hitchhike' on the success of the OpenAPI grass-roots evolved standard and provide a way to semantically describe web services.

6.2 Challenges

Information exchange between services would improve if several properties are machine-understandable. These properties include *syntactic* interoperability and *semantic* interoperability. Syntactic interoperability deals with data formats used to exchange the information in a well-defined syntax and encoding [89]. To achieve syntactic interoperability, the services can use standardized data formats (e.g. XML, JSON) and encoding formats (e.g. Unicode, ASCII). Semantic interoperability is an important property of web services since it allows not only parsing the information exchanged by the services, but also understanding the information to be exchanged [90]. The data can be processed, recognized, and exchanged from one system into another only when the semantics of data is defined and shared [91]. In this case, the services exchange their data, include the semantic meaning of the content. Several important challenges need to be addressed for the realization of machine-understandable Web Service descriptions. These challenges are presented as follows:

Service discovery (CH1). Finding one or more suitable services for an activity is the task of *service discovery*, which is usually done by searching within available web service description repositories. According to [92], web service discovery is *the act of locating a machine-processable description of a Web service-related resource that may have been previously unknown and that meets certain functional criteria.* It involves matching a set of functional and other criteria with a set of resource descriptions.

This matching process can be implemented by *syntactic matching* and *semantic*

[3] http://swagger.io/introducing-the-open-api-initiative/

matching. Syntactic matching is based on identifier names, while semantic matching is based on semantic descriptions provided by the web services. Syntactic matching has low matchmaking accuracy and limited automation support [93].

Service composition (CH2). Service composition is the process of aggregating multiple services into one single service to perform more complex operations [93]. Web service composition is a complicated task mainly because of the following reasons: 1) the rapid growth of available web services in recent years; 2) the fact that web services can be created and updated on the fly; and 3) organizations provide similar web services that are described with different underlying conceptual models [94].

6.3 Semantic OpenAPI Specification

In this section, we elaborate on how semantic technologies contribute to solving the challenges mentioned in the previous section, namely, service discovery and composition.

Addressing the challenges with the semantic approach

Semantics technologies play an important role related to the management of applications, devices, and services [95, 96]. According to [97], a fundamental component of the semantic web will be the markup of web services to make them machine-readable, use-apparent, and agent-ready. Therefore, our semantic approach is developed following the fine-grained principles of the Semantic Web and Linked Data. In the following, we show how our approach responds to the aforementioned challenges.

Service discovery (CH1) Employing RDF as the standard for describing web services enables them to be automatically queried by machines. This can be achieved using SPARQL,[4] as a query language recommended by W3C for RDF. Furthermore, a service registry can be created to host these semantic descriptions and serve as a central hub for discovering web services. As a result, machines can get all necessary information for the services using standardized mechanisms.

Service composition (CH2) Currently, OpenAPI only specifies which primitive types (e.g. DOUBLE or STRING) are expected as valid input and output data. To enable more 'semantic' information on what these types actually *(*contain), we propose to use the description property fields from the OpenAPI standard for adding URIs as semantic annotations. The advantage is that this will not 'break' the parser, and allows a non-intrusive way to add semantic information. For example, if an operation on a service requires an input parameter that corresponds to a municipality class in DBpedia, the input parameter description is appended with the link of *Municipality* class definition in DBpedia.[5] If an operation response introduces a *Dataset* class with a specific definition, it can provide a link to the newly introduced class, for example, *<http://www.openbudgets.eu/ontology/obeu/Dataset>*. Therefore, internet agents or third-party applications benefit from semantically annotated input and output and combine

[4] http://www.w3.org/TR/rdf-sparql-query/
[5] <http://dbpedia.org/ontology/Municipality>

several web services to perform complex operations in order to provide comprehensive results.

Architecture

As mentioned before, We propose to extend the current OpenAPI Specification with some additional constructs that both do not change the OpenAPI standard and add the necessary semantics for service discovery, composition, etc. From an architectural point of view, we expand the OpenAPI specification with some additional components. The proposed architecture is illustrated in Figure 6.1. These components are grouped as the *Semantic Layer* and offer functionalities for achieving semantically-enriched OpenAPI Specification. The existing OpenAPI Specification provides a list of properties and terms to describe an API based on OpenAPI standard, including constraints in the values of each property. *User API description* characterizes the API being developed or documented. It includes operations available in the API, also parameters required and the response provided in each operation. Based on this *User API description*, API developers write their API descriptions by using the various editors like the one provided by the Swagger Framework (1). The result of the editing process is an *API description* in the YAML or JSON format. Based on this written API description (2), the *Codegen* generates *Server Stub Code* (3). *Codegen* is accessible from the *Editor* and supports several programming languages and frameworks. The *Swagger UI* (part of the current Swagger Framework), automates the generation of documentation once the API code in the server has been annotated either manually or by utilizing the *Codegen* (4). The generated documentation is provided as a web page and can be explored interactively (5). The *OpenAPI Vocabulary* is based on the *OpenAPI specification*, and by using *OpenAPI vocabulary*, a semantic description of the *API Description* is created (6). The output is provided in a JSON-LD serialization format. The resulting *Semantic API Description* is later stored into a *Triple Store* as a registry, which enables a semantic discovery functionality for the registered services (7). Publishing API descriptions semantically is utilized as a basis for semantic web service discovery and composition by third party applications (8). These applications can be other web services, or scientific workflow systems as will be described later in section 6.4.

The OpenAPI Specification Vocabulary

There are two approaches for modeling semantic web services: a top-down approach and a bottom-up approach. In the *top-down* modeling approach, several generic ontologies such as *WSMO* [98] and *OWL-S* [99] have been designed to semantically describe web services before they are actually implemented. On the other hand, in the *bottom-up* modeling approach, semantic annotations are added to already available web services. This approach is implemented in *SAWSDL*, where a semantic layer for describing WSDL is created [86, 100].

Following the principles of a bottom-up approach, we develop a lightweight vocabulary for representing RESTful web services based on the OpenAPI Specification. Several

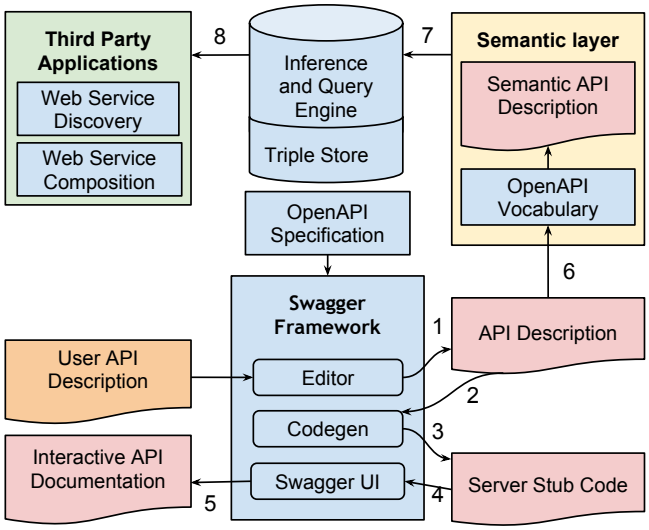

Figure 6.1: Architecture of semantic-enriched OpenAPI Specification.

vocabularies such as *Schema.org*[6] (`schema`), *Dublin Core*[7] (`dcterms`), *FOAF*[8] (`foaf`), and *Vcard*[9] (`vcard`) are used for defining the necessary concepts.

The diagram for the ontology we propose is provided in Figure 6.2, with `swg` used as the prefix for our OpenAPI ontology. This diagram illustrates a list of main classes and their relationships with the vocabulary. Each box on the figure is comprised of three main parts. The first part of the box depicts the name of the class. The second part describes the properties and related classes. The last part provides properties and related literals. The italicized properties and literals in the diagram show the corresponding properties in the OpenAPI Specification which do not have strict limitations on the allowed characters, i.e. they allow any kind of ASCII STRING.

OpenAPI uses *YAML* or *JSON* format to describe an API. Semantic lifting over this JSON specification can be done using the OpenAPI vocabulary,[10] providing a way to transform an OpenAPI-defined API specification. Additional code examples are available in the Github repository[11] as well as in Appendix C. Listing C.1 provides an example of how OpenAPI description is provided using YAML within an open budget domain. From the illustrated JSON format OpenAPI description in Listing C.2, semantified version can be written in JSON-LD format[12] as illustrated in Listing C.3.

[6] http://schema.org/
[7] http://purl.org/dc/terms/#
[8] http://xmlns.com/foaf/spec
[9] http://www.w3.org/2006/vcard/ns#
[10] http://vocabs.cs.uni-bonn.de/eis/oapi#
[11] https://github.com/fathoni/swg-sample
[12] https://www.w3.org/TR/json-ld/

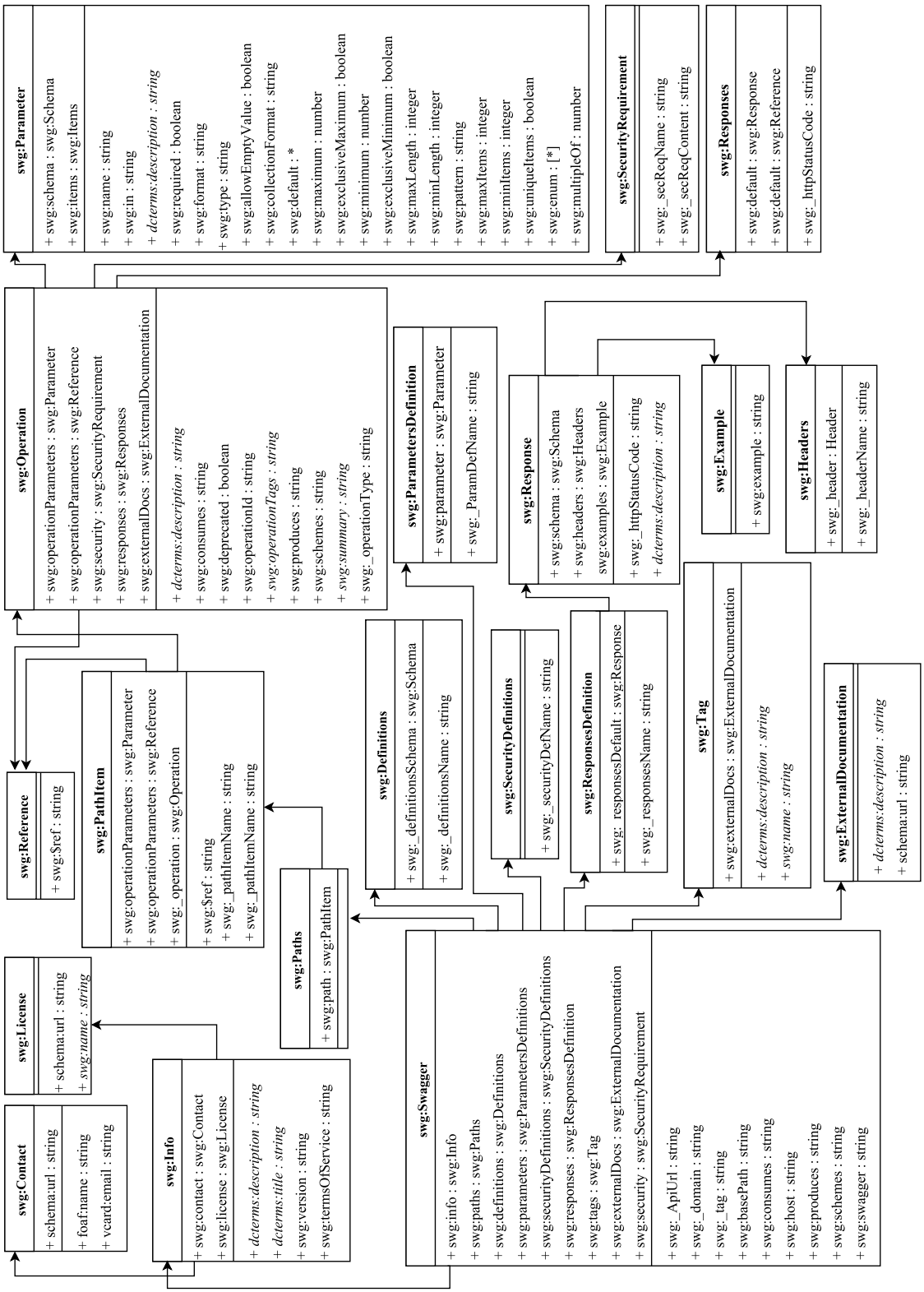

Figure 6.2: The ontology describing concepts of OpenAPI.

The relationship between these OpenAPI classes can also be seen in Figure 6.2. The `Swagger` class is the main class for describing any API. This class corresponds to the *Swagger* object in the OpenAPI specification. The property values of this class can be either a literal or another class. The `swagger` property is used to define the version of the OpenAPI Specification and is represented as a string literal. List of MIME types that the API produces and consumes are described through `produces` and `consumes` properties. The `host` property provides the domain that hosts the API and `basePath` property gives the location of the API service relative to the `host`. Some other properties in `Swagger` object accepts classes instead of literals as range, such as `paths`, `tags`, `definitions`, `info`, `responses`, `security` and `securityDefinitions`.

To be able to fully utilize the semantic descriptions added to the OpenAPI Specification, the vocabulary needs to be extended for several reasons. First, there is no specified property in the OpenAPI Specification which express a resource identifier. The concept of a resource identifier is of paramount importance in the linked data world. Hence, we explicitly specify a URI property for creating an API resource identifier from OpenAPI-described API endpoint. This URI is a concatenation between OpenAPI-defined `host` and `basePath` properties. Secondly, the OpenAPI specification has a description property that provides information about the API. The execution of SPARQL queries over large datasets using regular expressions in order to search specific patterns in the service descriptions results in poor performance. For this reason, we added two more properties, `_domain` and `_tag` property, to enrich the description of the API with additional information that limits the search space. The `_domain` property is limited by a static set of options specified by our extended schema. On the other hand, the `_tag` property is a flexible property that allows users to define their own API 'keywords' as long they are valid URIs. Discovering API via `_domain` and `_tag` property improve performance compared to searching directly through `description` property. Finally, according to the OpenAPI Specification, some properties can be described in non-fixed terms, such as the name of the parameter definition object, response name object, definition name object, and path item object. To address this issue, several properties are defined which hold the names of a non-fixed terms, i.e. `_paramDefName`, `_responseName`, `_definitionName` and `_pathItemName`.

The list of added properties is provided in Table 6.1. Extended properties that are not explicitly stated in OpenAPI specification are provided with "_" prefix. In the table, there are two main classifications for these extended properties. Firstly, the properties to hold the name of non-fixed terms in OpenAPI specification. These properties include the name of several objects, such as the parameter definition name, the MIME-type, and the type of operations. Secondly, some properties allow for additional metadata. As mentioned before, these are intended to improve the performance of linked data queries (such as `_domain` and `_tag`) or provide an identifier for the API (such as `_apiUrl`).

6.4 Use Case

Relevant to the sample code in Listing C.2, consider a case where an exact input of municipality and year as defined in DBpedia's `Year` and `Municipality` are needed. The

Table 6.1: Additional properties to support OpenAPI Specification Vocabulary.

Property	Domain	Range	Description
swg:_paramDefName	swg:ParametersDefinitions	xsd:string	{name} of field defining the Parameter Object.
swg:_responseName	swg:ResponsesDefinitions	xsd:string	The {name} of the Response Object, which maps into the response it defines.
swg:_mimeType	swg:Example	xsd:string	MIME type of provided Example object. This MIME type MUST be one of the Operation's produces values (either implicit or inherited).
swg:_securityDefName	swg:SecurityDefinitions	xsd:string	The {name} of the SecurityDefinitions Object.
swg:_apiUrl	swg:Swagger	xsd:string	The URL of the API. The value consisted of API host and basePath.
swg:_definitionName	swg:Definitions	xsd:string	The {name} of the Definitions Object.
swg:_domain	swg:Swagger	xsd:string	Selected domain for the API that has been predefined.
swg:_headerName	swg:Headers	xsd:string	The {name} of Header Object.
swg:_httpStatusCode	swg:Response	xsd:string	HTTP status code for the Response Object.
swg:_operationType	swg:Operation	xsd:string	The type of the Operation. Possible values include *get*, *put*, *post*, *delete*, *options*, *head* and *patch*.
swg:_pathItemName	swg:Paths	xsd:string	The relative {path} to an individual endpoint.
swg:_requiredSchema	swg:Definitions	xsd:string	The {name} of required parameter on a Definitions class.
swg:_scopeName	swg:Scopes	xsd:string	The {name} of Scopes that will be mapped to the Scopes short description.
swg:_securityReqName	swg:SecurityRequirement	xsd:string	The {name} of SecurityRequirement object. This name MUST correspond to a _securityDefName declared in SecurityDefinitions Object.
swg:_tag	swg:Swagger	xsd:string	Tag describing about the domain of the API.

API service validates if the given input is a type of Year and Municipality entity as described in the parameter definitions. A successful response should return a Dataset entity. The corresponding input and output entity is defined both in the JSON and in the semantic JSON-LD description of the API.

Another more complicated task is when a user wants to compare the spending of cities in Germany with a population of around 500,000 people and an area around 250,000 km^2. In many cases, web services are distributed across different hosts and organizations. This task requires the discovery of several web services from these different sources. For finding the area and population, data can be queried from publicly available datasets, such as DBpedia.[13] The data about spending for corresponding cities may have been distributed across several web services from several sources. In this case, a service discovery process is required to find related web services.

There might have been several available web services related to open government. Some

[13] http://wiki.dbpedia.org/

of them provide information about demography, while the other services provide information about spending and budget. However, we are only interested in finding the spending information about certain cities gathered from DBpedia that fits the specified criteria. We need to find which web services contain the necessary information for our case. Using our approach, this can be done in two ways. First, by inspecting the `description` properties in the API description. In the `description` property, a link to a class definition is used to describe semantically matched input/output of the services. This link is provided inside angle bracket ("<>"), for example `<http://dbpedia.org/ontology/Municipality>`. Another way is by querying the OpenAPI semantic registry utilizing `_domain` and `_tag` properties to support finding related services based on certain goal defined on the task.

Another use case is in the domain of Scientific Workflow systems. A Scientific Workflow is a graphical representation of a pipeline of executable processes for the purpose of scientific analysis. These workflows allow scientists with no programming knowledge to understand, execute, and share so-called `in silico experiments`, which means experimentation performed by computation. The wide variety of Scientific Workflow systems like KNIME [101], Taverna [102] and Pipeline-Pilot[14] have in common that they offer graphical workbench for scientists to create and share chains of executable components. The functionality of the components falls in various categories like data-transformation, data-preparation, data analysis, etc, and can both be part of the workbench or provided via local services and Web Services. The growing amount of scientific open data APIs is the motivation for this use case to make it easier for the scientific workflow community to integrate these APIs into their workflows.

6.5 Existing Approaches in Semantic Web Services

There exist many approaches in the research area of the Semantic Web Services. However, we will elaborate on those which are considered to have a strong influence in this field and are relevant for the goals of this chapter. The Web Application Description Language (WADL)[15] is designed to provide a description of HTTP-based web applications in XML format. These HTTP-based web applications are typically REST services. WADL strictly targets the HTTP(S) protocol, while WSDL is protocol-independent. This makes WADL simpler but has limited scope compared to WSDL [103]. WADL was submitted back in 2009 but in the meantime, W3C has no plan to use WADL as a recommendation.[16]

WSMF [104]: Web Service Modeling Framework is a modeling framework that describes several aspects of web services. WSMF provides four components including *ontologies*, *capabilities repositories*, Web services *descriptions*, and *mediators*.

OWL-S [105], formerly DAML-S, is a well-known ontology for creating semantic descriptions of the web services in the OWL standard. This ontology is interconnected with three subontologies: (1) *Profile* which responds to the question *what the service does*;

[14] http://accelrys.com/products/pipeline-pilot/
[15] http://www.w3.org/Submission/wadl/
[16] http://www.w3.org/Submission/2009/03/Comment

(2) *Process Model* responds to *how the service works*; and (3) *Grounding* responds to *how the service can be accessed*.

WSMO [106] is based on WSMF and consists of four main concepts: *Ontologies*, *Web Services*, *Goals* and *Mediators*. *Ontologies* provide terminology for the semantic description of WSMO components, i.e. resources and interchanged data. *Web Services* are identified as computational entities that can bring value in a specific domain. *Goals*, represent the objectives of the clients for certain functionality during Web Service consultation. *Mediators*, handle heterogeneity between interrelated elements by resolving mismatches of used terminologies of different levels, namely, *data level*, *protocol level* and *business process level*.

WSDL-S takes a bottom-up approach for semantic annotations of web services [107]. WSDL-S has three extension attributes that are used to associate the semantic annotations of the web service elements. First, the *modelReference* attribute provides the possibility to specify the semantic model of interrelation between a concept and a WSDL entity. Second, the *schemaMapping* attribute is used for handling the structural differences between XML Schema elements and their corresponding concepts presented with the semantic model. Third, the *category* attribute is included in the interface element to allow the organization of the information with the aim of publishing in different Web Service registries.

The Semantic Annotations for WSDL and XML Schema (SAWSDL) specification [108] is a reduced and homogenized version of the WSDL-S. It is a light-weight approach to annotate WSDL services with the objective of providing an extensible and agnostic solution regarding the ontologies and languages used for defining the conceptual models and their respective transformations. Therefore, users are forced to choose specific ontologies for semantically describing their services.

In order to address the lack of advocating a particular representation language for annotating services, a new version of SAWSDL called WSMO-Lite [109] has been developed. It supports concrete real-world challenges in intelligent service integration by addressing the following requirements: (1) identifying a simple vocabulary for semantic descriptions of services; (2) specifying the annotation mechanism for WSDL using this vocabulary; and (3) providing a bridge between WSDL, SAWSDL, and existing domain-specific vocabularies such as domain ontology models, classification schemes, etc. Swagger (OpenAPI) specification has not yet been supported by WSMO-Lite [110].

SWEET (Semantic Web sErvices Editing Tool) provides a way to insert semantics into an HTML web page describing web services [111]. By Using SWEET, users can annotate service properties on the web page describing web API properties. These semantic properties are in the form of hRESTS microformat tags and MicroWSMO model reference tags. RDF MicroWSMO description can then be extracted from annotated HTML pages. In our work, we annotate the OpenAPI JSON description instead of annotating the HTML pages using microformats. This is done in a simple way, by providing the context or link to related input/output entities on the JSON description in the developed API. By utilizing the developed vocabulary presented in this chapter, the RDF format can then be extracted into JSON-LD format and stored into a semantic service registry.

CHAPTER 7

Semantic Representation of Open Fiscal Data

In general, open data are frequently published in tabular formats, such as Microsoft Excel (XLS/XLSX) format, Comma-separated Values (CSV) format, Tab-separated Value (TSV) format or Open Document Spreadsheet (ODS) format. In a less-frequent circumstance, some datasets are also published in XML format, or even in a non-structured manner, such as PDF and Microsoft Word document (DOC/DOCX) format. With such heterogeneous formats being used by the open fiscal data publishers, processing fiscal data becomes a challenge. This chapter elaborates state-of-the-art data models that can be used to specifically represent open fiscal data. We compare the compatibility of heterogeneities elaborated in chapter 5 with state-of-the-art fiscal data models: the OpenBudgets.eu (OBEU) data model and Fiscal Data Package (FDP) which are designed specifically for representing budget and spending datasets. The comparison provides lessons learned for both datasets publishers and technical/research communities that deal with open data in budget and spending domain.

This chapter is based on the following publication:

- **Fathoni A. Musyaffa**, Fabrizio Orlandi, Maria-Esther Vidal, Hajira Jabeen. *Classifying Data Heterogeneity within Budget and Spending Open Data*. International Conference on Theory and Practice of Electronic Governance (ICEGOV) 2018. Galway, Ireland.

7.1 Available Data Model

In chapter 4, we discuss the issues and recommendations for open fiscal data quality. This is continued by chapter 5, elaborating the analysis result of open fiscal datasets heterogeneities. These heterogeneities can be minimized if the datasets are constructed following data models that comply with the particular specification, since one of the key requirement of government data quality, authority and governance is metadata

specification and data documentation standards [112]. Standardized data models can make the data more reusable. In the context of open fiscal data, two data model specifications exist, namely the Open Fiscal Data Package[1] for tabular datasets, or the OpenBudgets.eu data model[2] for RDF data. Third parties are proven to be willing to develop tools and services for consuming and analyzing government data [113]. Providing reusable open data can significantly reduce the costs of reuse, adaptation, and innovation for third parties.

Fiscal Data Package

Fiscal Data Package (FDP) is an evolving public fiscal data model and has been briefly discussed previously in section 3.2. Summarized from [114], FDP is designed based on the following modeling properties:

1. Consisted of main dataset/resource (in CSV format) and metadata (in JSON format) as core components. The usage of CSV and JSON utilizes open-standard.
2. Self-documenting metadata, with a progressive requirement. Some metadata are obligatory, but some are recommended/optional.
3. Designed with automated and standardized processing and analysis in mind.
4. Specifying detailed concepts common on budget and spending data (e.g., activity, entity, location). The FDP data model covers basic fiscal concepts, such as administrative and functional classifications, suppliers, amounts, etc.
5. Providing descriptors that define package metadata (name, country code, title, author, license, profiles, granularity, fiscal period), resource (column names and types), and models (mapping from CSV into FDP-defined logical models) such as measures and dimensions).
6. Online analytical processing (OLAP)-based design, which means the concepts of measures and multiple dimensions are taken into consideration.
7. Specifying some harmonized classifications, such as COFOG [23] [115] by the United Nations and GFSM [59] by International Monetary Fund. In FDP, non-harmonized classifications could be modeled as well.

OpenBudgets.eu Data Model

OpenBudgets.eu (OBEU) data model is an ontology for modeling budget and spending datasets into a linked data format [29]. It is developed by the OBEU consortium to model fiscal datasets semantically for the OpenBudgets.eu, an EU H2020 research project.[3]

[1] http://frictionlessdata.io/specs/fiscal-data-package/
[2] https://openbudgets.eu/resources/2016/11/17/open-budgets-data-model-and-landscape/
[3] http://openbudgets.eu/

OBEU ontology is based on the SKOS and DCV (see chapter 2), allowing concepts description within the concept scheme and multidimensional data representation using RDF. Therefore, it shares similar concept as DCV, as illustrated in more budget-context terms here. For example, the term *observation*, *measure*, *dimension*, and *attribute* [116], elaborated below:

- *Row*, *observation,* and *budget line*. Every *row* in a tabular file from a budget/spending datasets correspond to an *observation* (in DCV terms) or a *budget line* (in OBEU terms). An *observation* consists of an observed *value* (such as the amount of money spent), along with corresponding dimensions (such as for which office and functional usage this value is spent) and attributes (such as the currency of the value).

- Measure and amount. The *measure* defines the available value in a particular observation. In the budget and spending context, a measure typically represents the amount of money being budgeted/spent within an observation. A measure may also contain information such as population or budget/spending as the percentage of GDP.

- The *dimension* defines the measure in more detail, e.g., the classification to which the observation belongs. An observation in budget and spending datasets also typically contains a temporal dimension for the observed measure.

- The *attribute* provides more precise information on the observation, for example, metrics (e.g., currency: € or £), precision level, or the measurement unit (e.g., km or meter). The combination of dimensions make an observation unique, and the availability of attribute clarify the observation in more detail.

The OBEU data model considers the following modeling patterns [29]:

1. *Data Structure Definition* (DSD). A DSD is an additional file that provides detailed information regarding every dimension, measure, and attribute that is available in the datasets. A DSD is required to model datasets using OBEU data model.

2. *Component specification for budget/spending domain.* The OBEU data model specifies different dimensions, attributes, and measures that frequently occur in budget and spending datasets. There are 20 components defined within OBEU core data model, in which some are *abstract* components. Abstract components require data maintainers to extend these components for more fine-grained modeling.

3. *Support for coded dimensions/attributes.* Budget and spending datasets are often provided with classifications in the form of encoded notation along with its description. In the OBEU data model, these classifications are provided as a *code list*, represented using *Simple Knowledge Organization System* (SKOS) [36] vocabulary (see chapter 2).

4. *Integrity Constraints.* Several constraints introduced in the OBEU data model to avoid inconsistencies in data modeling, such as *namespace-hijacking*, mandatory component properties missing, properties instantiation, and wrong character case in DCV. The occurrence of these constraints can be checked using pipeline tools so that valid datasets transformation can be ensured.

5. *Lossless Mapping.* Mapping into OBEU's RDF data model should ideally preserve the information on the source of the original datasets.

6. *Dealing with multi-currency datasets.* The OBEU data model can handle datasets with multiple currencies by providing the currency as both dimension and attribute in each observation.

7. *Slices views.* OBEU data model supports *slice* views to ease data consumption. Slice allows viewing a piece of information from the dataset with regard to specified dimensions.

8. *Data normalization.* OBEU data model facilitates normalization in terms of *component attachment* and *schema implementation.* In the component attachment, the normalization is performed to make the mandatory properties available in the *observation* level, instead of *slice* or *dataset level.* In schema implementation, the normalized datasets are implemented using the star schema or snowflake schema which reduces data redundancy. This implementation optimizes storage but may affect the query performance.

9. *Datasets Versioning.* OBEU data model recommends using snapshots file only for budget phases. Minor fixes should not be provided as a snapshot. Instead, the fixes should be updated in place, as well as documented in the dataset's metadata.

10. *Optional properties.* Even though DCV is strict regarding the cardinality dimension, OBEU data model recognizes the existence of optional properties in the fiscal domain. However, optional properties do not identify observations. This means that if two rows are containing similar mandatory properties but having different optional properties, these rows are not regarded as unique rows. Since the uniqueness of rows in data cube is important, such a case may violate the data cube integrity constraints in DCV, which in turn also violates the OBEU data model integrity constraints as well.

11. *Classification versioning* (i.e., versioned code lists). Since the public administrations may publish some classifications annually, an extra effort to handle these annual versions should be done. Similar classifications across different years should be modeled on annually-different classifications. Connecting these classifications over the years should be done to provide links using relevant mapping properties, such as SKOS' *exactMatch property.*

12. *Metadata implementation.* OBEU recommends the usage of existing vocabularies (e.g., DCAT, DCAT-AP, FOAF, DC, etc.) to define the metadata of the datasets. Some mandatory metadata fields are defined in the OBEU data model.

7.2 Linking OBEU Data Model and FDP

The following Table 7.1 below compares enumerated heterogeneity (Section 4.3) with OBEU data model stack as well as FDP data model stack (Section 5). The plus '+' sign indicates the fact that the current data model able to represent heterogeneity among datasets, while the negative '-' sign represents otherwise and asterisk '*' sign represents limited support. The *stack* in this table refers to the respective data model as well as the included tools accompanying the data model. For example, the FDP stack would include the FDP data model itself as well as the Packager tool to transform the original CSV resource dataset into CSV and JSON format, i.e., the FDP data model. This table with additional explanatory comments is available online.[4]

Table 7.1: Support of heterogeneities on the state-of-the-art fiscal data models.

Data Model	**Heterogeneity**	**Subheterogeneity**	**OBEU DM Stack**	**FDP Stack**
No	**Heterogeneity**	**Subheterogeneity**	**1**	**2**
1	***CONTENT***			
1.1	***Measure***			
1.1.1	*Observation granularity*	Transaction	+	+
		Aggregation	+	+
1.1.2	*Funding source*	Single source funding	+	+
		Multiple source funding	+	-
1.1.3	*Sign representation*	Positive	+	+
		Positive and Negative	*	*
1.1.4	*Currency*	Single currency	+	+
		Multiple currency	+	-
1.1.5	*Time granularity*	Annual	+	+
		Non-annual cycle	+	+
1.2	***Classifications***			
1.2.1	*Insignificant classification hierarchy*	Existent	*	-
		Nonexistent	+	+
1.2.2	*Number of available classifications*	Standard classification	+	+
		Non-standard classification exist	+	*
1.2.3	*Classification structure*	Hierarchical	+	+
		Non-hierarchical	+	+
1.2.4	*Publication interval of classifications*	Once with occasional updates	+	-
		Everytime datasets published	*	+
1.2.5	*Classification Harmonization*	Harmonized	*	*

[4] https://goo.gl/o5H7Cx

Table 7.1 continued from previous page

Data Model	Heterogeneity	Subheterogeneity	OBEU DM Stack	FDP Stack
		Non-harmonized	*	*
1.2.6	*Classification Necessity*	Mandatory only	+	+
		Mandatory and optional	*	+
1.3	***Availability***			
1.3.1	*Budget phases*	Drafted	+	+
		Revised	+	+
		Approved	+	+
		Executed	+	+
1.3.2	*Observation description*	Description available	+	*
		Description unavailable	+	+
1.3.3	*Metadata availability*	Metadata available	+	+
		Metadata unavailable	-	-
1.3.4	*Budget direction*	Revenue	+	+
		Expenditure	+	+
9 2	***STRUCTURE***			
2.1	***Table Normalization***			
2.1.1	*Budget phase attachment*	Normalized	+	*
		Denormalized	+	+
2.1.2	*Budget direction*	Normalized	+	*
		Denormalized	+	+
2.1.3	*Classification attachment*	Normalized	+	-
		Denormalized	+	+
2.2	*Classification structure*			
2.2.1	*Classification notation*	Encoded	+	+
		Provided as plain label	+	+
2.2.2	*Abbreviated Classification*	Abbreviated	*	*
		Non-abbreviated	+	+
3	***SYNTACTICAL***			
3.1	***File format***	CSV	+	+
		Excel	+	*
		XML	+	-
		HTML	-	-
		PDF	-	-
		RDF	+	-
3.2	***Character encoding***	Encoding supported	Any	?
3.3	***Metadata***	Metadata type	DCAT	Data Package

7.3 Lesson Learned

After enumerating the heterogeneities encountered within open fiscal data, we provide several lessons learned from the compiled heterogeneities list. The lessons learned are particularly relevant for open fiscal data publishers as well as technical/scientific communities.

For Budget and Spending Data Publishers

Over the past few years, the technical and scientific communities have been working to provide sufficient tools and models for handling the open budget and spending data. In Table 7.1 above, it can be seen that some heterogeneities over datasets are either not yet supported or supported but with certain limitations. For example, OBEU stack has no or limited support for: measures with positive and negative values, datasets with insignificant classification hierarchy, datasets with classifications that are published periodically, datasets with harmonized and non-harmonized classifications, datasets with optional classifications, datasets without metadata, datasets with abbreviated classification labels and datasets with unstructured file formats. Therefore, if the datasets publishers want to make their published datasets compatible with OBEU stack, they should adapt their datasets for maximizing supported characteristics within OBEU stack. On the other hand, FDP stack has no or limited support for datasets with joint-funding amounts, datasets with positive and negative values, datasets with multiple currencies in a single amount column, datasets with insignificant classification hierarchy, datasets that are published with one-time published classifications, datasets with harmonized and non-harmonized classifications, datasets with a described budget line, datasets without metadata, datasets with abbreviated classification, datasets with normalized classifications, datasets with normalized budget phase, datasets with normalized budget direction and datasets published in a file format other than CSV. Similarly, datasets publishers are recommended to adapt their datasets characteristics so that it optimizes compatibility with FDP stack.

The choice of a particular stack depends upon the use case of the public administration. If the public administration expects their data to be modeled/consumed in a more flexible, descriptive way and intended to be analyzed in RDF, then their datasets have to be published in an OBEU stack compatible manner. If the datasets' publisher is more concerned about easy consumption without much technical skills required (albeit less descriptive), then the datasets' publisher is mostly interested in publishing their data to be compatible with FDP stack. FDP-packaged datasets can be transformed semi-automatically using an Extract-Transform-Load (ETL) pipeline [117].

For Technical and Scientific Communities

Table 7.1 shows that there are some heterogeneity issues which are not considered in the data model design yet, such as negative values interpretation (in both OBEU and FDP stack), multiple source funding (in FDP stack), multiple currencies (in FDP stack), insignificant classification hierarchy (in both stacks), nonstandard classification (in FDP stack), harmonized and non-harmonized classification (in both stacks), a classification which is published once - and therefore normalized (in FDP stack), classification which is published periodically (in OBEU stack), optional classification (in OBEU stack), datasets that provide observation description (in FDP stack), datasets with normalized budget phase, budget direction, and classification (in FDP stack), and datasets other than CSV format (in FDP stack). These known limitations against heterogeneity can be used as an

evaluation to improve the currently evolving budget and spending data model, as well as the technology stacks to process the budget and spending datasets.

CHAPTER 8

Semantic Uplifting of Open Fiscal Data

With the benefit of making multidimensional data having semantics, it is important to transform open fiscal datasets as linked data. However, adding semantics on fiscal data requires the availability of technical expertise within public administration bodies. Despite the availability of a fiscal data model that able to represent these datasets with semantics, the datasets need to be on a certain structure as we see in chapter 5 and chapter 7. Available, free, and open-source tools have been available for the semantification of these datasets. However, the complexity in performing the semantification of these datasets prevents dataset publishers to do so. In addition, the benefit of providing semantics on published fiscal data had not been widely known, making publishing datasets using semantics are pragmatically less relevant. By the end of this thesis, we aim to provide a proof of concept regarding the benefits of making datasets annotated with semantics. This chapter provides a building block to that goal by showing how semantics can be added from existing datasets. It provides some use cases on real-world datasets that we have transformed into RDF from its original raw formats (in CSV and XLSX). We give several example datasets gathered by the community or suggested by public administrators, followed by the analysis of these datasets, the transformation of the core datasets and code lists into RDF, and later, we present the final format of the datasets that have been transformed into RDF.

A brief part of this chapter is based on the following works:

- Jindřich Mynarz, Jakub Klímek, Marek Dudáš, Christiane Engels, **Fathoni A. Musyaffa**, and Vojtech Svátek, *Reusable transformations of Data Cube Vocabulary datasets from the fiscal domain*. Semstats 2016 in ISWC. Kobe, Japan.

8.1 Transformation Flow

The common flow of transforming the fiscal datasets into semantic, multidimensional format is shown in Figure 8.1. This process includes:

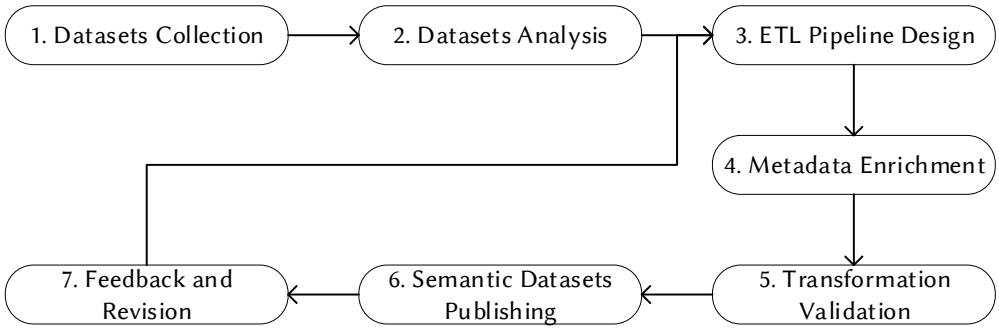

Figure 8.1: General process for fiscal datasets semantic lifting.

1. Collect datasets, along with documentation and classifications. The datasets can be collected from corresponding public administrations' open data portal, as well as contacting the public administrators from the city. These datasets, however, should be redistributable.

2. Analyze the datasets' structure, classifications, meaning, and correlation of each data item from the obtained datasets. Due to the heterogeneity of these datasets, this can be a demanding work that involves cooperation with public administrators to understand the datasets. Additionally, the datasets may be in a language that is not spoken by the data analyst.

3. Design an ETL pipeline to transform the datasets and accompanying classifications. The design of the ETL pipeline is very much different for each dataset, depending on the structure of the datasets that are being transformed.Through creating a different pipeline for different types of datasets, a set of common patterns in transforming datasets is recognized, hence common ETL pipeline fragments using LinkedPipes ETL are developed and provided in the work of [117].

4. Add additional metadata for transformed datasets, preferably using standardized metadata convention, such as DCAT-AP.[1] DCAT-AP is a short term for *DCAT Application Profile for data portals in Europe*. It is developed based on a vocabulary recommendation Data Catalog Vocabulary (DCAT) by Linked Data Group of W3C. DCAT [82, 118] is an RDF vocabulary developed to support the interoperability of web-published data catalogs. DCAT-AP is designed to specifically enhance the semantic interoperability of systems within European data portals.

5. Check the validity of the datasets using validator (e.g., DCV validator and some validator pipelines). DCV, as well as the OBEU data model, has several strict restrictions that have to be fulfilled. Some tools are available for validation, for example, DCV validation can be done via NoSPA-RDF Data Cube Validator[2] or

[1] https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe/release/11

[2] https://github.com/yyz1989/NoSPA-RDF-Data-Cube-Validator

via LinkedPipes ETL using DCV validator pipeline fragments[3] and OBEU data model constraints validation.[4] These LinkedPipes-based DCV and OBEU data model validation are described in [117].

6. Publish the datasets. Fiscal datasets that have been transformed into RDF using the OBEU data model are distributed by loading these datasets into a triple store, and then the SPARQL Endpoint can be shared so that the datasets can be queried by the interested parties. Additionally, these datasets can be distributed as an RDF dump, formatted in any of the RDF serialization formats (e.g., RDF/XML, turtle, JSON-LD, and so on). The datasets that we have transformed are provided publicly as an RDF dump in our Github datasets repository.[5]

7. Consider feedback and revise the transformation process if necessary. Since the process of data transformation may contain missing information and ambiguity, some feedback might be provided by the community. In this case, we update the data transformation and return to step 3 where the pipeline design is revised.

8.2 Transforming The Datasets

Transforming open fiscal data involves datasets collection, datasets analysis, ETL pipeline design, datasets metadata enrichment, transformation result validation, and sharing the transformation result. Afterward, datasets are revised according to feedback from data consumers. Each of these processes is elaborated in this section.

Collecting The Datasets

The datasets are collected based on suggestions by a fiscal data community using Github.[6] This community is involved in the OpenBudgets.eu project.[7] Additionally, we also work with the city of Bonn to analyze their dataset as a use case. Each dataset turns to have different structures and classifications, which provide representative examples of fiscal datasets characteristics: heterogeneous, independent, semi-structured, and rather complex. A comprehensive overview regarding the quality of this datasets can be seen on chapter 4, with a detailed spreadsheet containing how the quality of the dataset is calculated, and the source URL of the datasets are provided in an online Google Spreadsheet document.[8]

Datasets Analysis

To illustrate, each public administrations have their own system and business flow for their fiscal administration. For Example, in Bonn, the public administrators use

[3] https://github.com/openbudgets/pipeline-fragments/tree/master/dcv
[4] https://github.com/openbudgets/pipeline-fragments/tree/master/obeu
[5] https://github.com/openbudgets/datasets
[6] https://github.com/os-data/registry/issues
[7] http://openbudgets.eu/
[8] http://bit.ly/jdiq-datasheet-view

a rather advanced system via SAP,[9] resulting in advanced management of the public administration. Their business flow also uses a complex structure of code list, as well as very granular transaction/budgeting records available, to the extent that many budget line records have the measure of zero amount. Additionally, some budget lines have similar dimensions (see section 2.2) value, resulting in budget lines that are not unique. Having non-unique dimension values violates DCV integrity constraints, therefore, pre-processing to aggregate these budget line with similar dimensions are required. Additionally, some budget lines also have minus value of measure, which denotes that this is a *revenue* record.

Classifications are available in Bonn datasets in a composite manner: one numerical code is actually a joint of different numerical codes, each code has its own meaning (see section 5.2). These complexities and granularity hinder the understandability of the published datasets. Ideally, prior to publishing such datasets publicly, a deep analysis needs to be done to make the datasets easier to understand by common public and civil communities.

Another example is the datasets of Aragon, an autonomous community in Spain. The datasets[10] are published in an aggregate manner. Being an aggregate dataset, it has the drawback that it is not at a transactional level. However, the datasets are straightforward to understand, with classifications or code list is provided clearly.

Bonn's and Aragon's fiscal datasets are some datasets that we transform into the semantic format. Prior to transforming these datasets, we ask the domain experts or datasets publishers to decode some issues that make the datasets unclear. The extent of this discussion is very dependent on the data. Since Aragon datasets are straightforward, we only clarify the meaning of some classifications since there was no documentation in English. With Bonn datasets, however, it takes several meetings and requests datasets in a simplified form to be finally able to transform the datasets.

ETL Pipeline Design

Each dataset has its own pipelines and this is heavily dependent on its complexity. In this section, we provide examples of the pipelines we develop for several datasets. The transformation is done using LinkedPipes [119] an ETL tool able to perform semantic lifting from different sources using a graphical user interface, utilizing forms and specifically-defined SPARQL queries. The whole pipelines are available in our public Github repository of OpenBudgets.eu datasets.[11]

Bonn Transformation Pipelines

Figure 8.2 illustrates a sample of transformation pipeline. The whole pipelines are provided as a JSON-LD file which can be loaded to a LinkedPipes instance. There are five main tasks as numbered in the figure, consisting of:

[9] https://www.sap.com/corporate/en.html
[10] https://opendata.aragon.es/
[11] https://github.com/openbudgets/datasets

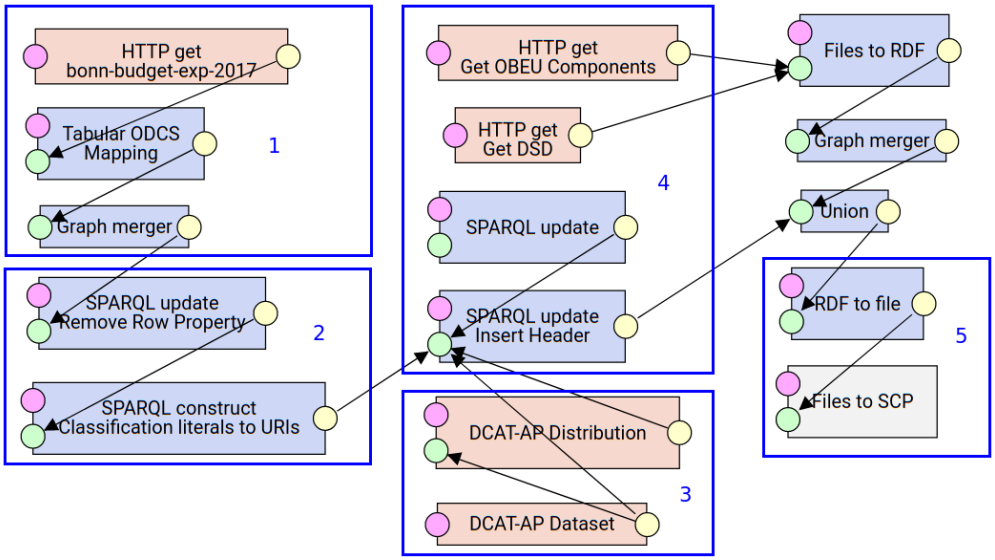

Figure 8.2: Semantic lifting pipeline of simplified Bonn 2017 datasets using LinkedPipes.

1. Download and column mapping. This task is illustrated in the first rectangle of Figure 8.2. Initially, raw data in the form of CSV or XLS is downloaded by providing links to the data source. However, if the dataset is not provided readily as stated in chapter 7, the data has to be first pre-processed and formatted so that it could accommodate the constraints imposed by DCV and OBEU data model.

2. Updating the mapped column with relevant properties using relevant SPARQL queries. For example, the SPARQL Construct in the second rectangle of Figure 8.2 is provided in the Listing 8.1. The query snippet shows how the observations are constructed, using `BIND` query that construct URIs following the desired URI prefix for each data item.

3. Providing standardized metadata for the datasets, as described by the EU's DCAT-AP specification [120] and the OBEU data model. LinkedPipes provide some components to perform this metadata enrichment using DCAT-AP's vocabulary.

4. Updating the datasets with the OBEU data model and Data Structure Definition (DSD). This is done to make the datasets conform with DCV and OBEU data model, as explained in the section 2.2 and chapter 7.

5. Serializing the resulting semantified data. In this case, the resulting data is then serialized into turtle format and then saved into either local or remote storage.

```
PREFIX  obeu: <http://data.openbudgets.eu/ontology/>
PREFIX  qb:   <http://purl.org/linked-data/cube#>
PREFIX  obeu-measure: <http://data.openbudgets.eu/ontology/dsd/measure/>
PREFIX  rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX  bonn-dsd:
↪  <http://data.openbudgets.eu/ontology/dsd/bonn-budget-simplified-updated/dimension/>
PREFIX  obeu-ds: <http://data.openbudgets.eu/resource/dataset/>
PREFIX  obeu-oc: <http://data.openbudgets.eu/resource/codelist/operation-character/>

CONSTRUCT {
  ?s rdf:type qb:Observation .
  ?s obeu:amount ?sum .
  ?s bonn-dsd:economicClassification ?economicUri .
  ?s bonn-dsd:profitCenter ?profitCenterUri .
  ?s bonn-dsd:functionalClassification ?functionalUri .
  ?s obeu:operationCharacter obeu-oc:expenditure .
  ?s qb:dataSet obeu-ds:bonn-budget-exp-2017 .}
WHERE {
  { SELECT  (MIN(?obs) AS ?s) (SUM(?amount) AS ?sum) ?profitCenter ?economic
    WHERE {
      ?obs  bonn-dsd:economicClassification  ?economic ;
            bonn-dsd:profitCenter  ?profitCenter ;
            obeu-measure:amount   ?amount}
    GROUP BY ?economic ?profitCenter}
BIND(substr(?profitCenter, 6, 9) AS ?pb)
BIND(uri(concat("http://data.openbudgets.eu/resource/codelist/kostenartenuebersicht_bonn/",
↪  ?economic)) AS ?economicUri)
BIND(uri(concat("http://data.openbudgets.eu/resource/codelist/profitcenter_bonn/",
↪  ?profitCenter)) AS ?profitCenterUri)
BIND(uri(concat("http://data.openbudgets.eu/resource/codelist/produktuebersicht_bonn/",
↪  ?pb)) AS ?functionalUri)}
```

Listing 8.1: SPARQL Query to construct URI for each observations from Bonn datasets.

In addition to datasets transformation to RDF, the accompanying classifications also need to be transformed. The way the classification transformed is also different case by case. In the case of the dataset from Bonn, the "*Profitcenter*" classification transformation is provided in Figure 8.3, which has four main components: 1) Data download and mapping, 2) removing unnecessary additional triples, 3) providing proper headers and concept hierarchies of the labels, and 4) serializing into certain RDF format and store the result into a remote host.

Aragon Transformation Pipelines

The way datasets transformed are different from each other. The public administration of Aragon also provides an out of the box - easy to understand, data portal. One of the datasets that they publish is an open budget and spending data. Here, we provide an example of Aragon datasets transformation pipelines, as shown in Figure 8.4, which has similar main components but differs in execution as well as its triple processing queries.

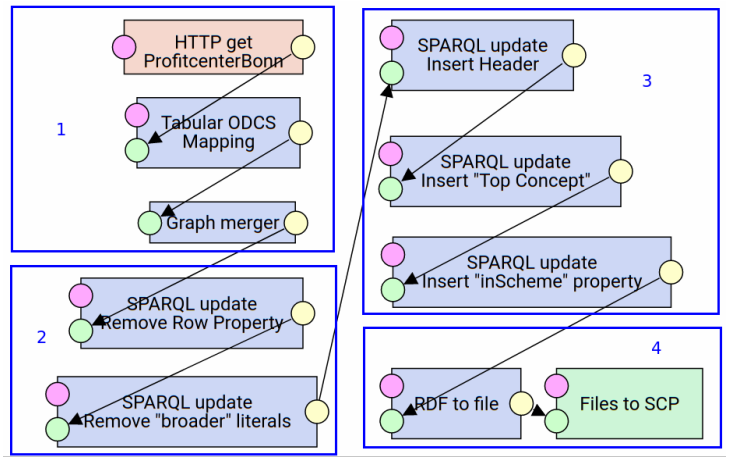

Figure 8.3: The transformation pipeline to semantify the *profitcenter* classification from Bonn.

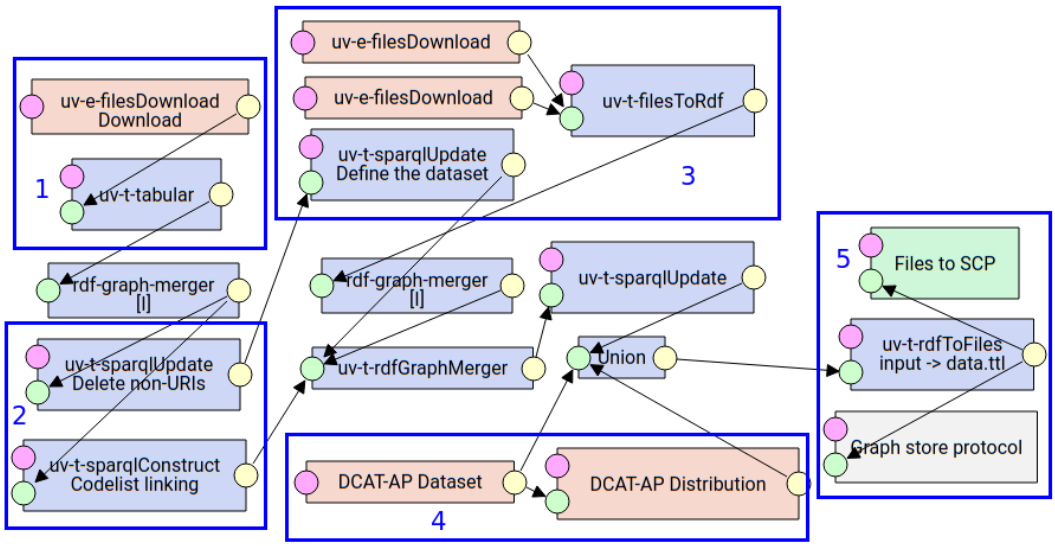

Figure 8.4: Semantic lifting pipeline using LinkedPipes from Aragon 2016 income datasets.

Metadata Enrichment

As mentioned in the previous section, the datasets are enriched with additional metadata to improve its interoperability with other datasets using DCAT-AP specification. DCAT-AP v1.1 to add metadata information items, implemented in DCAT-AP Dataset (as a mandatory application class) and Distribution component (as a recommended application profile class). These application profiles uses both Dublin Core Metadata terms (`dct`)[12] and Data Catalog Vocabulary (`dcat`).[13]

In DCAT-AP Dataset, the mandatory property include description (`dct:description`) and title (`dct:title`). Additionally, it is recommended to include contact point (`dcat:contactPoint`), dataset distribution (`dcat:dsitribution`), keyword/tag (`dcat:keyword`), publisher (`dct:publisher`) and theme/category (`dcat:theme`). For DCAT-AP Distribution, the only mandatory property is `dcat:accesURL`. It is recommended to have these properties: description(`dct:description`), distribution format (`dct:format`), and license (`dct:license`). LinkedPipes ETL has implemented these metadata application profile in their platform, as can be seen in the screenshot of Figure 8.5 and Figure 8.6.

Transformation Validation

The validation process is performed by checking if the datasets violate both DCV and OBEU integrity constraints. There are 22 DCV integrity constraints,[14] out of which 21 can be implemented using SPARQL CONSTRUCT queries [117]. The resulting validation errors are represented using RDF SPIN[15] to locate the problematic RDF resources. The OBEU data model integrity constraint is also checked using SPARQL CONSTRUCT Queries [117] via LinkedPipes pipeline fragments. It checks whether: 1) the component property of the code list is redefined, 2) the core namespace of OBEU data model is hijacked, 3) mandatory component property is missing, 4) whether a property is instantiated (as RDF allows classes instantiation only), 5) abstract property is used and 6) wrong character case of DCV is used.

The validation process of resulting transformation can be seen in Figure 8.7. The dataset, code list, OBEU, and DCV validation fragment are loaded into the LinkedPipes ETL tool. This process results in the OBEU and DCV validation reports.

Semantic Datasets Publishing

The datasets are published in two ways. First, the datasets are serialized using turtle format for readability and stored as flat files. These are then published in Github. Second, the datasets are loaded into the OBEU platform as will be described in chapter 9. Within the platform, a component used as a triple store is installed, in which Virtuoso[16] is

[12] http://purl.org/dc/terms/
[13] http://www.w3.org/ns/dcat#
[14] https://www.w3.org/TR/vocab-data-cube/#wf-rules
[15] https://spinrdf.org/
[16] https://virtuoso.openlinksw.com/

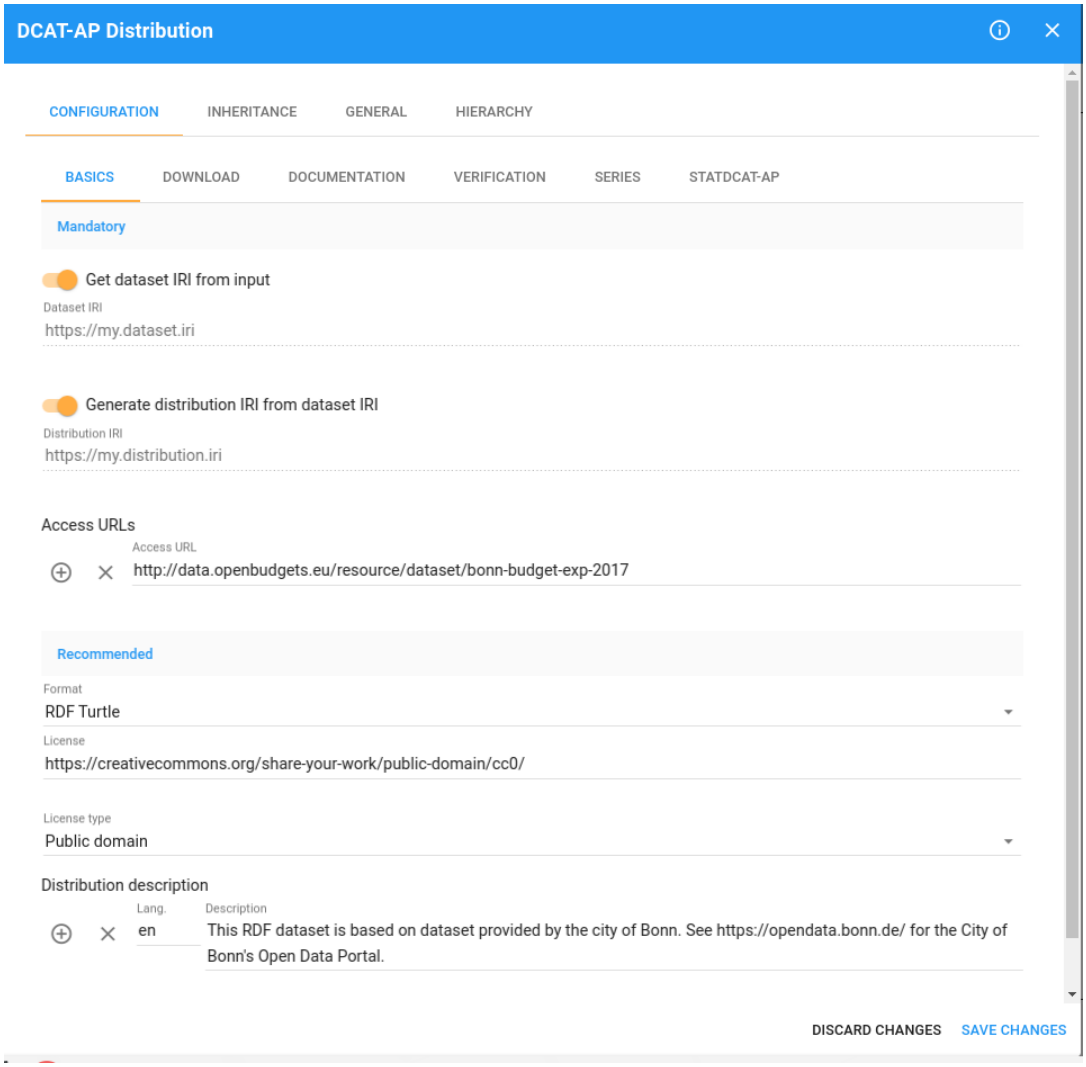

Figure 8.5: Implementation of DCAT-AP Distribution within LinkedPipes ETL.

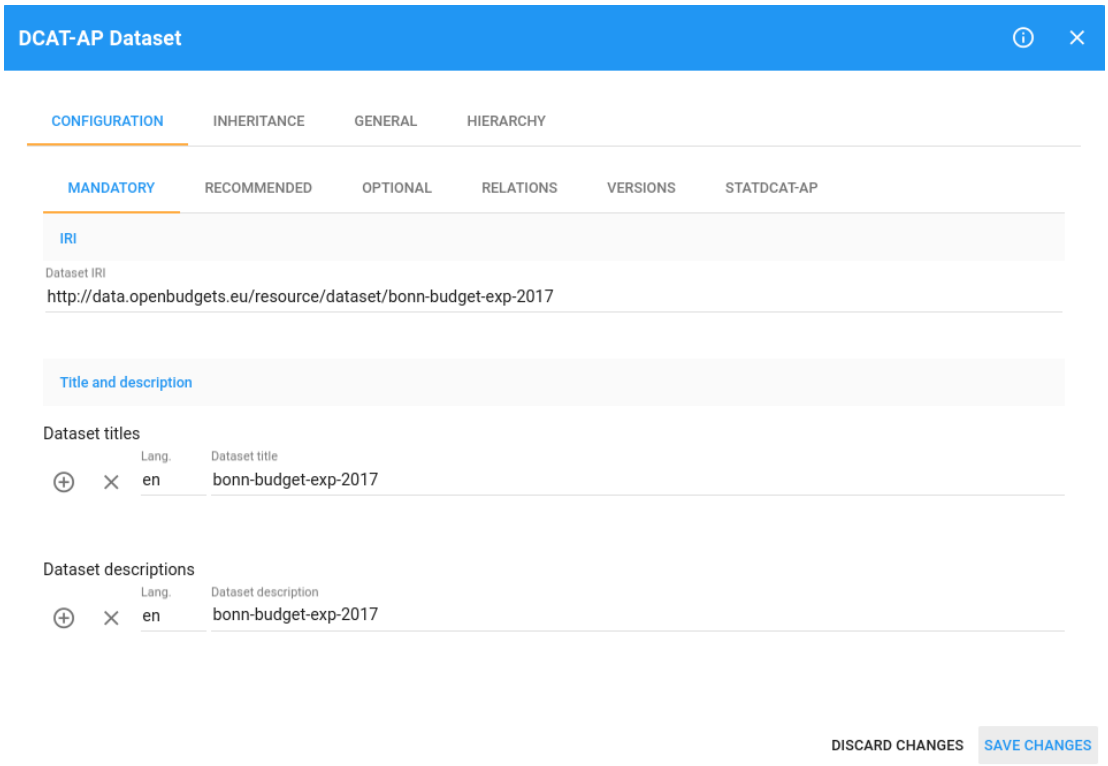

Figure 8.6: Implementation of DCAT-AP Dataset within LinkedPipes ETL.

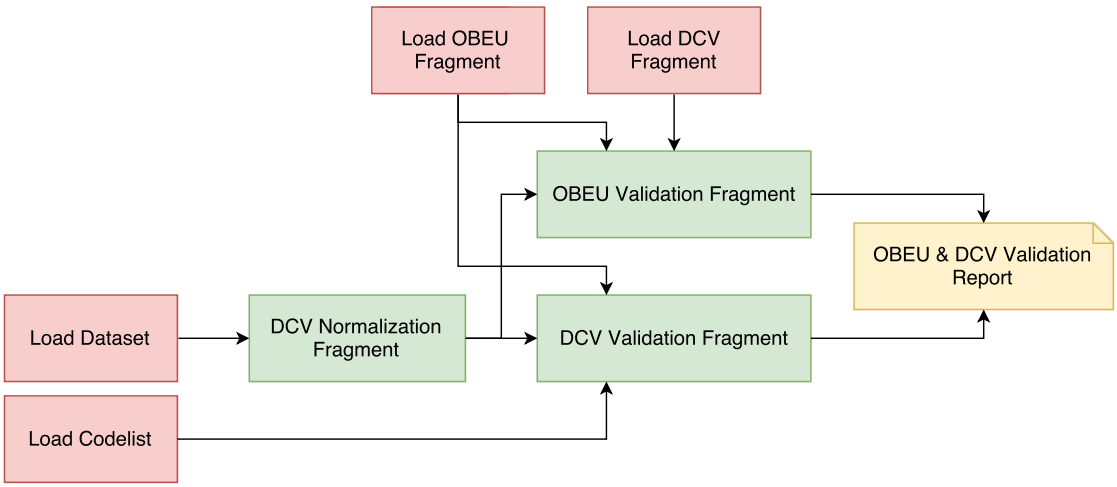

Figure 8.7: Validating the result of transformation from Bonn dataset.

installed for performance reason. The SPARQL endpoint from the platform is then shared publicly to enable public query over transformed and loaded fiscal data. Allowing SPARQL endpoints to be publicly accessible enables an advanced use case, which will be discussed in chapter 11.

Feedback Collection and Revision

Understanding the concept of DCV and OBEU data model takes a steep learning curve and hence this process may prone to errors for new adopters of these data models. The errors found on published datasets are then discussed as Github issue on the respective Github page. The resulting error pattern is collected and analyzed, which is then used as a basis to design a reusable OBEU data model validation pipeline [117]. Transformed pipelines of affected datasets are then revised accordingly.

8.3 Result

The final result of the transformation of the dataset is provided in our datasets repository listed in Github.[17] For datasets that are related to specific municipalities, we manage to have more than 10 million RDF Triples, distributed in more than 600 graphs. These triples are transformed from several different municipalities of Germany, Spain, and Greece. In addition to state-/city-based datasets, additional datasets are transformed, e.g., European Structural and Investment Funds, among others.

[17] https://github.com/openbudgets/datasets

CHAPTER 9

Open Fiscal Data Analytics Platform

Publishing open fiscal data allows CSOs, citizens, journalists, or stakeholders to gain knowledge about the data. However, deeper insights from the published data if the datasets can be compared across different public administrations that have similar characteristics. The linked data paradigm can help to harmonize and analyze the open budget and spending data. A major challenge, however, is to devise a software platform, which facilitates the harmonization of heterogeneous budget and spending data, while facilitating a variety of applications ranging from comparative analysis to participatory budgeting. In this chapter, we present the OpenBudgets.eu (OBEU) platform for linked open budget and spending data analysis. We materialize the conceptual open data architecture specifically for analyzing open budget and spending data on a platform. We propose a more fine-grained platform of linked budget and spending data analytics platform by considering requirements systematically collected from different sources. We collect requirements for a linked budget and spending data analytics platform, illustrate use cases using several actual datasets, and provide the platform design for a linked budget and spending data architecture.

This chapter is based on the following publication:

- **Fathoni A. Musyaffa**, Lavdim Halilaj, Yakun Li, Fabrizio Orlandi, Hajira Jabeen, Sören Auer, Maria-Esther Vidal, *OpenBudgets.eu: A Platform for Analyzing Semantic and Open Fiscal Data*. International Conference on Web Engineering (ICWE), 2018. Caceres, Spain.

9.1 Requirements

To collect the requirements for a linked open budget and spending data platform, we performed an analysis of several sources, namely open data life cycle [121] as well as several open data publishing guidelines (GODI [6], ODB [22], 5-star data ratings [27], Open Data Policy Guidelines [78]) and collected requirements through the OpenBudgets.eu

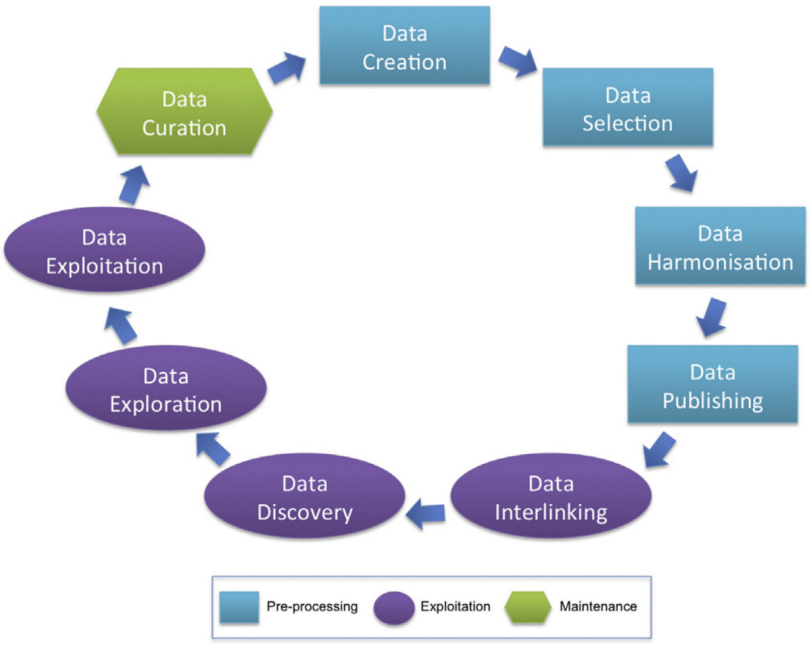

Figure 9.1: Open Data Life Cycle as proposed by Attard et. al.[121].

Community.[1]

There are several open data assessment methodologies as well as open data publishing guidelines. The W3C EGOV interest group provides recommendations and group notes,[2] which includes the Data Cube (DCV, see chapter 2) and Data Catalog (DCAT) vocabularies and the Internationalization Tag Set (ITS). The 5-Star data rating provides several suggestions, such as making the data to be available online, structured, in a non-proprietary format, provided with URIs, and linked to other data [27]. In addition, the Sunlight Foundation provides several comprehensive suggestions [78].

Furthermore, a questionnaire was conducted with the OpenBudgets.eu community. The respondents belong to different interest groups such as governance transparency, journalism, active public participation, e-government, technical implementation as well as research. In this gathering, 66 functional, 13 non-functional requirements, and 29 data quality indicators for a linked budget and spending platform are collected [122]. The summary of the gathered functional requirements can be found in the Table D.2 of Appendix D.

Attard et. al. [121] propose an open data life cycle (Figure 9.1) which consists of three main stages: *pre-processing*, *exploitation*, and *maintenance*. The pre-processing stage consists of data creation, selection, harmonization, and publishing. The exploitation stage consists of data interlinking, discovery, exploration, and exploitation. Finally, the maintenance stage consists of data curation. Based on the open data life cycle, open data

[1] www.openbudgets.eu
[2] https://www.w3.org/standards/techs/egov#w3c_all

assessments, and publishing guidelines, we have summarized several key requirements that should be implemented in the linked open budget and spending data platform.

Data Creation. Creating the datasets in public administrations is usually part of daily procedures. The main steps within the data creation are: providing documentation, providing provenance information, and ensuring that the datasets are authoritative.

Data Selection. Data selection involves the removal of existing private and personal data, as well as identification of conditions for publishing the data. Determining the list of available classifications (i.e., code lists, a list of predefined concepts that is used to group budget and items), checking for missing data, and enlisting available investment alternatives (in the context of participatory budgeting) are part of the requirements.

Data Harmonization. Making the datasets conformant with the open data publication standards is the focus of data harmonization. Steps within data harmonization include: creation of the RDF data model that supports budgets, revenues, incomes, transactions, classifications, amount, payer, payee and currency; acquisition of metadata; clarification of data usage license; semantic mapping of CSV data format to RDF; mapping of OpenSpending FDP data model to RDF; association of targeted amount to actual spending; and the linking of data items. Published datasets should ideally be provided as structured data in an *open format* using an *open license*.

Data Publishing. The main data publishing stage consists of different steps, such as data loading from CSV format or an API, providing kiosk mode on the data web page, as well as performing a customizable continuous integration, download option, and links to Freedom of Information Act/Access to Documents. Ideally, the published datasets should be easily and publicly accessible through an API as well as a bulk download; associated with license, contributors, and contact points information. The datasets should also be openly licensed and published in a sustainable manner, i.e., hosted on a government Open Data portal, an official website, or a preservable public platform (e.g., Github).

Data Interlinking. Data interlinking connects datasets and items within the datasets to other resources. The main step for datasets interlinking is a mapping between related classifications from different datasets, for example, mapping a functional classification (e.g., health, education, public infrastructure, etc) from a public administration with another functional classification published by different public administrations, which would enable comparative analysis (see chapter 11). Datasets should also be published as RDF and have a dereferenceable URI.

Data Discovery. The existence of open data should be discovered by data consumers. From the requirements perspective, data discovery can be enhanced by the availability of free-text search, the availability of semantic search, providing search result ranking, the availability of explorable processed datasets, availability of metadata, availability of the

feature to perform different levels of the query, and implementation of a user-friendly graphical user interface.

Data Exploration. To enable data exploration, simplified consuming options should be provided. The related steps for this requirement are: explained flow of budgeting process, tracking of budget version, availability of localized or translated data, querying by administrative regions or institutions, availability of search feature, availability of data exploration samples, visual exploration of both RDF and non-RDF data, availability of visualization suggestions, previewing the visualization, availability of geographical visualization, exporting and sharing high quality and indisputable visualization, tracking of user data processing workflow and cache processing data, budget comparison by using different dimensions (public administrations, time, and function), filtering (by spending or administration type), availability of top-level aggregation, and attachment of participatory budgeting results.

Data Exploitation. The next level of data cycle is exploiting the data, which is a more advanced step in consuming the data and allows users to provide analysis, mash-up, or some other innovations by using, reusing, or distributing the data. The requirements involved in the data exploitation stage include building custom visualization, performing exploit analysis, filtering commensurable objects, detecting outliers, extrapolating the data, aggregating the data by time interval, availability comparison between planned vs. spent money, normalizing by key metrics, differentiating between real vs. nominal value (e.g., inflation adjustments), providing contextual information, breaking down the budget and spending items, and attaching spending to participatory budgeting results.

Data Curation. Data curation is important to ensure data sustainability. Steps within data curation include pointing missing data, indexing both tabular and RDF graph data structures, as well as gathering budget votes for participatory budgeting. Datasets should ideally be published with detailed metadata and updated regularly and in a defined time interval. If possible, a version tracking for datasets should be provided.

9.2 Architecture

The high-level overview of the OBEU platform is provided in Figure 9.2. A more detailed data interaction between different components is provided in Figure 9.3, with most of the tools are developed during the duration of the OBEU project by OBEU partners, except tools within OpenSpending platform (which has been developed earlier and is being continuously developed) and Virtuoso Triple Store. As can be seen in Figure 9.2, there are five layers that build up the OBEU platform: data storage layer, data transformation layer, API layer, platform layer, and application layer. These layers are described in the following sections.

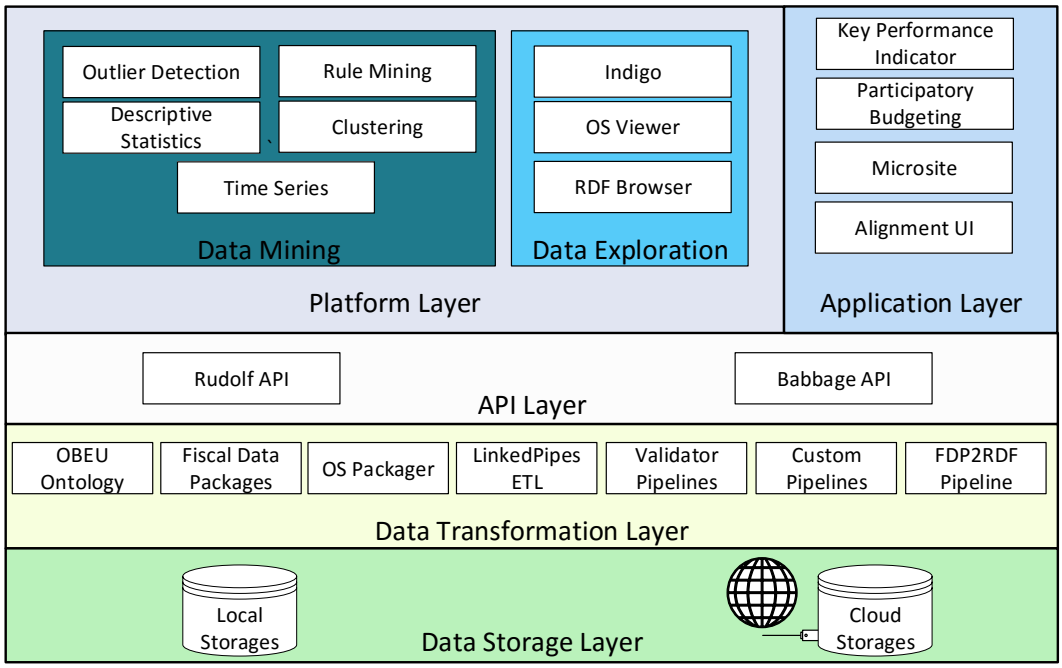

Figure 9.2: Logical Overview of the OBEU platform.

Data Storage Layer

The Virtuoso triple store is used to host all the graphs resulting from the data transformation pipelines. Non-RDF datasets coming from the OpenSpending (OS) packager interface are hosted in Amazon S3 cloud storage. This layer partially satisfies the requirements of Data Publishing and Curation in section 9.1.

Data Transformation Layer

Unifying heterogeneous budget and spending datasets is a challenging task due to their heterogeneity in terms of schema/structure, syntax, and format. Representing different open budget and spending datasets adhering to a unified and integrated data representation formalism significantly eases data analysis. As mentioned in chapter 7, there are two major data models for representing open budget and spending datasets: the OpenBudgets.eu (OBEU) data model [38] and the Fiscal Data Package (FDP) data model [114]. Both of these data models are used in this platform.

The ingestion of datasets can be performed through a step-by-step wizard using the *OS Packager*.[3] The usage of the OS Packager is recommended for users that do not have a strong technical background in transforming the data, i.e., those who are not familiar

[3] https://github.com/openspending/os-packager

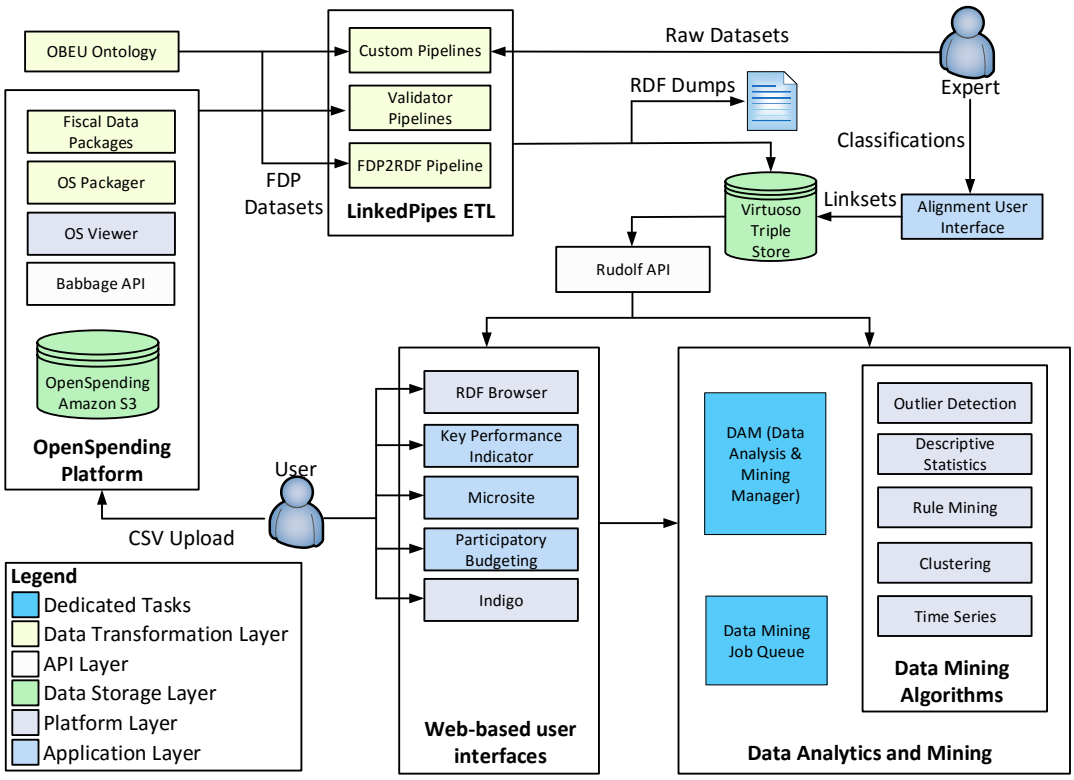

Figure 9.3: The Data Flow within the OpenBudgets.eu platform.

with linked data and SPARQL queries. With the OS Packager, the user annotates the datasets based on the schema of their data. These annotations are then saved as a JSON file. This JSON file, along with the original CSV format, makes up the *Fiscal Data Package* (FDP) data model. It should be noted that not all structures of the budget and spending data are supported by FDP. In such case, a manual data transformation should be done using an Extract-Transform-Load (ETL) tool. A comparison between supported features and limitation of the OBEU data model and FDP is given in chapter 7.

LinkedPipes ETL [119] is used for the ETL process. Using LinkedPipes requires some understanding of RDF concepts and constructing SPARQL queries. However, the users can flexibly arrange the components in such a way that fits the structure of the datasets to build up a *custom pipeline* for their own dataset transformation, so that all necessary information in their datasets can be represented in the *OBEU data model*. Datasets in the FDP format can also be transformed into the OBEU data model using a reusable *FDP2RDF* [117] pipeline, provided that the datasets have been correctly structured and annotated in FDP. Since building a transformation pipeline is an error-prone process, a reusable *validation pipeline* template [117] is provided to check the constraints imposed by the DCV and OBEU data model. Transformation pipelines from LinkedPipes can be exported into the JSON-LD format, which can then be imported in any LinkedPipes

instance.

On a side note, providing datasets with metadata significantly improves the accessibility of the datasets. FDP provides metadata when the user annotates and uploads the datasets using the OS Packager wizard. In the OBEU data model, users specify the metadata using several components, such as DCAT-AP distribution (for e.g., datasets access URL, format, license) and DCAT-AP datasets (for e.g., datasets title, description, IRI and contact point). The components in this layer facilitate Data Creation, Harmonization, and Interlinking.

API Layer

The *Rudolf API* [123] provides an API that fetches fiscal datasets from the OBEU RDF triple store. The Rudolf API derives data from the OS data store and serves the data for further tasks, such as data analytics and mining, as well as data visualization. OpenSpending's *Babbage API*[4] provides an OLAP-style implementation for querying database on PostgreSQL. This layer facilitates the Data Publishing and Discovery aspect.

Platform Layer

Indigo is the main dashboard that lets users choose available datasets to be explored (Figure 9.5). Users can then navigate through several other features, such as *Data Analytics and Mining* (DAM) and visualization. DAM provides a playground for scientists to experiment with the budget and spending datasets. Within the DAM component several algorithms are implemented, including several types of *outlier detection*, as well as *descriptive statistics*, *rule mining*, *clustering* and *time series* algorithms.

Two types of visualization are provided within the platform: standard and customized visualizations. A standard visualization is provided directly by the *OS Viewer*.[5] Customized visualizations can be easily integrated as well (e.g. as in the case of the city of Bonn[6]).

RDF browser[7] is designed to enable exploration of specific dataset entities using the particular URIs. By using the RDF Browser, users can inspect the relationship of items, in the form of URIs, within the datasets. The tools in this layer address the requirements of Data Discovery, Exploration, and Exploitation.

Application Layer

Alignment UI enables mapping between related concepts of classifications that are published by different public administrations. The interlinking of related concepts across different classifications enables comparative analysis across different datasets that have related concepts.

[4] https://github.com/openspending/babbage
[5] https://github.com/openspending/os-viewer
[6] https://github.com/shurkhovetskyy/obeu-vis
[7] https://github.com/okgreece/RDFBrowser

Microsite simplifies fiscal data website creation and embedding on public administration websites. Users are provided with configurable administrator dashboards to set which localization, data types, and visualization are embedded. Web page visitors can then comment on the showcased datasets.

Participatory Budgeting component allows public administrations to announce their budget plan and then let their citizens vote on their preferred budget allocation. The application within participatory budgeting components allows citizens to be more proactive in budget allocation decision-making.

Key Performance Indicator (KPI)[8] provides an analysis of fiscal performance from a specific dataset and organization. In KPI, users are also provided with a configurable administrator panel. The indicators examined within KPI include employment cost index to expenditure, total revenue to population, expenses per citizen, among many other indicators. The tools in this Application layer partially satisfy the requirements in several open data life cycles, including Data Selection, Publishing, Interlinking, Exploitation, and Curation.

9.3 Implementation

The Docker light-weight virtualization technology is used to integrate the components in different layers of the OBEU platform which is shown in section 9.2. The components are running within different Docker containers, and the access to different components is controlled by an Nginx web server which is also running within a Docker container. The internal communication between some components also goes through Nginx.

The management of the different docker containers is done by Dockerfile and Docker Compose. Dockerfile is used to build a Docker image and run Docker container, the configurations of Docker containers are done by Docker Compose. By using this management schema, the OBEU platform can be updated easily if there are some updates in any component, and the platform is also easily portable. Further documentation regarding integration and how to instantiate a new OBEU platform is accessible online.[9]

9.4 Evaluation

The OBEU platform is evaluated using (i) three large-scale trials conducted with municipalities, (ii) a survey on UI usability, and (iii) performance measurements. In addition, the evaluation is also conducted to see whether each requirement has been satisfied/partially satisfied/unsatisfied by developing and integrating related tools. This can be seen in Appendix D, with Table D.1 details the requirements alignment from Open Data Life Cycle [121], open fiscal data platform functionality requirements [122], and matching data quality factors [124]. Table D.2 elaborates OpenBudgets.eu platform functional

[8] https://github.com/okgreece/KPIs
[9] https://github.com/openbudgets/integration

requirements, associated tools, associated open data life cycle and its whether each mentioned requirement has been satisfied.

Use cases

To test the tools developed in the OBEU platform, large-scale trials were conducted in three municipalities: Bonn, Paris, and Thessaloníki. Seven different testing scenarios have been developed for each trial: (1) data ingestion with OS Packager, (2) automated data transformation to RDF from OS Packager, (3) ETL pipelines for RDF semantic lifting fiscal data using LinkedPipes, (4) Visualizations, (5) Microsite, (6) Data Analytics and Mining, and finally (7) Participatory Budgeting. Each municipality had to perform the same seven testing scenarios and was then asked afterward for comments and feedback. The main outcome of these testings is detailed in the project trials deliverable [125].

The first large-scale trial is implemented with the city of Bonn in Germany. Datasets from the city of Bonn are rich and complex, involving positive and negative values to indicate the expenditure and income, dimensions that do not uniquely define a budgeting item, as well as complex, nested classifications. The data structure available for the Bonn datasets was not supported by the current common structure in the OS Packager. Therefore, the dataset upload was performed in two scenarios: first by utilizing the custom pipeline to accommodate their data structure complexity, and second, by simplifying the initial datasets to adapt the supported structure by the OS Packager tool. Both scenarios successfully transformed the datasets to the OBEU data model, with the consideration that the first scenario needs more technical expertise. Embeddable visualization for the Bonn datasets was generated and tailored using the Microsite. The outlier detection algorithms, using either Local Outlier Factor (LOF) or frequency-based algorithms, have successfully detected unusual budgeting trends in the case of Beethoven's 250th birthday celebration in the budget year 2020. The City of Bonn has found that the Participatory Budgeting tool was easy to configure and use. However, implementing Participatory Budgeting tool for the citizen requires a large amount of political and bureaucratic work and consequently, the Participatory Budgeting tool is not used openly for now. The main feedback received from the city of Bonn was, that using the OBEU platform simplifies data ingestion (which would take a long time to comprehend), and once the datasets are properly ingested, the subsequent requirements in the data life cycle are adequately fulfilled.

The municipality of Paris is the second trial participant for the OBEU platform. Paris has already provided its datasets openly in a clean CSV format, and their datasets can be directly transformed to FDP using OS Packager. A custom pipeline was not necessary for Paris since the CSV and FDP format can be transformed into the OBEU RDF format using a reusable FDP2RDF pipeline. Visualizations have been tested and visualization within the Microsite is generated. According to this trial, the Microsite revealed to be the most visible component of the platform. Suggestions for improvement include making personalized data visualization available for other users, allowing for tabbed visualizations (default, expert, and visualization showcases), and allowing regulation of the Microsite by the communication/publication department. Data mining has not been used in the

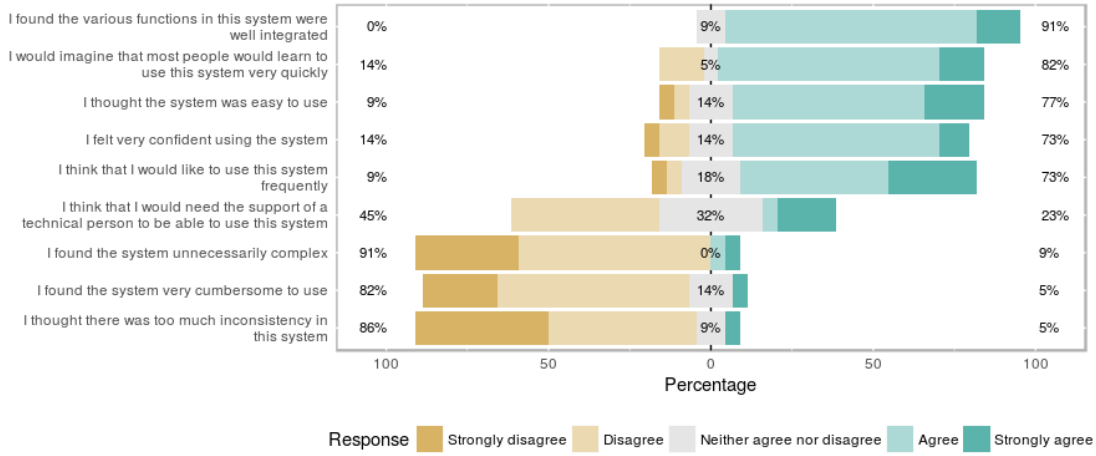

Figure 9.4: An Aggregated UI evaluation result from several OBEU tools.

Paris case, due to the lack of personnel with data mining expertise. Since Paris has also implemented its own participatory budgeting tool, the OBEU Participatory Budgeting tool has not been used in practice.

The last large-scale trial is the municipality of Thessaloníki. Data Ingestion with OpenSpending could be done easily since the municipality of Thessaloníki has already published Open Data which can be exported into different formats, including CSV that is already structurally compatible with OS Packager. A minor issue was found during the testing of the OpenSpending packager, which requires year/date columns in the datasets. Custom ETL pipelines were also created as a template so that other municipalities from Greece can reconfigure the pipeline and reuse it. The developed KPI visualizations offer rich financial performance indicators for Thessaloníki.[10] As with other municipalities, some of the data mining tools require domain expertise. However, under expert supervision, insights, and predictions over the data were found useful. An implementation of Participatory Budgeting was tested and currently planned for publication to the citizens of Thessaloníki.

Usability

Using a Likert-scale-based questionnaire, we evaluated several tools deployed within OBEU, namely OS Packager, OS Viewer, Microsite, KPI admin panel, KPI, and Linked-Pipes. The aggregated questionnaire result is summarized in Figure 9.4. The detailed, non-aggregated usability test result is available online.[11]

[10] http://kpi.okfn.gr/
[11] https://goo.gl/Kqkbc6

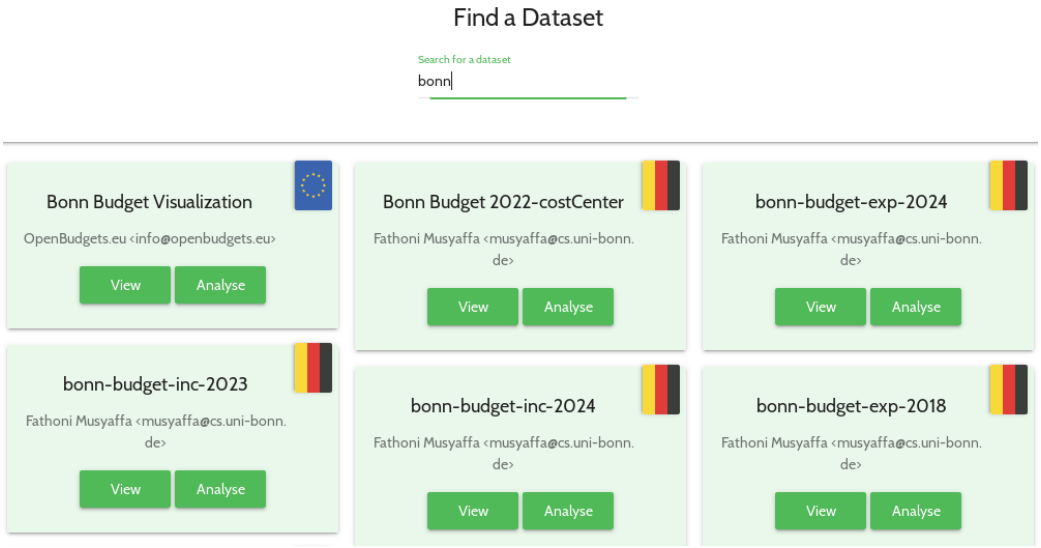

Figure 9.5: Search with a keyword in Indigo.

Performance

The OBEU platform is deployed on a server with CPU Intel®Xeon®CPU E5-2660 v3 @ 2.60GHz, 35 GB of RAM, 1 TB of disk and 4 GB Swap Memory. The Virtuoso triple store manages 12.6 million triples at the time of writing. There are 253 distinct datasets, 305 distinct classifications, 240 distinct data structure graphs, totaling 798 of distinct graphs.

The performance evaluation focuses on the data search and query performance. The entrance point of the platform for a normal user is Indigo, which sends a search request to the Rudolf API to load a certain size of datasets. Afterward, a user can search, as illustrated in Figure 9.5, e.g. using ”bonn” as a keyword. We use a script to initiate API calls and measure the runtime difference between a different number of datasets in the search request to the Rudolf API.

The runtime of searching through the Rudolf API with or without using a keyword is shown in Figure 9.6. We initiate 30 API calls for each data item size and plot the averaged run-time for each number of datasets. It can be seen that with the increasing number of datasets, the run-time does not deteriorate much.

```
PREFIX qb: <http://purl.org/linked-data/cube#>
SELECT DISTINCT ?s WHERE {?s a qb:DataSet}
```

Listing 9.1: SPARQL query that retrieves all OBEU datasets.

The evaluation of query performance is done by executing SPARQL queries against the SPARQL endpoint. To compare the performance of dataset listing using the Rudolf API

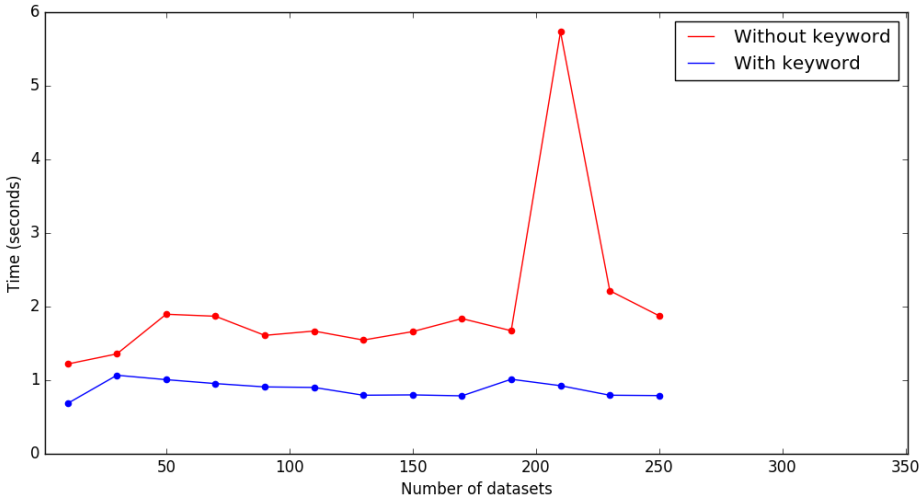

Figure 9.6: Runtime of searching through Rudolf API.

with a pure SPARQL query, we also measure the average time needed to fetch the list of datasets. The SPARQL query used to list the whole datasets is provided in Listing 9.1. The complete execution of this query takes 67 milliseconds on average. Compared with the Rudolf API execution time (see Figure 9.6), SPARQL querying is faster. Having an interface in Indigo which calls the Rudolf API is providing an easy-to-use interface for users, with some performance trade-off. However, the expense of loading time is justified when we consider the UI usability improvement as the evaluation shows in Figure 9.4.

Concluding Remarks for Part III: Data Management and Analytics for Open Fiscal Data

This part mainly discusses the topic of interoperability across datasets and fiscal data platform, aiming to answer this research question:

> **RQ3**: How can we improve the interoperability of open fiscal data by using a semantic data model?

Resuming from Part II, we continue with investigating how these datasets can be integrated. This is done in four different stages:

1. Obtain datasets from API endpoints that are provided using OpenAPI description. We annotate these descriptions with semantics to facilitate pulling the data from OpenAPI-based API endpoint.

2. Analyze compatibility of datasets with different fiscal data models.

3. Perform semantic lifting on open fiscal datasets.

4. Devise a platform to handle open fiscal data management and analytics.

In chapter 6, we presented an approach to describe Semantic Web Services by extending the OpenAPI specification in a non-intrusive way. The approach allows adding semantic information via an extensible and light-weight vocabulary that aims to enable automated service discovery, orchestration, and composition. As a result, internet agents and third-party applications can find and combine web service using their semantic descriptions to deliver comprehensive results.

Data models for open fiscal data representation is compared in chapter 7. We check whether heterogeneities found in open fiscal data are supported by state-of-the-art open fiscal data models. Lessons learned are provided for both datasets publishers and scientific/technical communities.

A detailed process of fiscal datasets transformation to RDF using the OBEU data model is elaborated in chapter 8. Due to its heterogeneous nature of fiscal datasets,creating one-pipeline-fits-all for these transformations is a non-trivial task. However, some transformation patterns can be reused to ease the transformation process.

A platform built for semantic open fiscal data analysis, OpenBudgets.eu, is described in chapter 9. Both conceptual architecture and implementation for a budget and spending data platform are provided to support the open fiscal data life cycle. This implementation addresses the challenges related to the open data life cycle. It integrates available relevant components and platforms in a micro-services architecture and extends them with extra tools, providing additional linked data capabilities. The platform has been evaluated with real application scenarios, usability, and performance tests.

Part IV

Enabling Comparative Analysis of Open Fiscal Data

CHAPTER 10

Scalable Interlinking of Multilingual Open Fiscal Data

Open budget data are among the most frequently published datasets of the open data ecosystem, intended to improve public administrations and government transparency. Unfortunately, the prospects of analysis across different open budget data remain limited due to schematic and linguistic differences. Budget and spending datasets are published together with descriptive classifications. Various public administrations typically publish the classifications and concepts in their regional languages. These classifications can be exploited to perform a more in-depth analysis, such as comparing similar budget or spending items across different, cross-lingual datasets. However, in order to enable such analysis, a mapping across the multilingual classifications of datasets is required. In this chapter, we present the framework for the Interlinking of Heterogeneous Multilingual Open Fiscal DaTA (IOTA). IOTA makes use of machine translation followed by string similarities to map concepts across different datasets. To the best of our knowledge, IOTA is the first framework to offer scalable implementation of string similarity using distributed computing. The results demonstrate the applicability of the proposed multilingual matching, the scalability of the proposed framework, and an in-depth comparison of string similarity measures.

This chapter is based on the following publication:

- **Fathoni A. Musyaffa**, Maria-Esther Vidal, Fabrizio Orlandi, Jens Lehmann, Hajira Jabeen. *IOTA: Interlinking of Heterogeneous Multilingual Open Fiscal DaTA*. Elsevier Journal of Expert Systems with Applications (ESWA). 2020.

10.1 Motivating example and use case

Open fiscal concepts published by different public administrations are often multilingual and there is no indication if two words have a similar meaning. For example, in Table 10.1, where datasets from the Aragon government (in Spanish) and from the municipality of

Thessaloniki (in Greek) does not indicate that the concepts within the table have similar or related meaning regarding the functional classification item *culture*. If a *mapping* exists between two concepts from different datasets, further analysis can be made possible.

Integrating classifications from different datasets allows at least two use cases. First, it allows a comparison of the allocated/spent budget for a particular classification item (for example, culture, public transport, and so on), even when the datasets' classifications are published in different languages. Second, the integrated classifications could be mapped to public semantic knowledge bases (e.g., Wikidata, DBpedia), to enrich the concepts with additional information (such as an instantiation of a class from a certain ontology and leverage word sense hierarchies). Both use cases allow a deeper understanding of the budget and spending datasets. More precisely, they allow data discovery and reusability which would provide actionable insight for public administrators, civil communities, NGOs, stakeholders, and most importantly, the citizens who fund the city with their taxes.

		Data A	**Data B**
Functional Classification	Language/Adm.	ES/Aragon	EL/Thesssaloniki
	Code	45	6471
	Label	Cultura	Έξοδα πολιτιστικών δραστηριοτήτων
Similar?	Translation	Culture	Cost of cultural activities

Table 10.1: A motivating example: functional classifications originating from Aragon (in Spanish) and from Thessaloniki (in Greek) which actually represent a similar concept of *culture* for the public budget. Each concept typically has its own code and label in the publisher's respective language, without indication that both concepts are, in fact, similar. Both classifications are published in separate spreadsheet documents.

10.2 Preliminaries

Owing to the structural complexity and the size of budget data, the automated cross-linking of these datasets is a challenging task. In order to do so effectively, it is important to develop mappings of similar attributes among different datasets available in different languages. Several efficient machine translation tools [126] exist to solve the problem of multilingual data. After the translation, various string similarity measures can be used to map similar concepts. However, the string comparison process is a computationally expensive task, especially when there is a high volume of concepts to be compared. Therefore, this task is not feasible for large scale data using a single machine. One of the recent in-memory distributed computing framework Apache Spark [44], can provide a scalable solution to solve complex tasks like mappings over large data. The above-mentioned challenges and existing technologies have inspired us to design and propose a framework that uses fiscal-data-classifications, machine-translation, and string-similarity in a distributed and scalable manner for interlinking of open fiscal data.

In this chapter, we present the IOTA Framework. IOTA uses a set of string similarity

Bag-based Similarity Measure	
Name	Description
TF/IDF	Takes into account term frequency to measure the similarity between two documents, and offset the similarity by the inverse document frequency so that commonly-appearing-terms' importance are discounted [128].
Phonetic-based Similarity Measure	
Soundex	To mimic pronunciation, Soundex replaces or removes characters from each compared strings, ended by examining processed strings. The steps are: 1) keep the first letter of each compared strings. 2) remove any occurrence of W, H, Y and vowels (A, E, I, O, U). 3) replace B, F, P, V with 1; C, G, J, K, Q, S, X, Z with 2; D, T with 3; L with 4; M, N with 5; R with 6. 4) remove any consequential identical digits (e.g., '22' to '2'). 5) keep only the first four characters but if the total length is less than four characters, the digit '0' is appended until it has four characters. 6) compare the processed strings which result in binary similarity score [129–131].
Editex	Editex is a Soundex similarity modification with different letter groups to represent a more accurate pronunciation and allows some characters to be on more than one of nine letter groups: group 0 = {A, E, I, O, U, Y}, 1 = {B, P}, 2 = {C, K, Q}, 3 = {D, T}, 4 = {L, R}, 5 = {M, N}, 6 = {G, J}, 7 = {F, P, V}, 8 = {S, X, Z}, 9 = {C, S, Z}; in which {W, H} is removed. Editex utilizes a Levenshtein-like similarity measure to compare processed strings [131].

Table 10.2: An overview of bag-based and phonetic-based string similarity metrics used in the experiment and applicable formula. Due to the complexity of some similarity measures, it is not possible to squeeze the summarized formula in this table. The similarity score of these algorithms is normalized by default. Soundex yields binary decision by default, while Editex needs the similarity score to be normalized.

measures to search the effective string similarity measures to find mappings between translated concepts. The *py_stringmatching* library[1] is used for string similarity measure calculation in IOTA. Five main similarity measure categories are presented and used for comparison in this chapter [127]:

- *Bag-based* (see Table 10.2) for similarity measures that collect tokens as bags in which a token in these similarity measures could appear multiple times.
- *Phonetic-based* (see Table 10.2) for similarity measures that mimic string pronunciation.
- *Sequence-based* (see Table 10.3) for similarity measures in which the inputs are considered as a sequence of characters.
- *Set-based* (see Table 10.4) for similarity measure in which the inputs are considered as tokens (i.e. words).
- *Hybrid-based* (see Table 10.4) for similarity measures which combines set-based and sequence-based similarity measures.

The overall list of compared similarity measures is provided in Table 10.5.

[1] https://pypi.org/project/py-stringmatching/

Sequence-based Similarity Measure		
Name	Description	Formula
Bag Distance	Counts characters in each string x and y as a multiset, subtracts the difference between elements in x and y as well as between the difference of elements between y and x, and chooses the maximum element count from these numbers [132].	$dBD(x,y) = max(\|x-y\|, \|y-x\|)$. And can be normalized by: $sE(x,y) = 1 - \frac{dBD(x,y)}{max(\|x\|,\|y\|))}$
Levenshtein	Measures the distance of two given strings based on how many minimum edit cost (insert/delete/substitute) are needed to make two strings identical [133]. Also known as *Edit Distance*.	Given *d(x,y)* is the edit distance between strings x , y, normalized Levenshtein similarity: $sL(x,y) = 1 - \frac{d(x,y)}{max(length(x), length(y))}$
Jaro	Counts how many common characters c are similar between strings x,y, and how many transpositions t are needed to make these common characters have a similar sequence [130, 134].	$sJaro(x,y) = \frac{1}{3 \times [\frac{c}{\|x\|} + \frac{c}{\|y\|} + \frac{c-\frac{t}{2}}{c}]}$
Jaro-Winkler	Improves *Jaro* by considering two extra parameters: the maximum length l of common prefix between two strings and the weight w considered for the prefix [130, 135].	$sJW(x,y) = (1 - l \times w) \times jaro(x,y) + l \times w$
Ratio	Utilizes parameters M as a total number of matches between elements in the strings x and y, and T as the total number of elements in both strings [136]. Score is normalized by dividing the result by 100.	$sR(x,y) = 2 \times \frac{M}{T} \times 100$
Partial Ratio	Compares the shorter string of length n with every sub-string of length n from the longer string. The maximum similarity score from these comparisons are provided as the partial ratio similarity score. Suppose between compared string x and y, x is the shorter string with length n. y is splitted into n-gram with length of n, resulting a set of tokens y with m member. $B_y = B_1, ..., B_m$ [136]. Score is normalized by dividing the result by 100.	$sPR(x,y) = max(\sum_{i=1}^{m} ratio(x, B_i))$
Partial Token Sort	Converts two strings into tokens, sorting the tokens, and calculates the partial ratio similarity score of calculated strings [136]. The score is normalized by dividing the result by 100.	
Token Sort	Splits two strings into tokens, sorts the tokens, and calculates the ratio similarity score [136]. The score is normalized by dividing the result by 100.	

Table 10.3: An overview of eight sequence-based string similarity measures used in the experiment and their respective formula. Some similarity measures (Token Sort and Partial Token Sort) use formula from other similarity measures. Some of the similarity scores are not normalized by default, those are Bag Distance, Levenshtein, Ratio, Partial Ratio, Partial Token Sort, and Token Sort similarity.

Set-based Similarity Measure		
Name	Description	Formula
Cosine-Ochiai	Computes the intersection between two sets of tokens, divided by the square root of the multiplication between the size of both token sets. This is A derivative of cosine's algorithm known as Ochiai coefficient [127, 137].	$sC(x,y) = \frac{\lvert B_x \cap B_y \rvert}{\sqrt{\lvert B_x \rvert . \lvert B_y \rvert}}$
Dice	Also known as Sørensen-Dice coefficient. It is calculated by twice the size of the intersection between two sets of tokens, divided by the size of both token sets [127, 138].	$sD(x,y) = 2 \times \frac{\lvert B_x \cap B_y \rvert}{\lvert B_x \rvert + \lvert B_y \rvert}$
Jaccard	The division between the intersection size of two sets and the union size across the sets [130, 139].	$sJacc(x,y) = \frac{\lvert B_x \cap B_y \rvert}{\lvert B_x \cup B_y \rvert}$
Overlap Coefficient	Indicates the overlap between two sets by dividing the intersection size between two token sets with the minimum size from of the two sets [140].	$simOC(x,y) = \frac{\lvert B_x \cap B_y \rvert}{min(\lvert B_x \rvert, \lvert B_y \rvert)}$
Tversky Index	The division between intersection size of the token sets with: the sum of intersection between sets, the number of items only available on the first token set multiplied by a coefficient α, and the number of element only available in the second token sets multiplied by a coefficient β [141].	$sT(x,y) = \frac{\lvert B_x \cap B_y \rvert}{\lvert B_x \cap B_y \rvert + \alpha \lvert B_x - B_y \rvert + \beta \lvert B_y - B_x \rvert}$
Hybrid-based Similarity Measure		
Generalized Jaccard	Calculated by 1) converting compared strings x, y into two sets B_x , B_y; 2) calculating the string similarity s between tokens across the two sets (hence Cartesian product is involved); 3) filtering the string similarity value s so that s is larger than specified threshold α. The result of this filtering is a bipartite graph mapping between B_x and B_y with similarity score $s > \alpha$ and collected into a graph M, which is then used to calculate the Generalized Jaccard similarity score; 4) getting the maximum similarity pairs s from graph M, and use the pair with maximum similarity s to calculate the final score. [130, 142].	$sGJ(x,y) = \frac{\sum_{(x_i,y_j) \in M} s(x_i, y_j)}{\lvert B_x \rvert + \lvert B_y \rvert + \lvert M \rvert}$
Monge-Elkan	Also requires specifying a string similarity as a parameter name. Calculated with the following steps: 1) compared strings x, y is tokenized to $x = A_1, ...A_n$ and $y = B_1, ...B_m$; 2) string similarity scores are counted against each token from the other set; 3) the maximum similarity score from each set is then taken from the two sets and then averaged. String similarity function $s'()$ is the chosen string matching similarity measure parameter [130, 143].	$sME(x,y) = \frac{1}{n} \sum_{i=1}^{n} \max_{j=1}^{m} s'(A_i, B_j)$
Soft TF/IDF	The calculation is done by 1) computing a similarity score between tokens, 2) filtering the tokens using a threshold, and 3) calculating similarity score using TF/IDF vectors along with filtered similarity score [144]. In this experiment, Jaro is used as the secondary string similarity measure.	

Table 10.4: An overview of five set-based and hybrid-based string similarity measures and their respective formula used in our experiment: Ochiai as a derivative of Cosine similarity (will be referred later here as Cosine), Dice (also known as Sørensen-Dice Coefficient), Jaccard, Overlap Coefficient and Tversky Index. In the set-based similarity metrics part, B_x and B_y are tokens generated respectively from compared strings x and y. All the resulting values from these similarity measures fall in the range of [0,1].

10.3 Existing Approaches

The work in this chapter involves several topics, ranging from open fiscal data analytics and platforms as well as multilingual datasets mapping. This section briefly covers the related existing approaches.

10.3.1 Open fiscal data analytics and platforms

The state-of-the-art in open fiscal data analytics have some limitations, with platforms related to open fiscal data have been mentioned in chapter 3. An important part of big data in an e-Government is the implementation of a robust architecture and data platform [112]. In chapter 9, we implement a platform for open spending and budget datasets. This platform provides a materialized and budget/spending-specific architecture for consuming Open Budget and Spending data. However, a missing component of the OpenBudgets.eu platform, as well as other platforms mentioned in chapter 3, is a mapping tool that could map concept labels from classifications by different publishers and in different languages.

Budget comparison across different public administrators could potentially be made if there is a *mapping across labels* from different languages. This mapping is a part of the *data interlinking* cycle, and according to [121], data interlinking is one of the eight elements within Open Data Life Cycle, and is particularly crucial for data exploitation stage. There are often similar concepts provided in different languages, but there is a rare chance that a mapping across these labels from different languages exists (see a related survey from chapter 4). Our work in this chapter addresses the mapping challenge by experimenting with a framework consisting of machine translation, multiple similarity measures, and a cluster computing architecture to find which similarity measures are more suited to create mappings for concepts originating from different languages.

10.3.2 Multilingual concept mapping

The mapping of concepts across multilingual concepts can be seen from different approaches. These approaches range from the area of distributional semantics, entity linking, as well as ontology-based data integration.

Distributional semantics

Firth [145] states that the words surrounding a word in question characterize its meaning. This is a basis for distributional semantics which uses the distributional properties in a large data sample for finding semantic similarities across language items. In the past few decades, several distributional semantic modeling approaches are developed, such as Latent Semantic Analysis (LSA) [146] and Hyperspace Analogue to Language (HAL) [147]. Another approach, word embedding, has been gaining popularity in the past few years. Word embedding maps words and phrases within a vocabulary to vectors of real numbers using a variety of methods. One of these methods uses a shallow neural network to

learn this mapping coined as Word2Vec [148]. Using Word2Vec, the semantic similarity between words of a similar language can be computed utilizing vector values that are being compared. Joulin et al. [149] have implemented the fast word embedding learning algorithm coined as FastTex and published the pre-trained models in multiple languages.[2] Aligning different languages into one vector space is done by MUSE [150]. The MUSE team has also published aligned word embedding vectors trained from different languages of Wikipedia.[3] These multilingual word embeddings that have been aligned into a single vector space can be further used to calculate the similarity between words from different languages. Yet, multilingual phrases similarity (instead of words) is not directly facilitated in the MUSE pre-trained model.

We have tested the aligned multilingual pre-trained vectors from MUSE for evaluating the quality of mappings between multilingual phrases. This is done by calculating the cosine similarity between averaged vectors of each concept from each language. If a phrase has multiple links, we select the phrases pair with the maximum similarity value. However, the result of using this averaged-vector approach from multilingual phrases is not satisfactory. For example, any language pair involving German language resulted in a maximum F-Measure value of 0.153 for CPV datasets. The following observations hold when working with pre-trained embeddings for specialized cases:

1. The pre-trained models are more generic and are not fully applicable to specialized fiscal data.
2. The training of models requires substantial amounts of training data, which is not widely available for fiscal data.
3. The effectiveness of word embeddings usage depends on the language. For example, in our experience, a word in German consists of several conjugated words that are not available in the publicly-available Wikipedia-based pre-trained word embedding vector index. Hence, it results in a word vector that can not be found in the vector index.
4. The published pre-trained word embedding vectors are, as the name suggests, based on words instead of phrases. Its application to the cross-lingual phrase similarity requires n-gram training from each respective language corpora. However, such data is not available in the fiscal domain.

Entity linking

Concept mappings can be done when concepts are represented as entities within knowledge bases. Pappu et al.[151] perform a lightweight multilingual entity extraction and linking using an approach they coined as Fast Entity Linker (FEL). The FEL approach detects mentions and retrieves entities, utilizing compact entity embedding that captures and searches several features used for entity disambiguation (e.g., click logs). Graph algorithms

[2] https://fasttext.cc/docs/en/crawl-vectors.html
[3] https://github.com/facebookresearch/MUSE/

and context-based retrieval on structured knowledge bases can also be utilized in detecting correct entities for multilingual settings. Moussallem et al. [152] present a multilingual, knowledge-base agnostic and deterministic entity linking approach (coined as Multilingual AGDISTIS or MAG) which combines context-based retrieval and graph algorithms on structured knowledge bases. MAG does not require mono-lingual models. In another work, labels surrounding a graph entity can also be used to find a matching entity from another language with the help of machine translation. Such work is done by [153], in which context found in an RDF graph is used to find links between similar entities from different languages. This is done by creating virtual documents from the labels found in the neighboring nodes of compared RDF entities. Virtual documents are then translated to a similar language and then compared with string similarity measures. All the approaches summarized above require the data and the context and/or published with additional information in the RDF format. These approaches are however not applicable to budget and spending datasets. Budget and spending classification data tends to be published as a spreadsheet in tabular form. Classification concepts in fiscal data tend to be provided in short phrases, i.e., not provided as entities on a knowledge graph with surrounding labels and properties. Hence, attempts to interlink fiscal data concepts as done by other previous works by [151–153] are not feasible due to the lack of entity/semantics surrounding published fiscal concepts. We consequently choose to investigate the use of string similarity measures instead to create a mapping between translated open fiscal data concepts.

Ontology-based data integration

According to [154], Ontology-Based Data Integration (OBDI) refers to the use of ontologies to capture implicit knowledge from different data sources and obtain the semantic interoperability from these heterogeneous sources. Wache et al. [154] also state that ontology can be used to integrate data with several approaches: (1) Single ontology - requires an ontology to integrate data. (2) Multiple ontologies - requires mapping of concepts across used ontologies. (3) Hybrid - uses one ontology as a base for underlying multiple ontologies used in data integration. In fiscal data context, it may be possible to integrate all these datasets multilingually using OBDI if the following is available: (a) Ontologies that able to represent different types of fiscal classifications. (b) An approach to mapp the different types of fiscal classifications into concept instantiation/assertions (i.e., A-Box) based on the specific ontologies. (c) A method to handle multilingualism during assertion mappings.

Ontology-based data integration for heterogeneous datasets requires a well-defined ontology for our use case, i.e., open fiscal data. Specifically for the main open fiscal data itself,[4] OpenBudgets.eu (OBEU) ontology has been developed by [116] based on DCV, as elaborated in chapter 7. The ontology covers how money allocated/spent for budget/spending are represented (i.e., measure), how attributes (e.g., the currency) can be represented, and how dimensions (i.e., classifications) can be modeled. However, a

[4] i.e., the datasets with stated budgeted amount

specific ontology[5] for representing concepts from different open fiscal data classifications is not yet available. Designing this specific ontology requires an international collaborative effort from open fiscal data experts to develop the ontology as well as to ensure that the ontology can cover their classification requirements.

Instead of publishing open fiscal data in the RDF format that adopts ontologies, open fiscal data is published in a tabular format without semantics. The dataset is accompanied by controlled vocabularies in the form of classifications. These classifications are mostly independently published by local/national public administrators. Despite the availability of standardized classifications published by interstate/supranational organizations, these classifications are not adapted by local/national public administrators (see chapter 4), since adapting these vocabularies requires alignment of concepts from their business process flow. This alignment requires efforts, resources, and an approach to handle fiscal data complexity. As we know, fiscal classification deals with concrete controlled concepts, while ontologies deal more in an abstract and formal level, hence it requires more expertise to apply or even develop ontologies for this domain. This demonstrates that the creation of ontologies is not feasible for now. Moreover, since embedding semantics on publishing fiscal data requires a steep learning curve, these ontologies may not be used by the data publishers. This is because the development and application of these ontologies require training, substantial efforts, and resources which are often neither feasible activities nor a priority for these public administrations.

10.3.3 Data interlinking frameworks

SILK Framework [155] is a data interlinking framework which, among others, consists of different string similarity measure implementations. Some common uses of SILK Framework include (1) link generation among data items across different sources of linked data, and (2) data transformation of structured data. SILK Framework provides a GUI to build a data interlinking pipeline. The framework allows the query of data items from SPARQL endpoints, as well as obtain data items from structured data formats (e.g., CSV). String similarity measures are also implemented within SILK Framework as plugins, which range from sequence-based and set-based string similarity measures. A distance threshold can be specified in the string similarity experiment on SILK Framework. Using SILK Framework, an experiment is done by [156], utilizing string similarity measures to map links between Central Product Classification (CPC, published by the UN) and Classifications of Products by Activity (CPA, published by the EU and derived from CPC). Four similarity measures are used in their experiment namely: Dice, Jaro, Jaro-Winkler, and Soft Jaccard, Jaro provides the best precision value when the threshold (i.e., distance) is set at 0.0, and Dice provides the best recall with a threshold set at 0.5. Depending on the similarity measures and distance threshold configuration, the usage of SILK Framework may impose an out-of-memory problem as we experienced with our initial experiment (will be elaborated in subsection 10.5.2). The fiscal classification

[5] in the sense that available classes, axioms, and properties that correlates the terms within the ontology are formally defined.

concepts may result in a large number of comparisons that require scalability, which is not covered by SILK Framework during the time we perform our experiment.

10.4 Approach

The architecture overview of IOTA is shown in Figure 10.1. Given any two fiscal datasets, their labels or concepts are identified according to the provided classifications. These classifications act as blocks for comparisons and the labels belonging to different classifications are not compared, avoiding any unnecessary comparisons and optimizing the performance. Since the labels are provided in the regional languages, an essential step is to translate the concepts based on the classification pairs. We used Google Translate[6] for the translation of concepts. These translated concepts are then post-processed for case correction and stored in HDFS. The string matching module is executed within Apache Spark (Figure 10.3), and this module reads the data for parallel string matching from the HDFS. The parallel string similarity assessment in IOTA is achieved by using *py_stringmatching* library [127, 130], and the parameters used in IOTA are detailed in Table 10.5. IOTA utilizes 19 generic string similarity assessment algorithms taken from five different similarity measure categories. The inspected similarity measures in this experiment are shortly described in Table 10.2, Table 10.3, Table 10.4, for bag-based and phonetic-based, sequence-based, as well as set and hybrid-based, respectively.

Another perspective of Figure 10.1 can be seen in Figure 10.2, explaining parameters used within IOTA Framework. Open fiscal data comes with respective classifications, which ideally should be aligned before integrating the data. However, this is not the case. For linking the similar concepts, we use and benchmark several similarity measures after string preprocessing and machine translation steps. A minimum threshold of γ is set to filter the overall similarity values that are allowed to be inspected as links. To investigate the optimum similarity threshold values, we set an iterative threshold *t(i)* where $\gamma \leq t(i) \leq 1$. The filtered result is then analyzed and evaluated, which are then interlinked via the RDF SKOS ontology for interlinking similar multilingual concepts.

The translated and processed set of concepts stored in HDFS for the two datasets are represented internally as RDDs as shown in Figure 10.3. The RDD is a parallel collection of records that can be processed in a distributed, parallel manner to achieve scalability. The entities presented in the two RDDs are then cross-compared with each other for the similarity assessment. For concept matching, we use a regular expression before applying the selected similarity assessment to extract the exact location of the labels of interest from each label. The output of this parallel operation is an RDD with the entity pairs and their similarity score. The final step is to filter out the scores in the distributed RDD, based on the provided threshold. The filtered result is stored back to HDFS, which is used for evaluating the performance. The use of classification based translation, matching, and parallel in-memory processing differentiates IOTA from other state-of-the-art string comparison frameworks.

[6] https://translate.google.com/

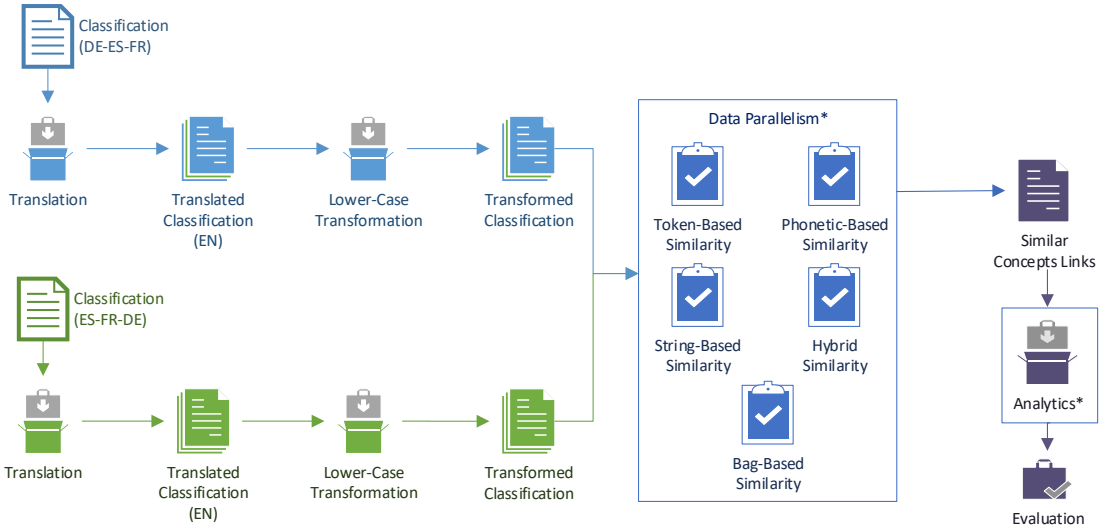

Figure 10.1: Our IOTA pipeline to map similar concepts from translated classification. Preprocessed, translated classifications from different languages and public administration are measured for their similarity scores. *The similarity measure comparison and analytics process utilizes Apache Spark for scalability.

10.5 Experiment and evaluation

In this chapter, we elaborate on the details regarding our experiments and evaluation. We begin with datasets used and evaluation metrics, followed by experiment configuration and result, both using SILK Framework as well as IOTA Framework.

10.5.1 Dataset and evaluation metrics

For the experiment, we use the European Union official procurement classification, CPV classification [24]. CPV is published in 24 different European languages. This dataset is comprised of 9454 concepts. Each concept in any language is associated with a unique key. Hence, the key can be used to identify a proper match between concepts.

The experiment starts with translating concepts from different languages using Google Translate. Three datasets originally from German, Spanish and French are translated into English. The translation result is then paired in three language pairs as shown in Table 10.6: German-Spanish, German-French, and Spanish-French.

The mappings between the two classifications are evaluated using recall, precision, and F-measure. To compute these measures, we calculate true positive, false positive, false negative, and true negative values by comparing the assigned and original values.

The basic combinations of actual vs assigned data category is a well-known concept in binary classification. *True positive* (*tp*) indicates the number of retrieved information that classified as correct and actually belongs to the correct result. In our case, the possible number of true positives is equal to the number of concepts in the datasets we

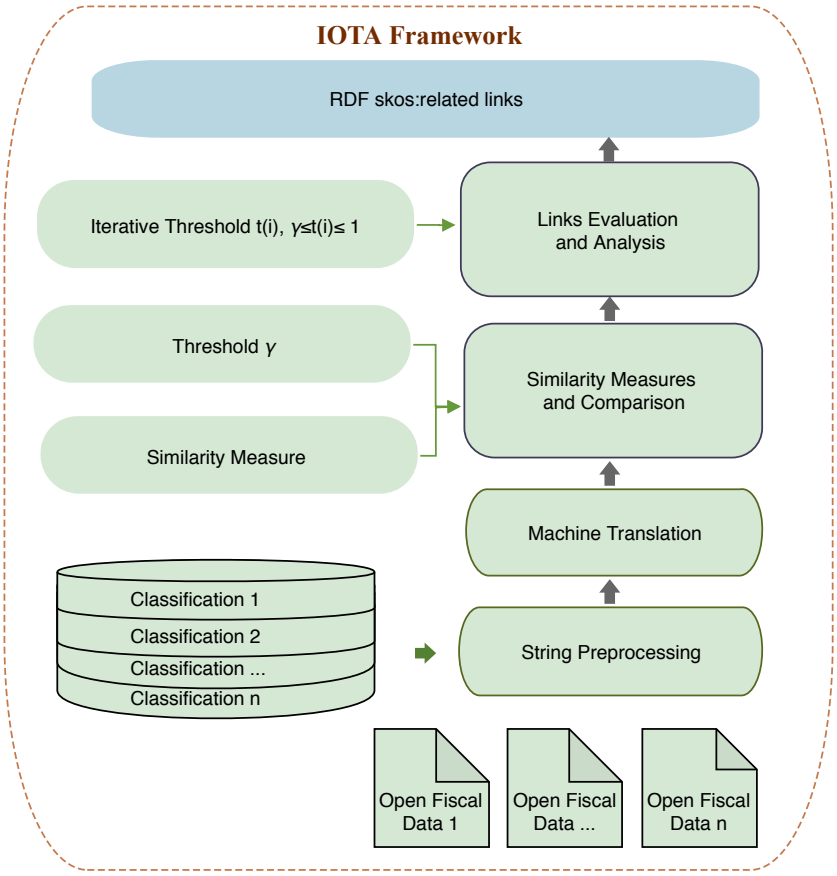

Figure 10.2: IOTA Framework takes out classification labels from different languages, as well as specific similarity measures and minimum threshold that can limit the similar string estimation. Later, we iterate from the minimum passing similarity threshold from γ to 1, to check which thresholds yield the highest F-Measure.

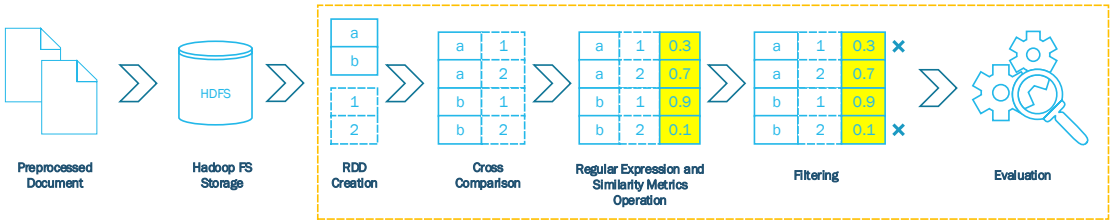

Figure 10.3: Distributed processing pipeline that we perform in our IOTA experiment. Preprocessed classification documents are stored within Hadoop FS, then Apache-Spark operations follow the next step: creating RDD data types out of stored documents, performing the cross computation, getting similarity score between concepts, filtering the scores and finished by evaluating the result.

Type	Similarity measure	Norm. range	Score ranging from [0-1] by default?	Non-norm. Score Range	Extra param. required?	Type of extra parameter	Default value (respectively)
Bag-based	TF/IDF	[0,1]	Yes	-	Yes	Corpus (list containing lists), Dampening (true/false)	None (only use tokens from compared strings), True
Hybrid-based	Generalized Jaccard	[0,1]	Yes	-	Yes	Similarity measure, Similarity threshold	Jaro, 0.5
	Monge-Elkan	[0,1]	Yes	-	Yes	Similarity measure	Jaro-Winkler
	Soft TF/IDF	[0,1]	Yes	-	Yes	Corpus (list containing lists), Similarity measure, Similarity threshold	None (only use tokens from compared strings), Jaro, 0.5
Phonetic-based	Soundex	[0/1]	Yes	-	No	-	-
	Editex	[0,1]	No	Integer	Yes	Match cost (weight when the correct char match), Group cost (weight when the char is in the same Editex group), Mismatch cost (weight when the char match incorrect), Local variant	0, 1, 2, False
Sequence-based	Bag distance	[0,1]	No	Integer	No	-	-
	Jaro*	[0,1]	Yes	-	No	-	-
	Jaro-winkler*	[0,1]	Yes	-	Yes	Prefix weight (weight for the prefix)	0.1
	Levenshtein*	[0,1]	No	Integer	No	-	-
	Partial Ratio	[0,1]	No	[0, 100]	No	-	-
	Ratio	[0,1]	No	[0, 100]	No	-	-
	Partial token sort	[0,1]	No	[0, 100]	Yes	Force ASCII (boolean to remove non-ASCII characters), Full process (boolean for preprocessing such as lower case transformation as well as removing leading/trailing white spaces)	True, True
	Token Sort	[0,1]	No	[0, 100]	Yes	Force ASCII (boolean to remove non-ASCII characters), Full process (boolean for preprocessing such as lower case transformation as well as removing leading/trailing white spaces)	True, True
Set-based	Cosine	[0,1]	Yes	-	No	-	-
	Dice	[0,1]	Yes	-	No	-	-
	Jaccard	[0,1]	Yes	-	No	-	-
	Overlap Coefficient	[0,1]	Yes	-	No	-	-
	Tversky Index	[0,1]	Yes	-	No	-	-

Table 10.5: The list of different similarity measures used within the IOTA framework. Similarity measures marked with asterisks (*) indicate a cythonized implementation in the *py_ stringmatching* library that speeds up the performance. The similarity score range from 0 to 1 for most similarity scores, except for Soundex similarity, which provides a true or false decision. Most of our experiments use default parameter values provided by the library, except for TF-IDF and Soft TF-IDF, in which we are using a corpus from the whole translated words instead of only compared, translated words.

Language Pairs	
German	Spanish
Spanish	French
French	German

Table 10.6: Language pairs used for our experiment. The pairs are chosen based on the availability in the datasets (Common Procurement Vocabulary by European Union) and how wide the EU languages are used.

are experimenting with. *False positive* (*fp*) indicates the number of the incorrect result but are classified as a correct result. *False negative* (*fn*) indicates information that is classified as false but it is not actually false. False-negative is computed based on the possible number of true positive links minus true positive links that are found, so

$$fn = |concept| - tp.$$

True negatives (*tn*) are the number of classes that are classified as false and are actually false. True negative is a result of the subtraction of Cartesian product cardinality between two sets in the compared concepts with the sum of true positives, false positives, and false negatives, so

$$tn = (|concept_1| \times |concept_2|) - (tp + fp + fn).$$

F-Measure is then calculated as a harmonic mean of precision and recall.

$$F - Measure = 2 * \frac{precision * recall}{precision + recall}$$

- *Recall* indicates how many correct items can be retrieved out of possible correct classes. Recall is also referred to as *sensitivity*.

$$recall = \frac{tp}{tp + fn}$$

- *Precision* indicates the portion of the retrieved concepts that are really relevant.

$$precision = \frac{tp}{tp + fp}$$

- *F-Measure* is a harmonic mean of precision and recall.

$$F - Measure = 2 * \frac{precision * recall}{precision + recall}$$

In this chapter, we attempt to answer the following four research questions ($\mathcal{RQ}$):

- $\mathcal{RQ}1$. Which string similarity measures provide the highest F-Measure in interlinking fiscal classification concepts?

- $\mathcal{RQ}2$. What is the impact of applying a similarity threshold for interlinking concepts between translated classification?

- $\mathcal{RQ}3$. How robust is the similarity measure performance when the language pairs are changed?

- $\mathcal{RQ}4$. Different similarity measures have different computational performance. Which similarity measures have faster computational performance, and is there any trade-off between faster computational performance and the resulting F-Measure?

10.5.2 Experimental configuration and result

We provide the details of the experiment configuration and experiment result for both SILK framework and IOTA Framework in this section. In the last parts of the section, we discuss the result of our IOTA Framework experiment.

SILK framework experiment configuration

In the initial experiment, we use German and Spanish concept from CPV classification. Computing similarity between translated string is done at first by utilizing SILK Framework which implements sequence-based similarity matching (*Jaro*, *Jaro Winkler*, *Levenshtein*, *Normalized Levenshtein*, *qGrams* and *Substring*) as well as set-based similarity matching (*Token-wise*, *Soft-Jaccard*, *Dice* and *Jaccard*). Other than performing tokenization for set-based similarity measure and changing the distance threshold, we use default parameters in SILK Framework. A comparison of strings leads to a *distance* threshold which is defined as the maximum distance two strings allowed to have. The more distance threshold value is set, the more links can be found, but there are more false-positive links discovered. The experiment result is then stored as an ontology alignment XML format,[7] which later converted to CSV and then processed for analytics using a python script. We use the latest stable version of SILK Framework v2.7.1, at the time of our experiment.

SILK framework experiment result

From our experiment using SILK Framework, the similarity measure that yields the biggest F-Measure score is Substring, with 0.501 F-Measure scores as the distance threshold is set to 0.2. In our particular use case, other similarity measures that provide a relatively good F-Measure score are qGrams (F-Measure = 0.453, distance threshold = 0.4) and Soft Jaccard (F-Measure = 0.446, distance threshold = 0.4). The result of the experiment using the SILK framework is provided in detail on Table 10.7. The corresponding F-Measure chart for the SILK experiment is provided in Figure 10.4.

On low distance thresholds (i.e. 0.0), it is fast to use SILK Framework for most similarity measures, except for some similarity measures that are failed at the 0.0 distance threshold. On Table 10.7, the fields that are marked as an asterisk (*) in the table indicate that those fields yield out of memory error during the experiment, hence an increase in the higher distance threshold can not be done. On the other hand, there is a problem with Cosine similarity measure during our experiment, and we can not proceed with any of the thresholds, which is indicated with dash (-) in Table 10.7. This limitation prevents our further experiment using higher distance thresholds. As a result, in similarity experiments

[7] http://alignapi.gforge.inria.fr/format.html

using SILK Framework, only limited thresholds can be presented. For this reason, we deviated from using SILK Framework for further experiments and continue with our IOTA framework for the experiment as we described in section 10.4.

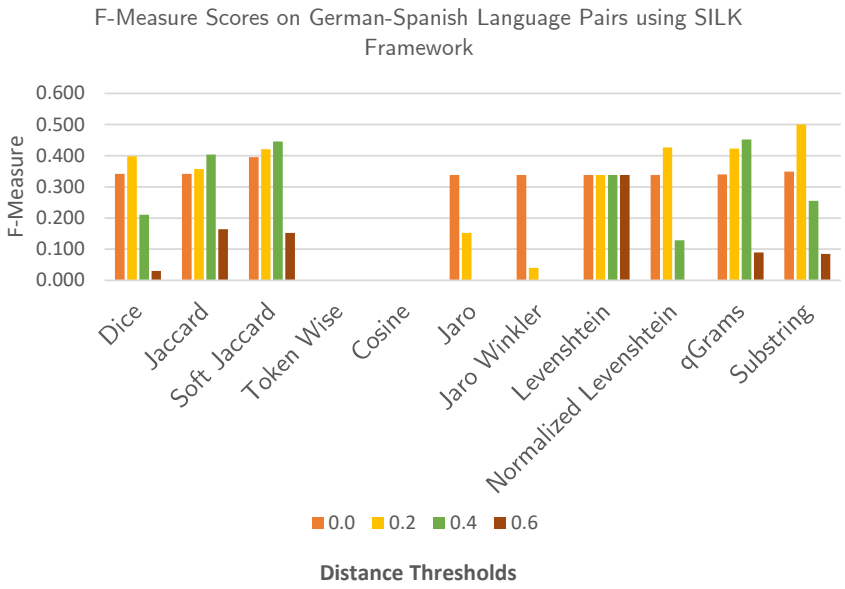

Figure 10.4: F-Measure chart of different similarity measures and *distance* thresholds experimented using SILK Framework. A blank space in the diagram indicates the unavailability of the F-Measure value for that particular similarity measure/filter mostly due to scalability reasons. In this comparison, *Substring* yields the highest F-Measure score, followed by *qGrams* and *Jaccard*.

IOTA Framework Evaluation Configuration

The evaluation is conducted on a cluster of three workstations, each consists of 256 GB RAM, and each has four AMD Optheron^{TM} 6376 2.3 GHz processors. Each processor has 16 cores, totaling 64 cores in each workstation. One workstation is used as a Spark driver, and two others are used as Spark workers. We use Apache Spark 2.3.1 on our cluster during our experiment.

IOTA Framework Experiment Result

IOTA Framework provides several experiment results. Execution time for each language pair in the cluster is compared in Figure 10.5. We present the result of F-Measure values from each language pairs experiment in Table 10.8, Table 10.9, and Table 10.10, respectively The more intense the color of the cell within those tables, the higher the F-Measure values are. The summarized top-10 F-Measure score for each similarity measure and the filter is summarized in Table 10.11. The charts for these F-Measure scores from Table 10.8 to Table 10.10 are provided in Figure 10.7, Figure 10.8, and Figure 10.9. The radar chart of the aggregated average score is provided in Figure 10.10, and broken down

Similarity Type	Similarity Measure	Distance Threshold	Found	TP	FP	FN	TN	Precision	Recall	F-Measure
Set-based	Dice	0.0	2,207	1,995	212	7,459	89,368,450	0.904	0.211	0.342
		0.2	3,952	2,671	1,281	6,783	89,367,381	0.676	0.283	0.398
		0.4	31,543	4,315	27,228	5,139	89,341,434	0.137	0.456	0.211
		0.6	417,618	6,531	411,087	2,923	88,957,575	0.016	0.691	0.031
	Jaccard	0.0	2,207	1,995	212	7,459	89,368,450	0.904	0.211	0.342
		0.2	2,372	2,114	258	7,340	89,368,404	0.891	0.224	0.358
		0.4	5,479	3,017	2,462	6,437	89,366,200	0.551	0.319	0.404
		0.6	46,383	4,597	41,786	4,857	89,326,876	0.099	0.486	0.165
	Soft Jaccard	0.0	2,844	2,432	412	7,022	89,368,250	0.855	0.257	0.396
		0.2	3,179	2,658	521	6,796	89,368,141	0.836	0.281	0.421
		0.4	7,237	3,722	3,515	5,732	89,365,147	0.514	0.394	0.446
		0.6	63,112	5,545	57,567	3,909	89,311,095	0.088	0.587	0.153
	Token Wise	0.0	*	*	*	*	*	*	*	*
	Cosine	0.0	-	-	-	-	-	-	-	-
Sequence-based	Jaro	0.0	2,179	1,968	211	7,486	89,368,451	0.903	0.208	0.338
		0.2	47,324	4,324	43,000	5,130	89,325,662	0.091	0.457	0.152
		0.4	*	*	*	*	*	*	*	*
	Jaro Winkler	0.0	2,179	1,968	211	7,486	89,368,451	0.903	0.208	0.338
		0.2	255,495	5,376	250,119	4,078	89,118,543	0.021	0.569	0.041
		0.4	*	*	*	*	*	*	*	*
	Levenshtein	0.0	2,179	1,968	211	7,486	89,368,451	0.903	0.208	0.338
		0.2	2,179	1,968	211	7,486	89,368,451	0.903	0.208	0.338
		0.4	2,179	1,968	211	7,486	89,368,451	0.903	0.208	0.338
		0.6	2,179	1,968	211	7,486	89,368,451	0.903	0.208	0.338
	Normalized Levenshtein	0.0	2,179	1,968	211	7,486	89,368,451	0.903	0.208	0.338
		0.2	4,922	3,066	1,856	6,388	89,366,806	0.623	0.324	0.427
		0.4	63,124	4,674	58,450	4,780	89,310,212	0.074	0.494	0.129
		0.6	*	*	*	*	*	*	*	*
	qGrams	0.0	2,192	1,981	211	7,473	89,368,451	0.904	0.210	0.340
		0.2	3,001	2,635	366	6,819	89,368,296	0.878	0.279	0.423
		0.4	9,836	4,365	5,471	5,089	89,363,191	0.444	0.462	0.453
		0.6	136,028	6,500	129,528	2,954	89,239,134	0.048	0.688	0.089
	Substring	0.0	2,262	2,043	219	7,411	89,368,443	0.903	0.216	0.349
		0.2	7,907	4,347	3,560	5,107	89,365,102	0.550	0.460	**0.501**
		0.4	35,881	5,791	30,090	3,663	89,338,572	0.161	0.613	0.255
		0.6	150,071	6,790	143,281	2,664	89,225,381	0.045	0.718	0.085

Table 10.7: Different similarity measures performance for mapping concepts originally from German and Spanish datasets using SILK Framework. Asterisk (*) sign indicates out of memory error, hence these algorithms are not scalable, while dash (-) indicates other errors during the experiment. Several similarity measures here are not robust to the change of distance thresholds.

into (1) bag-based, hybrid-based, and phonetic-based similarity measures; (2) sequence-based similarity measures; (3) set-based similarity measures. These figures and tables are used to highlight the findings from our experiment.

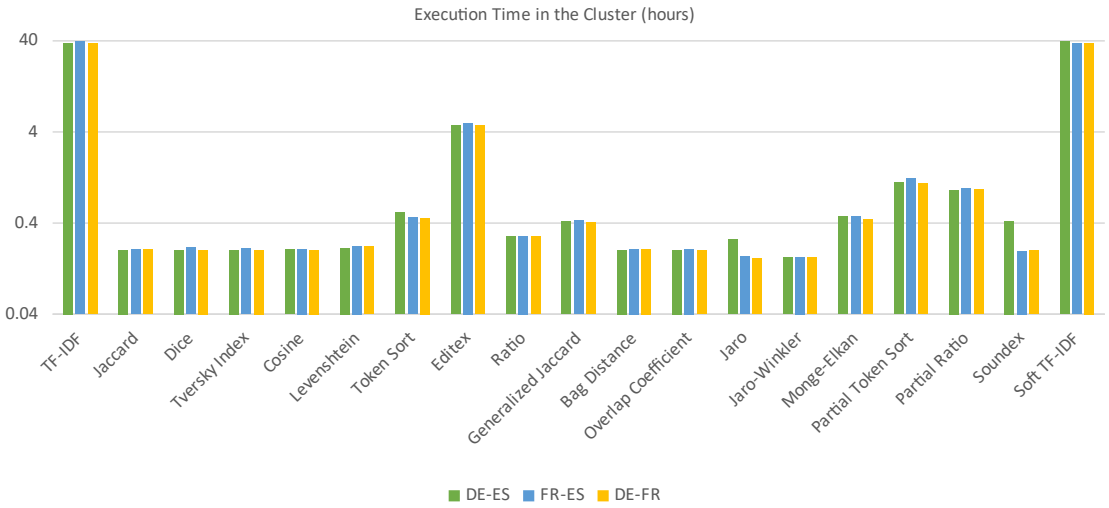

Figure 10.5: Execution time (hours, on a logarithmic scale) in our cluster. The cluster performs more than 89 million string comparisons. TF-IDF and Soft TF-IDF similarity measure have the longest execution time due to their complexity. Most of the other similarity measures provide decent computational performance.

10.5.3 Discussion

Our experiments with the IOTA framework show that finding a similar concept is reliable and scalable for all thresholds, even though some similarity measures used within IOTA need more time for the computation. Token Sort provides the IOTA framework the highest F-Measure score when the similarity threshold is estimated properly. TF-IDF provides a high average F-Measure score which is robust across similarity threshold change, yet TF-IDF needs significant computational resources which we discuss in detail on the following subsections.

Performance evaluation

In the discussion section, we categorize the similarity measures into three main categories: (1) bag-based, hybrid-based and phonetic-based similarity measures (2) sequence-based similarity measures and (3) set-based similarity measures. For each category, the F-Measure score is averaged by language pairs.

Hybrid similarity measures are designed to consider misspelled tokens [130]. In the hybrid similarity measure category, Generalized Jaccard provides the best F-Measure but it is sensitive to similarity threshold values. In this experiment, we use default parameter values in the *py_stringmatching* library (see Table 10.5), which uses Jaro as

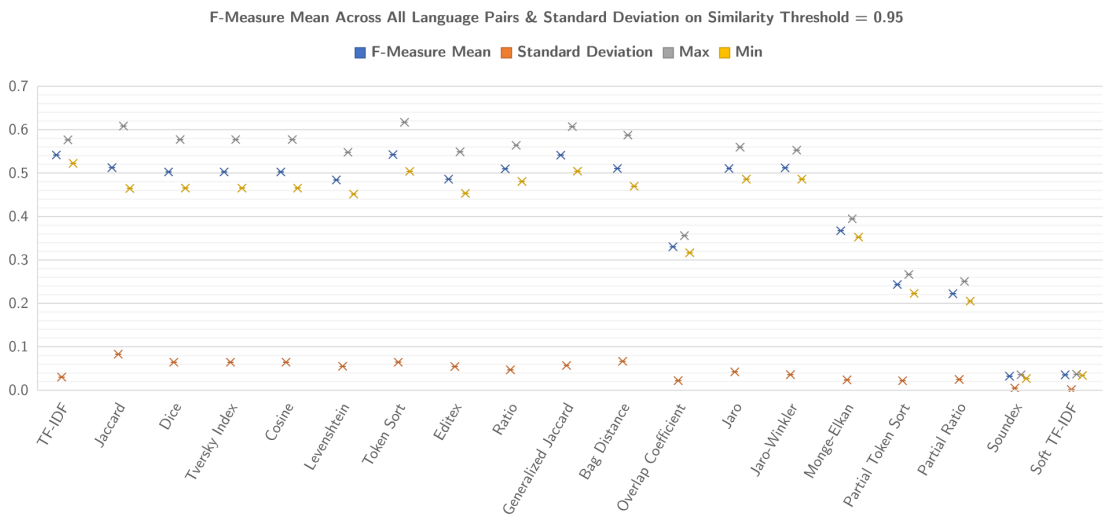

Figure 10.6: The plot for average F-Measure score, minimum F-measure score, maximum F-measure score, and sample standard deviation for each language as similarity threshold set to 0.95. Even though the TF-IDF similarity score takes a long time to compute, it has the minimum standard deviation with a relatively good F-Measure score compared to other similarity measures. On the other hand, Token Sort yields the maximum F-Measure and needs much less computational time compared to TF-IDF, but it has a high standard deviation.

German - Spanish											
Similarity Threshold	**0.50**	**0.55**	**0.60**	**0.65**	**0.70**	**0.75**	**0.80**	**0.85**	**0.90**	**0.95**	**1.00**
TF-IDF	0.280	0.337	0.388	0.437	0.479	0.511	0.532	0.543	0.540	0.522	0.455
Jaccard	0.241	0.364	0.424	0.487	0.496	0.501	0.494	0.476	0.466	0.465	0.465
Dice	0.067	0.152	0.191	0.241	0.363	0.424	0.487	0.501	0.482	0.465	0.465
Tversky Index	0.067	0.152	0.191	0.241	0.363	0.424	0.487	0.501	0.482	0.465	0.465
Cosine	0.092	0.145	0.205	0.240	0.335	0.422	0.498	0.501	0.482	0.465	0.457
Levenshtein	0.040	0.075	0.122	0.202	0.292	0.391	0.466	0.493	0.486	0.452	0.441
Token Sort	0.011	0.023	0.048	0.093	0.178	0.311	0.460	0.545	**0.548**	0.507	0.459
Editex	0.014	0.031	0.059	0.109	0.183	0.285	0.409	0.472	0.482	0.453	0.441
Ratio	0.012	0.023	0.043	0.077	0.143	0.244	0.376	0.478	0.510	0.485	0.441
Generalized Jaccard	0.003	0.004	0.009	0.021	0.047	0.100	0.219	0.377	0.482	0.513	0.465
Bag Distance	0.001	0.001	0.002	0.004	0.011	0.036	0.148	0.415	0.503	0.474	0.455
Overlap Coefficient	0.015	0.078	0.079	0.092	0.182	0.191	0.252	0.315	0.317	0.317	0.317
Jaro	0.000	0.000	0.001	0.004	0.019	0.058	0.133	0.296	0.463	0.487	0.441
Jaro-Winkler	0.000	0.000	0.001	0.003	0.010	0.023	0.035	0.070	0.249	0.486	0.441
Monge-Elkan	0.000	0.000	0.001	0.001	0.003	0.008	0.021	0.061	0.171	0.352	0.360
Partial Token Sort	0.001	0.003	0.006	0.011	0.022	0.036	0.078	0.164	0.221	0.223	0.216
Partial Ratio	0.001	0.003	0.006	0.011	0.021	0.032	0.068	0.141	0.195	0.205	0.202
Soundex	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027	0.027
Soft TF-IDF	0.001	0.001	0.002	0.002	0.003	0.005	0.008	0.013	0.021	0.034	0.042

Table 10.8: F-Measure values of different string similarity measures for mapping concepts originally from German and Spanish datasets. TF-IDF, Jaccard, and Dice have the best F-Measure scores when it is averaged by the similarity thresholds. Token-Sort provides the best F-Measure score as the similarity threshold is set to 0.90.

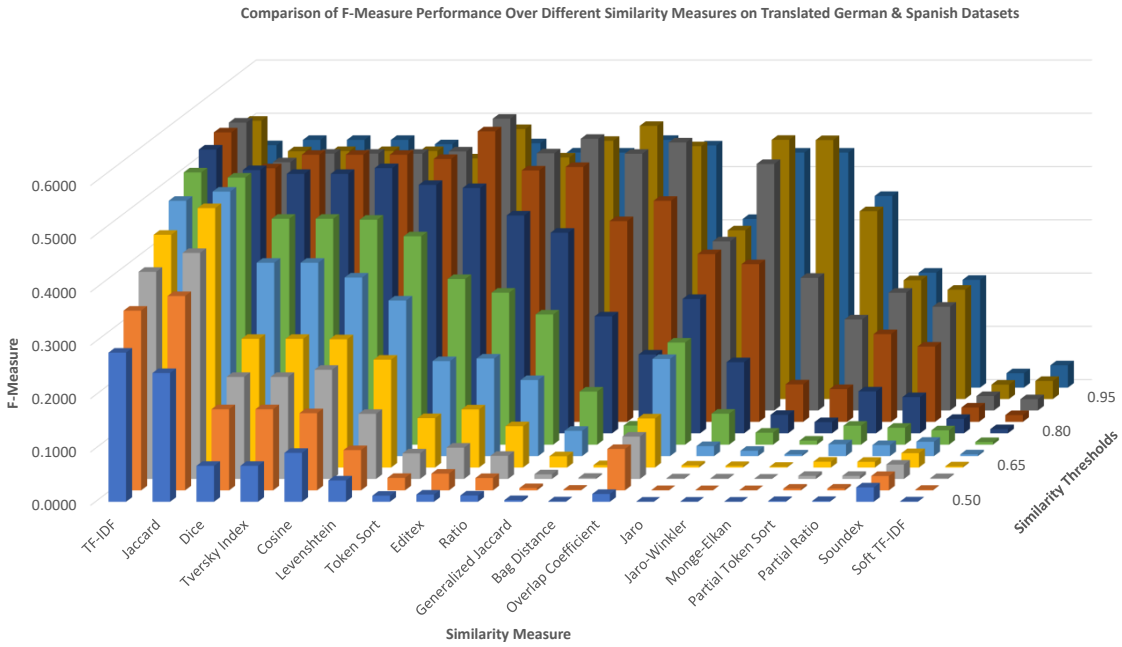

Figure 10.7: F-Measure chart of different similarity measures and filters for matching strings between translated German and Spanish datasets, as shown in Table 10.8. The performance reaches a peak as the similarity threshold is set to 0.90.

German - French											
Similarity Threshold	**0.5**	**0.55**	**0.6**	**0.65**	**0.7**	**0.75**	**0.8**	**0.85**	**0.9**	**0.95**	**1**
TF-IDF	0.283	0.340	0.393	0.442	0.485	0.517	0.537	**0.548**	0.544	0.526	0.456
Jaccard	0.282	0.413	0.468	0.513	0.511	0.512	0.497	0.478	0.466	0.465	0.465
Dice	0.073	0.178	0.234	0.282	0.412	0.468	0.513	0.512	0.484	0.466	0.465
Tversky Index	0.073	0.178	0.234	0.282	0.412	0.468	0.513	0.512	0.484	0.466	0.465
Cosine	0.106	0.169	0.242	0.281	0.371	0.466	0.515	0.512	0.485	0.466	0.456
Levenshtein	0.042	0.080	0.131	0.220	0.311	0.404	0.470	0.492	0.482	0.453	0.445
Token Sort	0.011	0.024	0.049	0.095	0.182	0.319	0.466	**0.548**	0.544	0.504	0.459
Editex	0.013	0.032	0.061	0.115	0.194	0.298	0.418	0.475	0.480	0.455	0.445
Ratio	0.012	0.023	0.045	0.081	0.152	0.263	0.395	0.486	0.502	0.481	0.445
Generalized Jaccard	0.003	0.004	0.009	0.021	0.048	0.102	0.229	0.394	0.486	0.504	0.465
Bag Distance	0.001	0.001	0.002	0.005	0.011	0.037	0.155	0.425	0.503	0.470	0.454
Overlap Coefficient	0.015	0.091	0.093	0.104	0.215	0.224	0.268	0.319	0.319	0.319	0.319
Jaro	0.000	0.000	0.001	0.004	0.020	0.073	0.170	0.345	0.483	0.486	0.445
Jaro-Winkler	0.000	0.000	0.001	0.003	0.011	0.027	0.046	0.100	0.309	0.497	0.445
Monge-Elkan	0.000	0.000	0.001	0.001	0.003	0.008	0.022	0.067	0.187	0.355	0.363
Partial Token Sort	0.001	0.003	0.006	0.011	0.022	0.034	0.080	0.184	0.260	0.266	0.261
Partial Ratio	0.001	0.003	0.005	0.010	0.019	0.030	0.067	0.154	0.232	0.251	0.249
Soundex	0.034	0.034	0.034	0.034	0.034	0.034	0.034	0.034	0.034	0.034	0.034
Soft TF-IDF	0.001	0.001	0.002	0.002	0.003	0.005	0.008	0.013	0.022	0.035	0.043

Table 10.9: F-Measure values of different string similarity measures for mapping concepts originally from German and French datasets. Token Sort remains the similarity measure that yields the highest F-Measure, although the optimum similarity threshold is 0.85 instead instead of 0.90.

French - Spanish											
Similarity Threshold	**0.50**	**0.55**	**0.60**	**0.65**	**0.70**	**0.75**	**0.80**	**0.85**	**0.90**	**0.95**	**1.00**
TF-IDF	0.297	0.360	0.420	0.474	0.523	0.566	0.598	0.615	0.618	0.608	0.564
Jaccard	0.227	0.396	0.452	0.556	0.595	0.605	0.605	0.587	0.578	0.577	0.577
Dice	0.055	0.135	0.178	0.227	0.395	0.452	0.556	0.605	0.592	0.577	0.577
Tversky Index	0.055	0.135	0.178	0.227	0.395	0.452	0.556	0.605	0.592	0.577	0.577
Cosine	0.080	0.128	0.192	0.226	0.356	0.450	0.568	0.605	0.593	0.577	0.562
Levenshtein	0.030	0.060	0.102	0.177	0.266	0.375	0.493	0.559	0.564	0.548	0.536
Token Sort	0.009	0.018	0.036	0.068	0.132	0.249	0.426	0.585	**0.645**	0.617	0.571
Editex	0.010	0.023	0.046	0.091	0.160	0.259	0.407	0.521	0.560	0.549	0.536
Ratio	0.009	0.017	0.032	0.061	0.117	0.214	0.357	0.501	0.571	0.564	0.536
Generalized Jaccard	0.003	0.004	0.009	0.021	0.044	0.093	0.213	0.397	0.558	0.607	0.577
Bag Distance	0.001	0.001	0.002	0.004	0.011	0.034	0.141	0.442	0.598	0.587	0.566
Overlap Coefficient	0.011	0.072	0.073	0.085	0.179	0.187	0.286	0.355	0.356	0.356	0.356
Jaro	0.000	0.000	0.001	0.003	0.018	0.069	0.154	0.321	0.520	0.560	0.536
Jaro-Winkler	0.000	0.000	0.001	0.003	0.010	0.026	0.047	0.103	0.286	0.553	0.536
Monge-Elkan	0.000	0.000	0.001	0.001	0.003	0.007	0.018	0.056	0.172	0.395	0.424
Partial Token Sort	0.001	0.003	0.005	0.009	0.018	0.030	0.068	0.151	0.227	0.240	0.234
Partial Ratio	0.001	0.003	0.005	0.008	0.016	0.025	0.053	0.114	0.189	0.211	0.208
Soundex	0.036	0.036	0.036	0.036	0.036	0.036	0.036	0.036	0.036	0.036	0.036
Soft TF-IDF	0.001	0.001	0.002	0.002	0.003	0.005	0.008	0.014	0.023	0.037	0.051

Table 10.10: F-Measure chart of different similarity measures and filters for matching strings between translated French and Spanish datasets. The French - Spanish language pair yields the highest F-Measure (0.645) compared to previous language pairs (0.548 for both of previous language pairs). Token Sort remains the best performing algorithms when the similarity threshold is properly set.

French-Spanish						German-French						German-Spanish					
Rk.	**FM**	**Prec.**	**Rec.**	**Similarity Measure**	**Thres.**	**Rk.**	**FM**	**Prec.**	**Rec.**	**Similarity Measure**	**Thres.**	**Rk.**	**FM**	**Prec.**	**Rec.**	**Similarity Measure**	**Thres.**
1	0.6450	0.8316	0.5269	Token Sort	0.90	1	0.5483	0.6501	0.4741	Token Sort	0.85	1	0.5476	0.8258	0.4096	Token Sort	0.90
2	0.6175	0.8763	0.4767	TF-IDF	0.90	2	0.5479	0.7684	0.4257	TF-IDF	0.85	2	0.5452	0.6425	0.4735	Token Sort	0.85
3	0.6170	0.9150	0.4654	Token Sort	0.95	3	0.5444	0.8486	0.4008	TF-IDF	0.90	3	0.5432	0.7802	0.4166	TF-IDF	0.85
4	0.6153	0.8048	0.4980	TF-IDF	0.85	4	0.5442	0.8187	0.4076	Token Sort	0.90	4	0.5402	0.8554	0.3948	TF-IDF	0.90
5	0.6084	0.9171	0.4552	TF-IDF	0.95	5	0.5366	0.6626	0.4508	TF-IDF	0.80	5	0.5325	0.6684	0.4425	TF-IDF	0.80

Table 10.11: Top five F-Measure (FM), Precision (Prec.), Recall (Rec.) scores and their corresponding similarity measure and thresholds (Thresh.). Token Sort and TF-IDF yield the highest F-Measure scores when the similarity thresholds is set from 0.80 upwards.

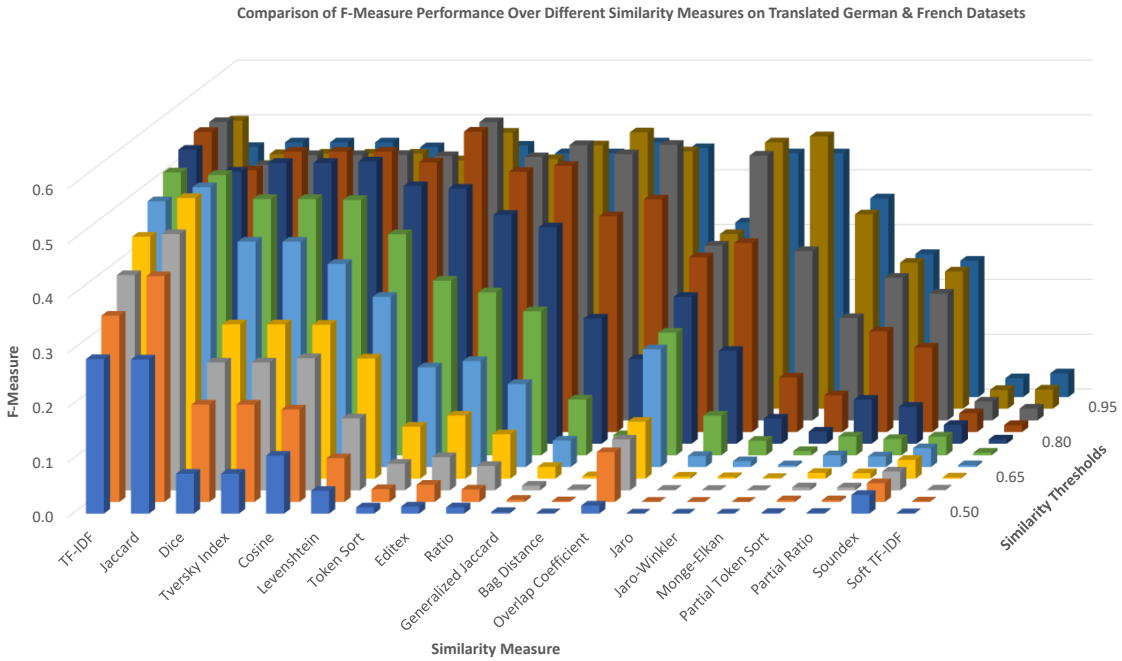

Figure 10.8: F-Measure chart of different similarity measures and filters for matching strings between translated German and French datasets, as shown in Table 10.9. There is no significant difference as compared to the chart from German-Spanish dataset (Figure 10.7).

secondary string similarity measure and 0.5 as the similarity threshold. The result can potentially be improved if a better performing, secondary, string similarity measure is chosen, for example, Levenshtein instead of Jaro (see Tables and Diagrams for the IOTA Framework experiment result in subsection 10.5.2). The same tuning can also be made for Monge-Elkan and Soft TF-IDF to improve their result. The only bag-based similarity measure, TF-IDF, provides better F-Measure value than Generalized Jaccard and other hybrid-based similarity measures. TF-IDF can capture insignificant words from the set of given corpus, and use it to make the result more relevant.

Phonetic-based algorithms are intended to match similarly sounding words. Intuitively, phonetic algorithms do not fit in our particular use case for matching translated multilingual concepts. As can be seen in the result section, Soundex has low F-Measure scores across language pairs due to its binary distinction of similar concepts. Editex, despite belongs to the phonetic algorithm category, provides a much higher F-Measure score as compared to Soundex for phonetic similarity measures. This might be because Editex improves the character grouping on Soundex and combines it with Levenshtein-like similarity measure (see Table 10.2), making the Soundex similarity measures are much less restrictive. The difference between the two is highlighted in Figure 10.10(a), which shows a low average F-Measure score for Soundex, and in contrast, Editex is surprisingly decent for this task.

The best F-Measure score for the sequence-based similarity measure category is provided

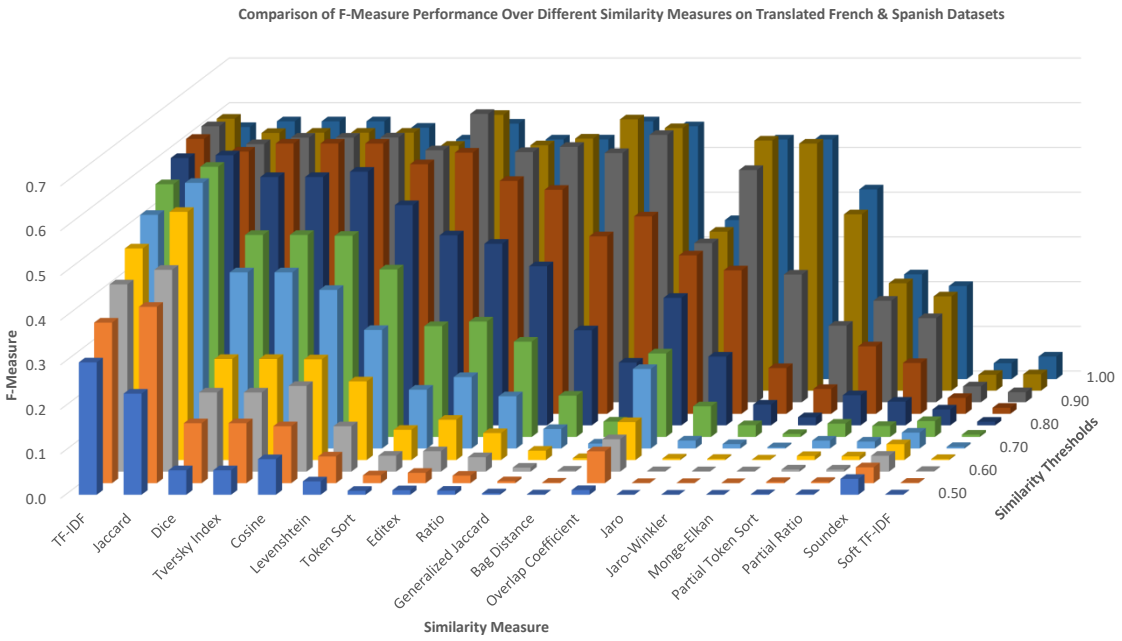

Figure 10.9: F-Measure chart of different similarity measures and filters for matching strings between translated French and Spanish datasets, as shown in Table 10.10. It consistently has similar patterns with Figure 10.7 and Figure 10.8 regarding similarity thresholds and similarity measures.

by Token Sort, as illustrated in Figure 10.10(b). In our use case, this makes sense because the result of the translation may end up as similar words, but arranged in a different phrase. After removing non-ASCII characters, removing trailing white spaces, and sorting these phrases, ratio similarity measure counts the ratio between matching elements and total elements from both compared strings. In overall algorithm category, Token Sort provides the best F-Measure score across all the language pairs we have experimented with, as shown in Table 10.11, as well as the diagrams in Figure 10.7, Figure 10.8, and Figure 10.9.

In the set-based similarity measure category, many similarity measures share close results as can be seen from Figure 10.10(c). This may be due to the similar formulation of the set-based similarity measures, which are mostly based on the size of shared tokens. The highest average F-Measure value across three language pairs, 0.54, is shared by many set-based similarity measures, such as Cosine-Ochiai, Dice, and Tversky Index, when the filter configuration is set to 0.85. The same F-Measure value (0.54) is also found on Jaccard when the filter value is set to be 0.75.

Similarity thresholds

Generally, the effective similarity threshold values, i.e., the filter that provides a consistently good F-Measure value could not be determined as it can be seen from Table 10.8,

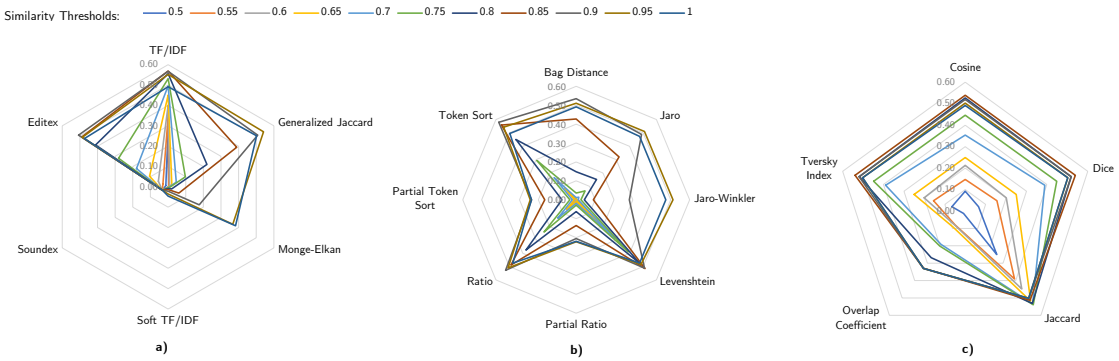

Figure 10.10: Average F-Measure on different similarity measures and filters across three language pairs. The further the threshold line from the center of the polygon, the more F-Measure score it has. (a) Bag-/Hybrid-/Phonetic-based similarity measures have various performance. TF-IDF provides a robust performance against most similarity thresholds while Editex and Generalized Jaccard also provide a decent performance but not robust to threshold change. (b) Sequence-based similarity measures are sensitive to threshold change, with Token Sort provides the highest F-Measure. (c) Set-based similarity measures yield similar performance, and are also sensitive to threshold change.

Table 10.9, and Table 10.10. The values of the F-Measure score in these tables show that some similarity measures are sensitive to similarity threshold values, except, for example, TF-IDF and Jaccard which still provide a relatively high F-Measure score across different similarity thresholds values, and Soft TF-IDF which provides low F-Measure score values across similarity thresholds. Figure 10.6 shows that Token Sort has a high standard deviation and sensitive to similarity threshold change, but also yield the highest F-Measure on three language pairs. Token Sort is sensitive to change on similarity thresholds. It is also observed that the highest filter score does not guarantee that the F-Measure will be higher, but the F-Measure score tends to be higher with the higher filter score.

Language pairs

There is a difference in F-Measure performance across different translated language pairs. The best language pair in our experiment is FR-ES, followed by DE-FR, and at last DE-ES. The correlation across different language pairs are very high, as illustrated in four different similarity threshold in Figure 10.11(a), Figure 10.11(b), Figure 10.11(c), and Figure 10.11(d) for similarity threshold = 0.85, 0.90, 0.95 and 1.00 respectively. The lowest Spearman correlation value between all the language pairs is between German-French and French-Spanish with a value of 0.882 when the similarity threshold is set to 0.95. This indicates that regardless of language pairs, using the language pairs we experiment with, the resulting F-measure score from the combination of similarity measures and threshold matrix stays highly correlated.

Sim. Threshold: 0.85			
	DE-ES	DE-FR	FR-ES
DE-ES	1.00	1.00	0.96
DE-FR	1.00	1.00	0.96
FR-ES	0.96	0.96	1.00

(a)

Sim. Threshold: 0.9			
	DE-ES	DE-FR	FR-ES
DE-ES	1.00	0.92	0.90
DE-FR	0.92	1.00	0.91
FR-ES	0.90	0.91	1.00

(b)

Sim. Threshold: 0.95			
	DE-ES	DE-FR	FR-ES
DE-ES	1.00	1.00	0.88
DE-FR	1.00	1.00	0.88
FR-ES	0.88	0.88	1.00

(c)

Sim. Threshold: 1			
	DE-ES	DE-FR	FR-ES
DE-ES	1.00	1.00	0.98
DE-FR	1.00	1.00	0.98
FR-ES	0.98	0.98	1.00

(d)

Figure 10.11: Spearman Correlation between different language pairs for different thresholds: a) 0.85, b) 0.90, c), 0.95, d) 1.00. Each language pairs are positively and strongly correlated to each other, with the lowest value of correlation score 0.882 between German-French and French-Spanish when the similarity threshold is set to 0.95.

Execution time

Execution time for each similarity varies to a large difference, depending on the complexity of the measures. There are 89.3 million comparisons performed to map the concepts in our experiment, hence the required computation process could take a long time. This can be seen in the logarithmic chart on Figure 10.5. The execution time in the cluster highlights that two of the tested similarity measures, TF-IDF and Soft TF-IDF, performed very slow compared to the other similarity measures. TF-IDF and Soft TF-IDF build a corpus first and use the corpus along with compared labels, instead of directly using the strings or tokens from the compared labels done by other similarity measures. Despite the slow performance, TF-IDF provides great F-Measure values and those values are robust to the change of filter values since, by nature, TF-IDF discounts the importance of less-determining words. Jaro and Jaro-Winkler are implemented using Cython[8] within the library (indicated with an asterisk in Table 10.5) so, in theory, these similarity measures should have a faster performance compared to pure-Python implementation. Cython is designed to approach the performance of C as a compiled programming language, instead of the Python as an interpreted programming language. However, cythonized implementation on those similarity measures does not significantly perform differently compared to other similarity measures that are not implemented with Cython. Most of the similarity measures took about only several minutes or less than an hour to run in the cluster, especially set-based similarity measures.

The robustness of TF-IDF can be used as a default choice for linking between multilingual concepts. However, the computational complexity for TF-IDF can be an issue if the datasets to be linked is high in volume. Token sort yields the highest F-Measure score in our experiment, and computational complexity is far less costly compared to TF-IDF, but a proper similarity threshold needs to be carefully approximated for Token Sort to yield the best result.

[8] https://cython.org/

CHAPTER 11

Comparative Analysis of Open Fiscal Data

Despite the increasing size of the open fiscal datasets being published, the level of analytics done on top of these datasets is still limited. There is a plethora of tools and ontologies for open fiscal data e.g., transformation, linking, multilingual integration, and classification. These existing technologies enable the development of a pipeline that could be used for comparative analysis of open fiscal data. In this work, we also contribute for improving the data quality, data analysis, data integration, fiscal data platform and fiscal concept mappings as elaborated in the previous chapters. In this chapter, we demonstrate the comparative analysis over linked open fiscal data, Open fiscal data are cleaned, analyzed, transformed (i.e., semantically lifted), and have their related concept labels connected across different public administrations so budget/spending items from related concepts can be queried. Additionally, the information on linked open data (e.g., DBpedia) has been used to provide additional context for the analysis. We provide a proof-of-concept and demonstrate that such a cross-comparison is possible using the existing tools.

This chapter is based on the following publication:

- **Fathoni A. Musyaffa**, J. Lehmann, H. Jabeen. *Cross Administration Comparative Analysis of Open Fiscal Data*. International Conference on Theory and Practice of Electronic Governance (ICEGOV) 2020. Athens, Greece.

11.1 Requirements

Publishing fiscal datasets is one of the first key steps to be transparent in regard to the financial management of the public administration. With increasing volume of available fiscal data, analyzing open fiscal datasets has more potential to engage the public, for example, by performing comparative analysis across cities that shares similar properties. Yet to enable comparative analysis, several steps need to be done, such as:

1) Ensuring the data are published by considering several factors [6], [68], [78], and even better, if specific quality factors for open fiscal data are considered (see chapter 4).

2) Providing representations that support semantics for open fiscal data, such as the OpenBudgets.eu (OBEU) ontology [30] for RDF datasets.

3) Making standardized concepts (also referred to as *classification*, *code list* or *vocabulary*) across fiscal datasets available and reused whenever it is relevant and applicable for the published fiscal data. The standardized concepts are typically published by an interstate organization such as the European Union and the United Nations. Reusing standardized concepts for fiscal data is unfortunately not yet a common practice.

4) Making available link sets that maps similar or related concepts across datasets from different public administrations. This can be based on different datasets that are published by different organizations but share similar topics/labels across their classifications. The link sets should also be available in the RDF format.

5) Making different datasets have similar metrics (e.g., similar currency) and granularities (e.g., similar temporal units for each fiscal records).

In the current state, the above steps are not being followed, making the cross-comparative analysis of open fiscal data a challenging task. The comparative analysis could help civil communities, journalists, and citizens to analyze public budgeting performance and help in highlighting best practices in public administration budgetary practices. For example, a person could look up a budget for expenditure (e.g., elementary school funding) for similar cities (similar by e.g., population size) and see how respective public administrations allocate their budgets for that particular expenditure.

11.2 Motivation

The motivation of this pipeline is to compare budgets and spending from two different public administrations with similar properties. For example, from the data on DBpedia, it can be seen that the city of Bonn has a similar population size as the municipality of Thessaloniki. From this information, comparing the budget allocation for both cities can be interesting, particularly when the labels of that budget item have a similar meaning. This is illustrated in Figure 11.1. Here, we compare the budget for conceptually related items: "Referat Stadtförderung" in German and "Έξοδα ενημέρωσης και προβολής δραστηριοτήτων του Δήμου" in Greek which according to Google Translate, both are related to *promotions*. Both concepts belong to functional classification. In this case, the budget allocation of two public administrations having similar properties can be seen and compared. This use case can provide an additional analysis approach for the citizen, civil organization, and journalists that are interested to mash up open fiscal data with the available linked open data from, e.g., DBpedia or Wikidata.

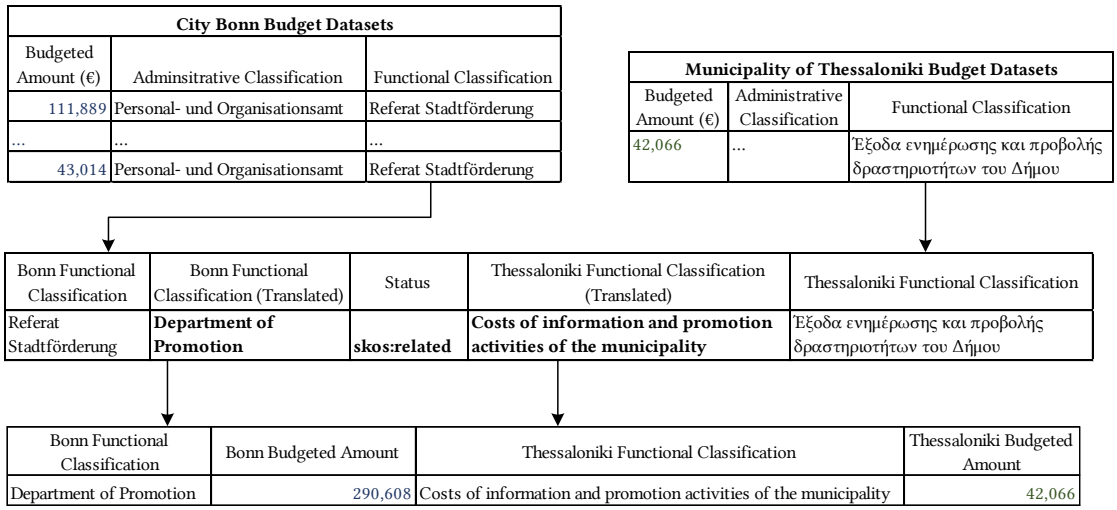

Figure 11.1: Comparative analysis of open budget data that are represented in different languages.

11.3 Pipeline

Our pipeline is illustrated in Figure 11.2. Available datasets and classifications are analyzed to ensure proper modeling according to the OBEU ontology. Meanwhile, the classifications coming from different public administrations are analyzed, translated, and mapped for related links. These mappings are then subsequently evaluated. These related links that are confirmed to be relevant are passed along with the classifications and datasets for transformation into the RDF format. The transformation results in datasets, classifications, and link sets which are then stored in a triple store. Additional information is needed to get an additional context, which is used to find which datasets to be compared with. This is done by a federated query using external linked data service in DBpedia.[1] Stored data are then queried for comparative analysis. A more detailed approach is provided in the following sub-sections.

Datasets, Analysis, and Transformation

The semantic lifting process in general is elaborated in chapter 8. There are two datasets that we use for the experimen in this chapter: the expenditure budget from the city of Bonn and the expenditure budget from the municipality of Thessaloniki. For datasets from Bonn, we obtained the data directly from the responsible city officers for the data. We clarify both the main budget datasets and the accompanying classifications from Bonn datasets. After the clarification process, a transformation is performed to produce an RDF representation of Bonn datasets that are compatible with the OBEU ontology. LinkedPipes ETL tool [62] is used to perform the transformation, which allows loading the datasets from tabular formats, adding metadata over the datasets that

[1] https://wiki.dbpedia.org/OnlineAccess

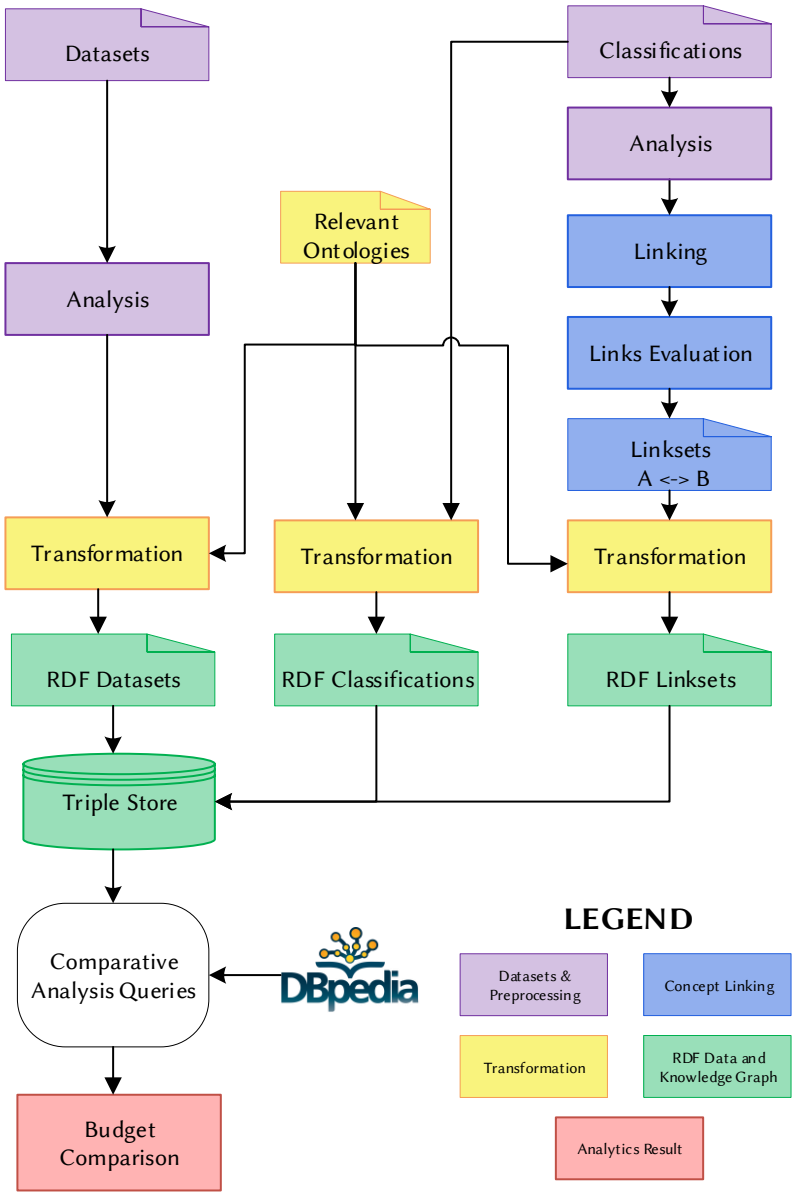

Figure 11.2: The flow to analyze, map, transform, store and query open fiscal datasets.

conform with DCAT-AP specification, and performing semantic lifting of the data into RDF with SPARQL queries. The transformation pipeline for the Bonn dataset can be seen in Figure 11.3. Each box in Figure 11.3 has its own roles, such as (1) download the dataset, (2) map fields/columns in the records into a specific property, (3) merge data, (4) construct necessary triple statements, (5) insert metadata and data structure definition, (6) combine the data and, (7) materialize the datasets into a flattened file.

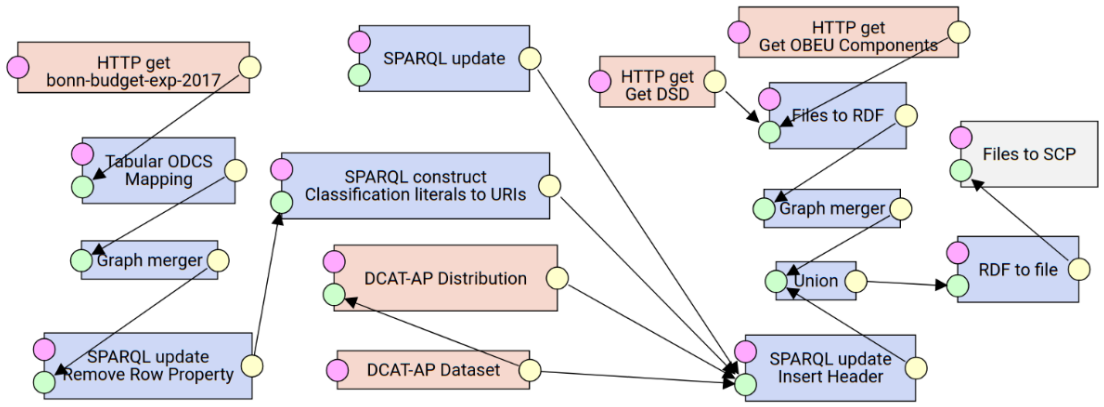

Figure 11.3: Bonn expenditure dataset 2017 transformation pipeline. The CSV raw data CSV is mapped column-wise to the OBEU ontology properties. The data are transformed further and enriched by using SPARQL statements to follow OBEU data model requirements and constraints.

These transformation pipelines can be found in a GitHub repository[2] and can be inspected and executed online using the LinkedPipes Demo website.[3] The Thessaloniki expenditure datasets[4] are available in their open data portal. A transformed version of the datasets represented in the OBEU ontology is provided in the GitHub repository as well.[5]

Concept Mapping

Concept mapping among the two datasets is done utilizing Apache Spark [44] and *py_stringmatching,*[6] a string-matching library as elaborated in detail within chapter 10. To recapitulate, we perform benchmarking of several string similarity measures from different categories: string-based similarity measures, set-based similarity measures, hybrid measures (a combination of both string and set-based similarity measure), phonetic similarity measures, and bag-based similarity measure. We use the European Union's Common Procurement Vocabulary (CPV) classification [24] for the gold standard, which has human-translated labels in 24 different European languages. We then use Google Translate to translate the labels from other languages (German, French and Spanish labels of CPV datasets labels) into English and label the translation based on RFC 6497 – BCP 47 Extension T [157]. For example, making use of the extension specification, "en-t-de" denotes that the label content is in English, but it is obtained by transforming and translating the labels which were previously available in German. We performed 19 different string similarity measures computation from the translated labels and then check: (1) which similarity measures yield the highest F-Measure score, (2) which similarity

[2] https://git.io/JejR1
[3] https://demo.etl.linkedpipes.com/#/pipelines
[4] https://gaiacrmkea.c-gaia.gr/city_thessaloniki/index.php
[5] https://git.io/JejRM
[6] https://git.io/JejRy

measures have the best-performance, and (3) how robust these similarity measures against changes in similarity thresholds. From our experiment, we know that the TF-IDF similarity measure provides the best F-Measure performance. We reuse the conclusion from this mapping experiment for this chapter, therefore, we use TF-IDF similarity measures to predict relation links in the Thessaloniki and Bonn budget datasets. The final result of the concept mapping process is link sets. Link sets explicitly state that a concept of a functional classification from Thessaloniki is *related* to a particular concept of functional classification from Bonn.

Data Storage

The results of datasets transformation and related links that have been transformed to RDF are stored in a triple store, a database for data represented in RDF formats. The data within the triple store is queried using SPARQL queries. All data from previous operations are stored in the triple store, those are: (1) transformed datasets from the city of Bonn, (2) transformed datasets from the municipality of Thessaloniki, (3) functional classifications from both public administrations, and (4) produced link sets. We use Apache Jena Fuseki[7] as the triple store.

Comparative Analysis

Datasets, classifications, and link sets that are stored in the triple store are queried for comparative analysis. The query decision can be based on relevant properties available from open knowledge bases. For example, DBpedia and the *total population* property within the DBpedia page of compared cities/municipalities. Figure 11.4 illustrates the non-exhaustive DBpedia properties that are relevant to be used as a comparison point for open fiscal data. These properties for comparative analysis can be from different public administration level: countries (e.g., currency, GDP, GDP per capita, GDP per capita rank, GINI score, Human Development Index, HDI change), states, and cities (metro area size, urban area size, metro population, urban size, state, province, etc.). Some properties are shared between different public administration levels.

Relevant information can be obtained using the properties of each public administration. For example, information regarding the list of money allocated from related functional concepts coming from public administrations that have a similar total population size. Municipality of Thessaloniki and the city of Bonn have a similar population number according to DBpedia. Therefore, the amount of money that each public administration's functional classification concepts between the two public administrations can be compared based on this fact.

[7] https://jena.apache.org/documentation/fuseki2/

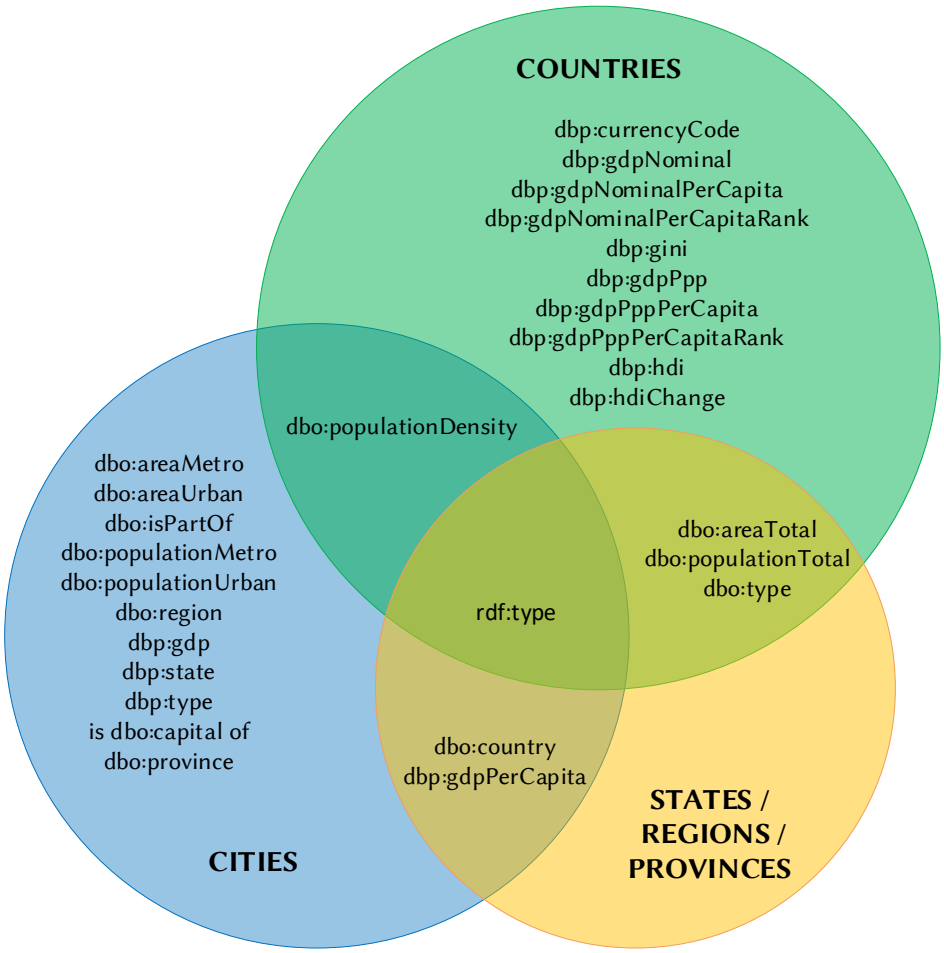

Figure 11.4: Relevant DBpedia properties to enrich OFD for further comparative analysis. The prefix `dbo` refers to `<http://dbpedia.org/ontology/>`.

11.4 Analysis

The datasets used in the experiment have different characteristics in terms of e.g., classification types availability and the way data are transformed. In terms of *classification types*, different datasets include a different number of classifications, for example, classification datasets from the municipality of Thessaloniki are comprised of administrative classification and functional classification. The unique code enumeration and labels (i.e., the primary key in database terms) for these classifications is not entirely clear from Thessaloniki's data portal, but the list is available and can be obtained via correspondence with the dataset's Github repository maintainer. The list of classification from Bonn datasets is not publicly available either, thus the data were also available through correspondence with the officials from the city with a public domain license. We mirror this dataset into Github. Additionally, datasets from the city of Bonn have more classifications: business

area, economic classification, and one local classification named as a *profitcenter*, which we need to preprocess since *profitcenter* is a composite of administrative and functional classification.

These data are also different in terms of *transformation modeling*. Since the datasets have different classification and budget phase availability, the datasets from different public administrations are modeled in a slightly different manner in the OBEU ontology. Specifically, *observation* provides a granular representation of the financial record. In the case of the city of Bonn's datasets, an observation consists of only one amount of expenditure. On the other hand, slice provides a coarse representation of a public administration record. It may consist of several observations, combined with several different dimensions. In the Municipality of Thessaloniki's case, one record contains several dimensions of different classification types that are modeled as a slice. This slice has several amounts of expenditure values in which each value represents different budget phases (drafted, revised, approved, and executed).

Table 11.1: An example of functional classification for the Thessaloniki dataset.

Code	**Original Label (EL)**	**English-translated Label**
641	ΕΞΟΔΑ ΜΕΤΑΦΟΡΩΝ	TRANSPORT COSTS
6411	Έξοδα κίνησης ιδιόκτητων κεταφορικών μέσων (καύσιμα λιπαντικά διόδια κ.λ.π.)	Expenses motion ketaforikon owned media (fuel oils tolls etc.)
6412	Έξοδα μεταφοράς αγαθών φορτοεκφορτωτικά	Transport costs stevedores goods
6413	Μεταφορές προσώπων	transport of persons
6414	Μεταφορές εν γένει	Transport generally

Bonn and Thessaloniki datasets have both functional classification and administrative classification. For this experiment, we are using functional classification as a comparison point between two datasets. As for the mapping process, Thessaloniki functional classification consists of 394 concepts. The functional classification for the Municipality of Thessaloniki contains a hierarchical concept, as can be seen in Table 11.1. In Table 11.1, the concept of transport cost is divided into four concepts: (1) Cost of transport of privately-owned and paid media (fuel, toll, lubricants, etc.) (2) freight forwarding costs, transport of persons, and general transport. The translation as we can see from the table is obtained from Google Sheet's translation feature. At the time of our experiment, the translation of Google Sheet has a subordinate quality compared to its Google Translate web version (see the English-translated label from Table 11.1). Bonn functional classification consists of 183 concepts. The functional classification of Bonn is also provided hierarchically as well (see Table 11.2). The concept of transport for Bonn datasets have more sub-concepts compared to Thessaloniki's concepts of transport. There is also a hierarchy in this classification, 4-digits concepts which code has "0" suffix is a more general concept, followed by codes with similar three-digit prefix as sub-concepts. The different granularity of concepts in both tables illustrate how obtaining *exactly similar* links is still a challenge, and hence we proceed with *related* links instead.

Since there are multiple observations with the same functional concept spanned over different values, the aggregation operation needs to be performed. For example, a func-

Table 11.2: Another example of functional classifications published by the Municipality of Bonn.

Code	Original Label (DE)	English-translated Label
1200	PB12 Verkehrsflächen und -anlagen, ÖPNV	pb12 traffic areas and facilities, public transport
1207	Verkehrsplanung	traffic planning
1201	Gemeindestraßen	local roads
1202	Kreisstraßen	county roads
1203	Landesstraßen	country roads
1204	Bundesstraßen	federal roads
1205	Parkeinrichtungen	park facilities
1206	ÖPNV	public transport
1208	Straßenreinigung und Winterdienst	street cleaning and winter services

tional classification concept *transport* could be distributed among different administrative offices. Here, all budget/spending items are summed from different administration offices, enabling one-to-one comparison of related labels from different municipalities.

The transformation is done using the latest LinkedPipes version,[8] with Apache Jena Fuseki v 3.12.0 as the triple store. Each transformation pipelines are available on GitHub (Bonn[9] and Thessaloniki[10]). The link mapping part utilizes the translated concept using Google Translate via Google Sheet and the result is fed into our concept mapping framework that uses Apache Spark 2.3.1 and the py_stringmatching library v0.4.1. The detail of the concept mapping part is discussed in chapter 10.

11.5 Result and Discussion

Querying available datasets that have similar contextual properties (e.g., as seen in Figure 11.4) can be done using DBpedia's SPARQL service, as illustrated in Listing 11.1. Here, we select distinct datasets from the local triple store whose public administration has a total population between 300.000 – 400.000 people. The result of this query listed in Table 11.3, which shows the available datasets URI in our local triple store, organization (city) URI, and the population size of the city obtained from DBpedia entries.

```
PREFIX  dbo:  <http://dbpedia.org/ontology/>
PREFIX  qb:   <http://purl.org/linked-data/cube#>
PREFIX  obeu-dimension: <http://data.openbudgets.eu/ontology/dsd/dimension/>
SELECT DISTINCT  ?dataset ?organization ?populationTotal
    WHERE {?dataset a qb:DataSet ;
                    obeu-dimension:organization  ?organization
           SERVICE <http://dbpedia.org/sparql?default-graph-uri=http://dbpedia.org>
             {?organization dbo:populationTotal  ?populationTotal}}
```

Listing 11.1: Querying available datasets based on specific values (e.g. population size) available in DBpedia.

[8] https://git.io/JejRS
[9] https://git.io/JejR1
[10] https://git.io/JejR7

Table 11.3: The resulting query of available datasets that fulfil certain population numbers in DBpedia.

@prefix obeu-ds: <http://data.openbudgets.eu/resource/dataset/> .
@prefix dbr: http://dbpedia.org/resource/> .

Dataset	Organization	PopulationTotal
obeu-ds:bonn-budget-exp-2017	dbr:Bonn	311287
obeu-ds:bonn-budget-exp-2018	dbr:Bonn	311287
obeu-ds:bonn-budget-exp-2019	dbr:Bonn	311287
obeu-ds:budget-thessaloniki-expenditure-2017	dbr:Thessaloniki	385406
obeu-ds:budget-thessaloniki-expenditure-2018	dbr:Thessaloniki	385406
obeu-ds:budget-thessaloniki-expenditure-2019	dbr:Thessaloniki	385406

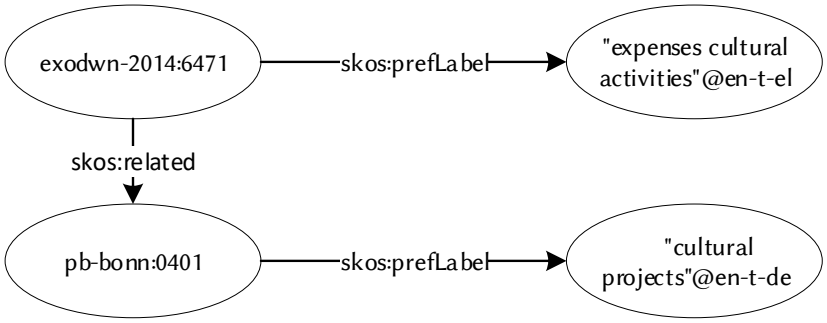

Figure 11.5: An illustration of a relation between concepts from the city of Bonn and the municipality of Thessaloniki.

Result

The mapping experiment results in 87 related links. The links are associated with `skos:related` property. Figure 11.5 illustrates the `skos:related` link across concepts that are related to culture from Bonn and Thessaloniki, with each `skos:prefLabel` indicates that the concepts have labels in English translated from respective original languages.

The transformation result is loaded into the triple store. This consists of expenditure budget datasets and functional classifications from the city of Bonn (2017-2019) and the Municipality of Thessaloniki (2015-2019), as well as created link sets from the mapping experiment. In total, there are 219.220 triples obtained from this experiment.

Listing 11.2 provides an example of a query to obtain the amount of money budgeted for similar items on the datasets found to have similar contextual properties. The SPARQL snippets in Listing 11.2 uses a subquery to fetch a set of observations in Bonn datasets that are known to have related functional classification labels compared to Thessaloniki datasets. Here, the set of observations is restricted to a particular fiscal year (2017), which is specified using `http://reference.data.gov.uk/id/year/2017` URI. As each of the related functional classification items may span over several observations in both of the datasets, an aggregation operation is performed by summing the amount of budgeted

money for that particular functional classification concept. The final result is then filtered by the language of labels available in each related concept. In this case, since the labels are transformed by translating from Greek and German to English, `en-t-el` and `en-t-de` language code are respectively used as a restriction to clarify that those are the result of translation operation from respective language codes.

```
PREFIX  qb:   <http://purl.org/linked-data/cube#>
PREFIX  gr-dimension:
↪  <http://data.openbudgets.eu/ontology/dsd/greek-municipalities/dimension/>
PREFIX  obeu-budgetphase: <http://data.openbudgets.eu/resource/codelist/budget-phase/>
PREFIX  obeu-measure: <http://data.openbudgets.eu/ontology/dsd/measure/>
PREFIX  bonn-dimension:
↪  <http://data.openbudgets.eu/ontology/dsd/bonn-budget-simplified-updated/dimension/>
PREFIX  xsd:  <http://www.w3.org/2001/XMLSchema#>
PREFIX  skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX  ukref-year: <http://reference.data.gov.uk/id/year/>
PREFIX  obeu-dimension: <http://data.openbudgets.eu/ontology/dsd/dimension/>

SELECT  ?bnFC ?bnLabel ?thSliceFC ?thLabel (xsd:decimal(?bnAmountTotal) AS
↪  ?bnAmountTotalDec) (SUM(?thAmount) AS ?thAmountTotalDec)
WHERE
  { ?thDataset a qb:DataSet ;
         obeu-dimension:fiscalYear ukref-year:2017 ;
         qb:slice ?thSlice .
    ?thSlice a qb:Slice ;
             gr-dimension:economicClassification ?thSliceFC ;
             qb:observation ?thObs .
    ?thObs a qb:Observation ;
         gr-dimension:budgetPhase obeu-budgetphase:approved ;
             obeu-measure:amount ?thAmount .
        ?thSliceFC skos:related ?bnFC ;
             skos:prefLabel ?thLabel .
    ?bnFC skos:prefLabel ?bnLabel
      { SELECT  ?bnFC (SUM(?bnAmount) AS ?bnAmountTotal)
    WHERE
    { ?bnObs a qb:Observation ;
             bonn-dimension:functionalClassification ?bnFC ;
             obeu-measure:amount ?bnAmount ;
             qb:dataSet ?bnDataSet .
          ?bnDataSet obeu-dimension:fiscalYear ukref-year:2017}
          GROUP BY ?bnFC}
       FILTER ( lang(?thLabel) = "en-t-el" )
       FILTER ( lang(?bnLabel) = "en-t-de" )}
GROUP BY ?thSliceFC ?bnFC ?bnAmountTotal ?thLabel ?bnLabel
```

Listing 11.2: An example of SPARQL query to perform a comparative analysis between Bonn and Thessaloniki datasets. Subquery was used to aggregate functional classification amount - which initially was distributed across different budget lines.

The result of the query is sampled in Table 11.4 with the following columns: Bonn functional concept URI, translated concept labels from Bonn datasets, Thessaloniki functional

Table 11.4: An example of comparative analysis query result.

@prefix bn-cl-pu: «http://data.openbudgets.eu/resource/codelist/produktuebersicht_bonn/>
@prefix th-cl-koe: <http://data.openbudgets.eu/resource/codelist/kae-ota-exodwn-2014/>
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>

Bonn Concepts URI	**Concept Labels - English Translation**	**Thessaloniki Concepts URI**	**Concept Labels - English Translation**	**Amount Approved: Bonn**	**Amount Appoved: Thessaloniki**
bn-cl-pu:0802	”sports promotion”@en-t-de	th-cl-koe:6472	”expenses sports”@en-t-el	”2,198,574.75” ^^xsd:decimal	”15,054.9” ^^xsd:decimal
bn-cl-pu:0119	”department of promotion”@en-t-de	th-cl-koe:6431	”costs of information and promotion activities of the municipality”@en-t-el	”290,608.375” ^^xsd:decimal	”42,065.87” ^^xsd:decimal
bn-cl-pu:0124	”administrative organization and it applications”@en-t-de	th-cl-koe:6266	”maintenance of software applications”@en-t-el	”5,007,714.5” ^^xsd:decimal	”63,819.04” ^^xsd:decimal
bn-cl-pu:0401	”cultural projects”@en-t-de	th-cl-koe:6471	”expenses cultural activities”@en-t-el	”842,904.75” ^^xsd:decimal	”392,930.09” ^^xsd:decimal

concept URI, translated concept labels from Thessaloniki datasets, the approved budget amount of Bonn datasets, and approved budget amount from the City of Thessaloniki. For example, knowing the fact that both Thessaloniki and Bonn have the population size around 350,000 – 400,000, from the initial DBpedia query (Listing 11.1) we can compare that cultural expense listed as code `0401` in Bonn is allocated at 842,904 € while the expense for cultural activities listed as code `6471` allocated for the Municipality of Thessaloniki is 392,930 €. This comparison is illustrated in Figure 11.6.

The comparative analysis experiment results in 47 related links. The result of the comparison is affected greatly by the noise in the datasets (e.g., the budget amount is in zero) as well as the quality of generated related links. The mapping, query result, as well as the whole resulting experiment is provided in our GitHub repository.[11]

Lessons Learned

This chapter presents efforts that have enabled a comparative analysis of open fiscal data, providing a proof of concept highlighting the potential in open fiscal data analysis by exploiting the growing linked open data knowledge bases (e.g., DBpedia, Wikidata). To enable a wider scale adoption for publishing open linked data to be integrated into the linked open data cloud, there are several points that we learn:

- Different public administrations have different legislation, business process, and data flow. Therefore, each dataset is most probably different and tends to be complex. We

[11] https://github.com/fathoni/icegov2020-ofd-analysis

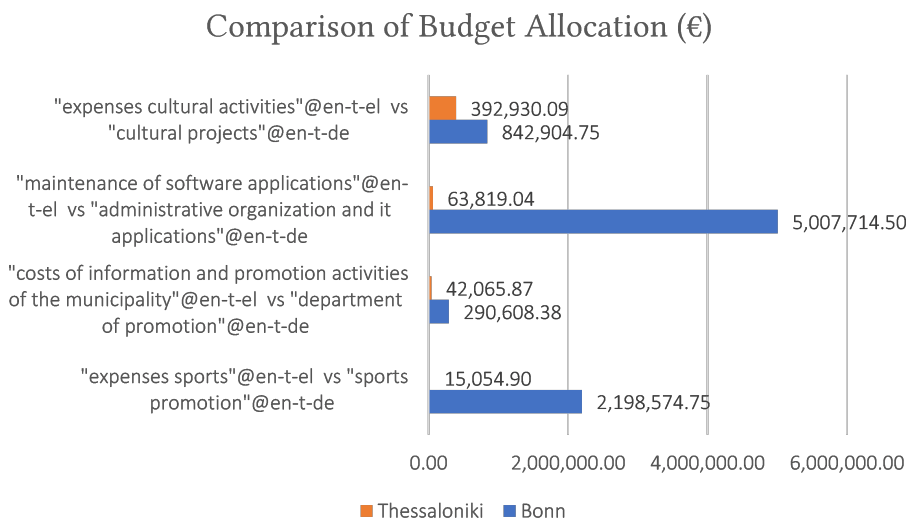

Figure 11.6: A visualized comparison of related and aggregated budget from both public administrations.

suggest that a careful data simplification process should be done prior to publishing if such datasets are initially complicated (e.g., contain positive/negative values, composite classification items). The datasets should be documented mentioning what each column in the datasets contains. The available classification should be explained clearly.

- Applying additional technical processes for enabling datasets publishing as Linked Open Data is a good practice and desirable, however, there is a different capacity for public administrations to invest in such technical expertise. In this case, the attempt for public administrators to publish good quality open datasets (see [1], [68], [17], [78]) with an open license can help civic and research communities to analyze and disseminate the datasets. The civic and research communities often have the technical capacity to understand, reuse, and publish the data. Good quality data would encourage innovation from these communities.

- Several classifications have been published by interstate organizations. However, the adoption of these classifications is not yet a widespread practice. Reusing published concepts to publish data helps improving data integration and consumption process.

- With the rise of AI approaches, the need for a structured knowledge base in the form of linked is growing. The size of the information available in initiatives such as DBpedia and Wikidata is hence expanding. Publishing Linked Open Fiscal Data enables data consumers to utilize more context from these growing knowledge bases for a more advanced analysis.

Concluding Remarks for Part IV: Enabling Comparative Analysis of Open Fiscal Data

In this part, we deal with the final Research Question:

> **RQ4**: How can we facilitate the comparative analysis of heterogeneous fiscal datasets?

Having open fiscal datasets published by several public administrations, the amount of datasets available is increasing. A different type of analytic can be performed, such as comparative analytics that compares budget/spending across public administrations. This requires a mapping of fiscal concepts across public administration and enriching the datasets with open knowledge bases.

In chapter 10, we present the IOTA framework that interlinks multilingual fiscal data by making use of the fiscal classifications. IOTA is designed using the distributed in-memory scalable platform (Apache Spark) to deal with the complex task of string similarity assessment for a large number of concepts. IOTA utilizes nineteen different similarity measures to assess the similarities of the concepts. We test the performance of IOTA over data containing the three translated language pairs. We find that the best similarity measure with the relatively low computational cost is Token Sort. It provides the highest F-Measure score when the similarity threshold is properly approximated. TF-IDF also provides a high F-Measure across different similarity thresholds at the expense of significantly longer execution time. The correlation between language pairs shows a consistently high and positive correlation. IOTA can be easily adapted to be used for other use cases and domains.

In chapter 11, we demonstrate a proof of concept that enables comparative analysis of open budget and spending data. This involves the usage of the OBEU ontology to enable a unified semantic representation of open fiscal data, using information available on public knowledge bases to enrich the context of the datasets and to create relation links between similar concepts across datasets.

Part V

Epilogue

CHAPTER 12

Conclusions and Future Direction

Open fiscal data are published openly to enhance the transparency and accountability of public administration, motivated by increasing trust for the governance which will enhance political participation from their citizen. Understanding fiscal datasets, however, requires expertise in fiscal and technical domains. Facilitating the gap between the community and released open fiscal data with strategies, methods, technologies, and research on open fiscal data is the main aim of this study.

This thesis consists of different parts. Part II focuses on the state of open fiscal datasets. Here we perform a thorough analysis of open fiscal data from diverse public administrations in terms of language, geographical locations, and public administration levels. We survey several existing open data publishing recommendations. Since these recommendations are for the general open data domain, we provide additional factors that contribute to good quality open data specifically in the fiscal domain. We check each collected datasets with open fiscal data quality factors, score, and finally rank them. We also find common patterns of heterogeneities, which we organize and classify hierarchically.

In Part III, we perform various tasks with regard to unifying the heterogeneous datasets from different sources into one single, semantic data format by using the OpenBudgets.eu ontology. The ontology is based on the statistical data cube (DCV) ontology. We develop an ontology for semantifying OpenAPI-formatted API endpoints. We provide an overview of state-of-the-art open fiscal data models, as well as comparing heterogeneity items we found upon analyzing fiscal datasets from Part II with state-of-the-art data models. This is followed by the semantification of available fiscal data using the semantic OBEU data model to support further analysis. At last, we devise an architecture for the open fiscal data domain, incorporating common tasks in the public fiscal domain such as data ingestion, transformation, storage, query, visualization, and participatory budgeting.

Part IV provides a framework for the mapping of fiscal concepts from different languages. We utilize machine translation to translate concepts into English as an intermediary language, followed by computing similarity scores using 19 similarity measures formulas through distributed computing. The study is then used for creating comparative fiscal data analytics across different public administrations.

12.1 Answering Research Questions

In this section, we provide the answers to the research questions mentioned in section 1.3, and summarize our contributions.

Research Question 1

> **RQ1**: What are the requirements for publishing high quality open fiscal data?

There exist several guidelines on how high-quality open data is published. The quality factors published in these guidelines aligned well with fiscal open data. Some of the quality factors requirement might be difficult to satisfy. For example, publishing open fiscal data as linked open data can be difficult for public administrations given that the steep technological commitments a public administrator need to make. However, many of these quality factors are an easy requirement to address. This is true especially for publishing datasets by clearly mentioning its license, or also important, in an open license. Another easy change that can be made by public administrators is publishing their open data in a structured open format. Our contributions towards this research question is an assessment framework that allows fiscal domain-specific factors to be used within the assessment process. We also investigate open data assessment methodologies for generic domains with our framework and then carefully analyze open fiscal datasets from various public administrations on different administration levels. We reported the state of datasets checked against 23 quality factors.

Research Question 2

> **RQ2**: What types of data heterogeneity problems occur with open fiscal data?

The factor that hinders a common representation of open fiscal datasets is the heterogeneity, as we found that these datasets can be published on a very different structure regardless of similar file formats that are being used. For example, a spreadsheet file can contain fiscal data, the different classification being used, and how granular the budgeting transactions are recorded. On a conceptual level, this poses a challenge on how public fiscal datasets can be presented uniformly across public administrations on different scope/level of administrations. In this regard, we contribute by providing an enumeration of these heterogeneities after conducting a detailed survey and analysis across these public fiscal datasets. Since there have been already existing open public fiscal data models that aim to universally represent these datasets, we contribute further by assessing whether these data formats can support these individual heterogeneity items. We found that none of these data models completely support all the heterogeneities that we found, hence we provide recommendations for technical CSOs/NGOs, academics in the open data domain, and open fiscal data publishers to deal with these heterogeneities.

Research Question 3

> **RQ3**: How can we improve the interoperability of open fiscal data by using a semantic data model?

This question is linked to the RQ2, in which we provided an overview of data heterogeneities and its supporting data formats. A semantic ontology, namely OBEU, has been already developed to represent open fiscal data. We use this ontology for semantically representing open fiscal data. The ontology also supports OLAP data cube, hence open fiscal data that are transformed into this format can be analyzed based on the intended dimensions. The transformation is done using an ETL pipeline, with the raw fiscal datasets (mostly spreadsheet documents) as the input and semantic format that conforms with OBEU ontology as the output. As fiscal data can have a lot of different structures, it is difficult to create a one-pipeline-fits-all transformation, but there are patterns of these pipeline fragments that can be reused. Once the datasets are transformed into a semantic representation, we can make the datasets interoperable. This is also supported by providing an architecture that able to make use of interoperable data. In addition, we also perform the semantification of standardized API endpoints to facilitate gathering more data from supporting data portals.

Research Question 4

> **RQ4**: How can we facilitate the comparative analysis of heterogeneous fiscal datasets?

Making datasets comparable requires the availability a comparison point that can be used for making comparative analysis. Most open public fiscal data are published with fiscal concepts surrounding them, and we contribute by devising a framework to create mappings from translated fiscal concepts coming from different datasets. This framework can also be applied to datasets from different domains. We provide a report on the performance of different string similarity measures to create a similarity mapping from translated concepts. At last, we contribute by providing a proof of concept that allows us to perform comparative analysis from open fiscal data. This combines the contributions from previous research questions into a comprehensive solution.

12.2 Future Works

In this section, we present some suggestions to improve this work in both short-term and long-term directions.

12.2.1 Short-Term Works

In this thesis, we contribute on the investigations, ideas, and solutions to make the open fiscal data more understandable and reusable by the community. This is done by considering fiscal data cycle from the very beginning of fiscal data publication until how

it is being used for making a more insightful analytics. Here, we elaborate further on how our contributions can be extended in the short term.

- **Devising a semi-automatic mechanism to assess OFD quality**

  The work in chapter 4 has provided a ground for creating an assessment of open fiscal datasets. To improve this work, further contribution in this area can be done, such as :

  1. Providing a semi-automatic quality assessment tool for open fiscal datasets so that public administrators could easily evaluate their fiscal data themselves, which may include a file format development to represent the assessment result. For example, the assessment portal ideally contains a simple set of statements along with a short explanation for each statement regarding the quality factors. The quality score of the published datasets should appear by the end of the assessment, along with its ranking compared to other published fiscal datasets. Additionally, it also shows how compatible the assessed datasets with the state-of-the-art fiscal data models, as shown in chapter 7.
  2. Performing studies which assess how adherence to the proposed publication guidelines actually influences open fiscal datasets' consumption. This can be done through surveys, as well as the reuse rate of the published open fiscal data.

- **Demonstrating and publishing best practices in OFD analytics**

  Making datasets conform to standards and recommendation guidelines takes a lot of efforts. Spending extra resources for these efforts may not seem worth the investments, if a clear use case and analytical enhancement can not be visibly seen. Therefore, providing demonstrations and proof of concepts showing that publishing standard-abiding open fiscal data can create values could motivate data publishers to provide their datasets in a high-quality manner. This value ranges from transparency, accountability, fiscal participation, and more advanced analytics. These analytics can be in the form of visualization (as shown in chapter 9) as well as comparative analytics (as shown in chapter 11).

- **Investigating approaches to improve OFD standard adoption rate**

  Despite the existing recommendation of best practices on publishing open linked data, the number of open fiscal datasets published following the best practices is rather small, as discussed in chapter 4. A standard on publishing open fiscal data should be agreed upon, published, and followed. This is, however, a challenging task to persuade public administrations giving up a small part of their sovereignty by adhering to standardized OFD publication practices for greater benefit. This can be done via an organization from the supra-national level, by building a network within these initiatives and communities that are able to enforce open fiscal data publishing in a standardized manner. Several initiatives exist, such as the publication of DCAT-AP by the EU, or the Open Government Partnership (OGP).

- **Exploiting OpenAPI-specified public API endpoints further**

  A list of publicly available OpenAPI-specified API endpoints is available in the SwaggerHub registry. This registry is searchable, and as of April 2020, there are 93 open data registered in SwaggerHub.[1] These endpoints can be further semantified to support service orchestration as discussed in chapter 6. The work regarding OpenAPI semantification be improved by 1) refining and expanding the OpenAPI vocabulary with support for a complete list of the elements from OpenAPI Specification, 2) automating RDF extraction from annotated JSON OpenAPI description into JSON-LD format in which the resulting extraction can be provided in a semantic service registry, and 3) studying and providing a solution for interoperability problems that exist between different versions of the OpenAPI specification using this semantic approach.

- **Devising automated semantic-lifting framework**

  An attempt to create a unified pipeline to semantify open fiscal datasets was not satisfactory due to the heterogeneous format and structure of open fiscal datasets. Once we have a standardized format that is agreed and used by open data publishers, it would be an easier task to perform a semantic lifting of standardized datasets using e.g., a unified pipeline or process to a semantic data model.

12.2.2 Long-Term Works

This thesis focused specifically on linked open data in the fiscal domain, covering the data quality assessment, heterogeneity challenges, data modeling and integration, platform architecture, concepts interlinking framework, and wrapped up with the proof of concepts for fiscal data analytics. This work has contributed in several aspects of linked fiscal data analytics, as summarized in section 12.1, improved upon the vast number of previous research contributions by investigating, analyzing, adapting, devising, and proposing contributions specifically for open fiscal data domain. Thus, the future developments of related fields mentioned in this thesis would also benefit the research on open fiscal data domain. These include the advancements in the field of natural language processing, machine learning, named entity recognition, named entity disambiguation, named entity linking, and big data.

- **Entity recognition and disambiguation for multilingual open datasets**

  In this thesis, we do not experiment with Named Entity Recognition (NER) and Named Entity Disambiguation (NED). NER and NED are emerging fields, with new datasets, algorithms, resources, and optimization methods keep evolving. These developments can help to build a robust NER/NED approach for open fiscal data, in which similar concepts from different language can be recognized as one entity properly, and organized into a hierarchical taxonomy using e.g., SKOS vocabulary.

[1] https://app.swaggerhub.com/search?query=%20open%20data

For each type of classifications, a single information structure can be provided in different languages.

- **Improving cross-lingual entity linking**

  In chapter 10, we use the translation result and analyzed using different string similarity algorithms in a distributed computing environment, devising the IOTA framework. This work could be extended in several ways. First, the performance of other machine translation tools could be compared. Second, it could be evaluated if additional pre-processing tasks can improve the performance and accuracy of the mappings (such as stop words removal, in which additional stop words can be added specifically for open fiscal data domain). Third, for hybrid similarity measures, it can be assessed which specific string similarity measure combination pairs would yield the best F-measure value.

  Our approach with the IOTA framework does not take into account machine learning and natural language processing approach, e.g., word embedding and its derivative approach, which potentially yield a better linking result. Even though our initial experiment using direct cross-lingual word embedding using aligned multilingual Facebook MUSE word embedding was not satisfactory, there may be different approaches that can be used to map cross-lingual fiscal concept with higher F-measure value.

- **Entity linking between multilingual OFD concepts and Knowledge Graphs**

  The vast amount of information on the open knowledge graph keeps expanding and it has a great potential to be utilized for further analysis. Investigating an approach to link the concepts from open fiscal data classifications to popular semantic knowledge bases, such as DBpedia and Wikidata, can make open fiscal datasets concepts richer with information from external sources, which can be used as a comparison point as shown in chapter 11.

Bibliography

[1] OKI, *Place Overview: Global Open Data Index*, 2015, URL: http://index.okfn.org/place/ (visited on 31/01/2017) (cit. on pp. 3, 18, 19, 47, 151).

[2] UK Cabinet Office, *Policy paper: G8 Open Data Charter and Technical Annex*, 2013, URL: https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex#technical-annex (visited on 31/01/2017) (cit. on p. 3).

[3] Open Knowledge International, *Government Budget - Global Open Data Index*, 2015, URL: https://index.okfn.org/dataset/budget/ (visited on 20/01/2019) (cit. on pp. 3, 53).

[4] American Heritage Dictionaries Editors, *The American heritage dictionary of the English language, Fifth Edition*, Fiscal, Houghton Mifflin Harcourt, 2016 (cit. on p. 3).

[5] The World Wide Web Foundation, *Open Data Barometer 4th Edition — Global Report*, The World Wide Web Foundation, 2017, URL: https://opendatabarometer.org/doc/4thEdition/ODB-4thEdition-GlobalReport.pdf (cit. on pp. 3, 4).

[6] Open Knowledge International, *Place Overview: Global Open Data Index*, 2017, URL: http://index.okfn.org/place/ (visited on 30/01/2019) (cit. on pp. 3, 18, 97, 139).

[7] N. Shadbolt, K. O'Hara, T. Berners-Lee, N. Gibbins, H. Glaser, W. Hall and M. C. Schraefel, *Linked Open Government Data: Lessons from Data.gov.uk.*, IEEE Intelligent Systems **27** (2012) (cit. on pp. 3, 22, 46).

[8] A. F. Tygel, J. Attard, F. Orlandi, M. L. M. Campos and S. Auer, "How Much? is not Enough: an Analysis of Open Budget Initiatives", *Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance*, ACM, 2016 276 (cit. on pp. 3, 4).

[9] A. Graft, S. Verhulst and A. Young, *Open Data's Impact - Brazil's Open Budget Transparency Portal: Making Public How Public Money Is Spent*, tech. rep., Govlab, 2016, URL: http://odimpact.org/files/case-study-brazil.pdf (visited on 11/06/2016) (cit. on pp. 3, 4).

[10] N. Huijboom and T. Van den Broek,
Open data: an international comparison of strategies,
European journal of ePractice **12** (2011) 4 (cit. on pp. 3, 4).

[11] The World Bank, *Starting an Open Data Initiative*, 2015,
URL: http://opendatatoolkit.worldbank.org/en/ (visited on 31/01/2017)
(cit. on p. 3).

[12] S. J. Piotrowski and G. G. Van Ryzin,
Citizen attitudes toward transparency in local government,
The American Review of Public Administration **37** (2007) 306 (cit. on p. 4).

[13] R. Stapenhurst and M. O'Brien, *Accountability in governance*,
The World Bank, Washington DC (2008) (cit. on p. 4).

[14] P. Srinivasan, *GovLab Index on Open Data – 2016 Edition*, 2016,
URL: http://thegovlab.org/govlab-index-on-open-data-2016-edition/
(visited on 08/02/2017) (cit. on p. 4).

[15] M. Ballard, *Poor data quality hindering government open data programme*, 2014,
URL: http://www.computerweekly.com/news/2240227682/Poor-data-quality-hindering-government-open-data-transparency-programme (visited on
08/02/2017) (cit. on pp. 4, 53).

[16] J. Hendler, *Data integration for heterogenous datasets*, Big data **2** (2014) 205
(cit. on p. 4).

[17] F. A. Musyaffa, F. Orlandi, H. Jabeen and M.-E. Vidal,
"Classifying Data Heterogeneity within Budget and Spending Open Data.",
ICEGOV, ACM, 2018 81, URL: http://dblp.uni-trier.de/db/conf/icegov/icegov2018.html#MusyaffaOJV18 (cit. on pp. 4, 151).

[18] R. B. Denhardt, *Public administration: an action orientation*, en, 7th Ed,
Cengage Learning, 2013, ISBN: 978-1-133-93921-4 (cit. on p. 17).

[19] D. F. Kettl, *Politics of the Administrative Process*, Cq Press, 2016 (cit. on p. 17).

[20] D. H. Rosenbloom, *Public administration: an action orientation*, en,
McGraw-Hill, 1993, ISBN: 978-0-070-53937-2 (cit. on p. 17).

[21] Open Knowledge Foundation, *What is open?*, 2012,
URL: https://okfn.org/opendata/ (visited on 13/02/2020) (cit. on p. 18).

[22] Open Data Barometer, *Global Report: Open Data Barometer*, 2017, URL:
http://opendatabarometer.org/4thedition/report/ (visited on 29/12/2017)
(cit. on pp. 18, 97).

[23] United Nations Statistics Division (UNSD),
Classification of the Functions of Government (COFOG), 1999,
URL: https://unstats.un.org/unsd/iiss/Classification-of-the-Functions-of-Government-COFOG.ashx (visited on 27/05/2019) (cit. on pp. 20, 38, 78).

[24] European Commission, *Information System for European Public Procurement: Common Procurement Vocabulary*, 2008,
URL: https://simap.ted.europa.eu/cpv (visited on 27/05/2019)
(cit. on pp. 20, 123, 143).

[25] S. Bratt, *Semantic web and other W3C technologies to watch*,
Talks at W3C, January, 2007,
URL: hhttps://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb/#(1)
(cit. on p. 21).

[26] F. Manola and E. Miller, eds., *RDF Primer*, W3C Recommendation,
World Wide Web Consortium, 2004, URL: http://www.w3.org/TR/rdf-primer/
(cit. on p. 20).

[27] T. Berners-Lee, *Linked Data - Design Issues*, (2010), URL:
https://www.w3.org/DesignIssues/LinkedData.html (visited on 28/03/2017)
(cit. on pp. 22, 46, 47, 49, 97, 98).

[28] S. S. Alonso and E. G. Barriocanal, *Making use of upper ontologies to foster interoperability between SKOS concept schemes.*,
Online Information Review **30** (2009) 263,
URL: http://dblp.uni-trier.de/db/journals/oir/oir30.html#AlonsoB06
(cit. on p. 22).

[29] M. Dudáš, J. Klimek, J. Kucera, J. Mynarz, L. Sedmihradska and J. Zbranek,
Deliverable 1.5: Final release of data definitions for public finance data, 2016,
URL: http://openbudgets.eu/assets/deliverables/D1.5.pdf
(cit. on pp. 23, 78, 79).

[30] M. DUDÁŠ, J. KLÍMEK, J. KUČERA, J. MYNARZ, L. SEDMIHRADSKÁ,
J. ZBRANEK and B. SEEGER,
The OpenBudgets Data Model and The Surrounding Landscape, (),
URL: https://openbudgets.eu/resources/2016/11/17/open-budgets-data-model-and-landscape/ (cit. on pp. 23, 140).

[31] D. R. (ed), *The Organization Ontology*, https://www.w3.org/TR/vocab-org/,
(Accessed on 06/08/2020), 2014 (cit. on p. 23).

[32] F. Bauer and M. Kaltenböck, *Linked open data: The essentials*,
Edition mono/monochrom, Vienna (2011) (cit. on pp. 23, 46).

[33] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z. Ives,
"DBpedia: A Nucleus for a Web of Open Data",
Proceedings of 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference (ISWC+ASWC 2007), 2008 722 (cit. on p. 23).

[34] D. Vrandečić and M. Krötzsch, *Wikidata: A Free Collaborative Knowledgebase*,
Commun. ACM **57** (2014) 78, ISSN: 0001-0782,
URL: http://doi.acm.org/10.1145/2629489 (cit. on p. 24).

[35] lod-cloud.net, *The Linked Open Data Cloud*, https://lod-cloud.net/, (Accessed on 06/09/2020) (cit. on p. 24).

[36] A. Miles and S. Bechhofer, *SKOS Simple Knowledge Organization System Reference*, World Wide Web Consortium, Working Draft WD-skos-reference-20080829, 2008 (cit. on pp. 26, 79).

[37] R. Cyganiak, D. Reynolds and J. Tennison, *The RDF Data Cube Vocabulary*, URL: https://www.w3.org/TR/vocab-data-cube/ (cit. on pp. 28, 29).

[38] J. Mynarz, V. Svátek, S. Karampatakis, J. Klímek and C. Bratsas, "Modeling fiscal data with the Data Cube Vocabulary.", *SEMANTiCS (Posters, Demos, SuCCESS)*, vol. 1695, CEUR Workshop Proceedings, CEUR-WS.org, 2016, URL: http://dblp.uni-trier.de/db/conf/i-semantics/semantics2016p.html#MynarzSKKB16 (cit. on pp. 29, 101).

[39] E. Prud'hommeaux and A. Seaborne, *SPARQL Query Language for RDF*, W3C Recommentation, W3C, 2008, URL: http://www.w3.org/TR/rdf-sparql-query/ (cit. on p. 29).

[40] A. Lauret, *OpenAPI Map*, http://openapi-map.apihandyman.io/?version=3.0, (Accessed on 06/08/2020) (cit. on pp. 30, 31).

[41] NIST Big Data Working Group (NBD-WG), *NIST Big Data Definitions and Taxonomies*, (2013), URL: http://bigdatawg.nist.gov/_uploadfiles/M0142_v1_3364795506.docx (visited on 18/02/2020) (cit. on p. 32).

[42] O. Curé and G. Blin, *RDF Database Systems: Triples Storage and SPARQL Query Processing*, Morgan Kaufmann, 2015, ISBN: 978-0-12-799957-9 (cit. on p. 32).

[43] J. L. Hennessy and D. A. Patterson, *Computer architecture: a quantitative approach*, Elsevier, 2011 (cit. on p. 32).

[44] M. Zaharia, R. S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. J. Franklin, A. Ghodsi, J. Gonzalez, S. Shenker and I. Stoica, *Apache Spark: A Unified Engine for Big Data Processing*, Commun. ACM **59** (2016) 56, ISSN: 0001-0782, URL: http://doi.acm.org/10.1145/2934664 (cit. on pp. 32, 114, 143).

[45] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauly, M. J. Franklin, S. Shenker and I. Stoica, "Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing.", *NSDI*, ed. by S. D. Gribble and D. Katabi, USENIX Association, 2012 15, URL: http://dblp.uni-trier.de/db/conf/nsdi/nsdi2012.html#ZahariaCDDMMFSS12 (cit. on p. 32).

[46] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker and I. Stoica, "Spark: Cluster Computing with Working Sets.", *HotCloud*, ed. by E. M. Nahum and D. Xu, USENIX Association, 2010, URL: http://dblp.uni-trier.de/db/conf/hotcloud/hotcloud2010.html#ZahariaCFSS10 (cit. on p. 32).

[47] C. S. of Civil Society Report, *Civicus State of Civil Society Report 2017 - Year in Review: New Democratic Crisis and Civic Space*, (Accessed on 03/16/2020), 2017, URL: https://civicus.org/documents/reports-and-publications/SOCS/2017/year-in-review/new-democratic-crisis.pdf (cit. on p. 33).

[48] International Budget Partnership, *Open Budget Survey 2017*, (2018), URL: https://www.internationalbudget.org/sites/default/files/2020-04/2017_Report_EN.pdf (cit. on pp. 33–37).

[49] OKI, *Global Open Data Index - Methodology*, 2015, URL: http://index.okfn.org/methodology/ (visited on 16/12/2016) (cit. on pp. 33, 37, 46, 47).

[50] International Budget Partnership, *Open Budget Survey 2019*, (2020), URL: https://www.internationalbudget.org/sites/default/files/2020-04/2019_Report_EN.pdf (cit. on pp. 34–36).

[51] The UN Guiding Principles Reporting Framework, *CIVIL SOCIETY ORGANIZATIONS (CSOS) : UN Guiding Principles Reporting Framework*, (Accessed on 03/16/2020), URL: https://www.ungpreporting.org/glossary/civil-society-organizations-csos/ (cit. on p. 35).

[52] Caucus of Development NGO Networks (CODE-NGO), *Civil Society Index: A Philippine Assessment Report*, (Accessed on 03/16/2020), 2011, URL: http://www.ombudsman.gov.ph/UNDP4/wp-content/uploads/2013/03/CSI-Report_Bookv.pdf (cit. on p. 35).

[53] J. Gray, *Open budget data: Mapping the landscape*, Available at SSRN 2654878 (2015) (cit. on p. 38).

[54] *The Open Definition - Open Definition - Defining Open in Open Data, Open Content and Open Knowledge*, (Accessed on 03/09/2020), URL: https://opendefinition.org/ (cit. on p. 38).

[55] *The 8 Principles of Open Government Data (OpenGovData.org)*, (Accessed on 03/09/2020), URL: https://opengovdata.org/ (cit. on p. 38).

[56] *Linked Data Glossary*, (Accessed on 03/09/2020), URL: https://dvcs.w3.org/hg/gld/raw-file/default/glossary/index.html#x5-star-linked-open-data (cit. on p. 38).

[57] J. Wonderlich, *Ten principles for opening up government information*, Washington, DC: Sunlight Foundation. August **11** (2010) 2010 (cit. on pp. 38, 46–48).

[58] C. Dener and S. Young (Sandy) Min, *Financial management information systems and open budget data: do governments report on where the money goes?*, The World Bank, 2013 (cit. on p. 38).

[59] M. S. De Clerck and T. Wickens, *Government Finance Statistics Manual 2014*, International Monetary Fund, 2014 (cit. on pp. 38, 78).

[60] W. Kim and J. Seo, *Classifying schematic and data heterogeneity in multidatabase systems*, Computer **24** (1991) 12, ISSN: 1558-0814 (cit. on p. 38).

[61] A. L. Machado and J. M. P. de Oliveira, "DIGO: An Open Data Architecture for e-Government.", *EDOCW*, IEEE Computer Society, 2011 448, ISBN: 978-1-4577-0869-5, URL: http://dblp.uni-trier.de/db/conf/edoc/edoc2011w.html#MachadoO11 (cit. on p. 39).

[62] K. Höffner, M. Martin and J. Lehmann, *LinkedSpending: OpenSpending becomes Linked Open Data.*, Semantic Web **7** (2016) 95, URL: http://dblp.uni-trier.de/db/journals/semweb/semweb7.html#HoffnerML16 (cit. on pp. 39, 141).

[63] P. Espinoza-Arias, M. J. Fernández-Ruiz, V. Morlán-Plo, R. Notivol-Bezares and O. Corcho, *The Zaragoza's Knowledge Graph: Open Data to Harness the City Knowledge*, Information **11** (2020) 129, URL: https://doi.org/10.3390%2Finfo11030129 (cit. on pp. 40, 41).

[64] J. Gracia, E. Montiel-Ponsoda, P. Cimiano, A. Gómez-Pérez, P. Buitelaar and J. P. McCrae, *Challenges for the multilingual Web of Data.*, J. Web Semant. **11** (2012) 63, URL: http://dblp.uni-trier.de/db/journals/ws/ws11.html#GraciaMCGBM12 (cit. on p. 41).

[65] B. Ahrendt, *What are the costs and benefits of XBRL in the financial services industry?*, Erasmus University, 2009 (cit. on p. 45).

[66] D. Lucas, *Evaluating the government as a source of systemic risk*, Journal of Financial Perspectives **2** (2014) 45, URL: https://EconPapers.repec.org/RePEc:ris:jofipe:0048 (cit. on pp. 45, 46).

[67] D. Lucas, *Valuation of government policies and projects*, Annu. Rev. Financ. Econ. **4** (2012) 39 (cit. on p. 45).

[68] T. Davies, *Open data barometer: 2013 global report*, World Wide Web Foundation and Open Data Institute (2013) (cit. on pp. 46, 47, 139, 151).

[69] M. Peters, A. Alberts and B. Seeger,
European Structural Funds - A Data Quality Index, 2017,
URL: http://openbudgets.eu/post/2017/04/04/esif-data-quality/
(visited on 03/05/2017) (cit. on p. 46).

[70] OKI, *How to Open up Data*, 2015,
URL: http://opendatahandbook.org/guide/en/how-to-open-up-data/
(visited on 07/12/2016) (cit. on pp. 46, 52).

[71] J. Tauberer, *Open government data*, Joshua Tauberer, 2012,
URL: https://opengovdata.io/ (cit. on pp. 46, 52).

[72] D. Dietrich, J. Gray, T. McNamara, A. Poikola, P. Pollock, J. Tait and T. Zijlstra,
Open data handbook, 2009 (cit. on pp. 46, 47).

[73] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann and S. Auer,
Quality assessment for linked data: A survey, Semantic Web **7** (2015) 63
(cit. on pp. 46, 47).

[74] CMMI Institute, *Data Management Maturity Model V1.0*, 2014,
URL: http://cmmiinstitute.com/data-management-maturity (visited on
12/10/2017) (cit. on p. 46).

[75] A. Fung, M. Graham and D. Weil,
Full disclosure: The perils and promise of transparency,
Cambridge University Press, 2007 (cit. on p. 46).

[76] A. Vetrò, M. Torchiano, C. Minotas Orozco, G. Procaccianti, R. Iemma and
F. Morando, *An Exploratory Empirical Assessment of Italian Open Government
Data Quality With an eye to enabling linked open data*, (2014) (cit. on p. 47).

[77] R. Caplan, T. Davies, A. Wadud, S. Verhulst, J. Alonso and H. Farhan, *Towards
common methods for assessing open data: workshop report & draft framework*,
World Wide Web Foundation (2014) (cit. on p. 47).

[78] Sunlight Foundation, *Open Data Policy Guidelines*, 2014,
URL: http://sunlightfoundation.com/opendataguidelines/ (visited on
07/12/2016) (cit. on pp. 47, 52, 97, 98, 139, 151).

[79] OKI, *Open Data Handbook: File Formats*, 2015,
URL: http://opendatahandbook.org/guide/en/appendices/file-formats/
(visited on 31/03/2017) (cit. on p. 47).

[80] W. Jaquith, *API versus Bulk Data*, 2015,
URL: https://how-to.usopendata.org/en/latest/The-Basics-of-Open-Data/API-vs-Bulk-Data/ (visited on 19/12/2016) (cit. on pp. 49, 51).

[81] R. Guenther and J. Radebaugh, *Understanding metadata*,
National Information Standard Organization (NISO) Press, Bethesda, USA (2004)
(cit. on p. 51).

[82] F. Maali, J. Erickson and P. Archer, *Data catalog vocabulary (DCAT)*,
W3C Recommendation (2014) (cit. on pp. 52, 61, 86).

[83] UK HM Government, *Open Data White Paper Unleashing the Potential*, 2012, URL: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/78946/CM8353_acc.pdf (visited on 07/12/2016) (cit. on p. 52).

[84] R. Pollock, *We need distributed revision/version control for data*, (2010), URL: https://blog.okfn.org/2010/07/12/we-need-distributed-revisionversion-control-for-data/ (cit. on p. 52).

[85] R. Czyborra, *ISO 8859 Alphabet Soup*, 1998, URL: http://czyborra.com/charsets/iso8859.html (cit. on p. 61).

[86] D. Fensel, F. M. Facca, E. P. B. Simperl and I. Toma, *Semantic Web Services*. Springer, 2011, ISBN: 978-3-642-19192-3 (cit. on pp. 67, 70).

[87] N. A. B. Gray, "Performance of Java middleware - Java RMI, JAXRPC, and CORBA", *Australasian Workshop on Software and System Architectures (ASWEC 2005)*, 2005 (cit. on p. 68).

[88] R. Tsouroplis, M. Petychakis, I. Alvertis, E. Biliri and D. Askounis, "Community-based API Builder to manage APIs and their connections with Cloud-based Services", *Proceedings of the CAiSE 2015 Forum at the 27th International Conference on Advanced Information Systems Engineering co-located with 27th International Conference on Advanced Information Systems Engineering (CAiSE 2015), Stockholm, Sweden, June 10th, 2015*. 2015 (cit. on p. 68).

[89] H. van der Veer and A. Wiles, *Achieving technical interoperability*, European Telecommunications Standards Institute (2008) (cit. on p. 68).

[90] K. P. Sycara, M. Paolucci, A. Ankolekar and N. Srinivasan, *Automated discovery, interaction and composition of Semantic Web services.*, J. Web Sem. **1** (2005) (cit. on p. 68).

[91] H. Kubicek, R. Cimander and H. J. Scholl, *Organizational Interoperability in E-Government - Lessons from 77 European Good-Practice Cases*. Springer, 2011, ISBN: 978-3-642-22501-7 (cit. on p. 68).

[92] D. B. et al, *Web Services Architecture*, 2004 - accessed January 15, 2016 (cit. on p. 68).

[93] Q. Z. Sheng, X. Qiao, A. V. Vasilakos, C. Szabo, S. Bourne and X. Xu, *Web services composition: A decade's overview.*, Inf. Sci. **280** (2014) (cit. on p. 69).

[94] J. Rao and X. Su, "A Survey of Automated Web Service Composition Methods.", *SWSWPC*, ed. by J. Cardoso and A. P. Sheth, vol. 3387, Lecture Notes in Computer Science, Springer, 2005, ISBN: 3-540-24328-3 (cit. on p. 69).

[95] M. Thoma, T. Braun, C. Magerkurth and A.-F. Antonescu, "Managing things and services with semantics: A survey", *Network Operations and Management Symposium (NOMS), 2014 IEEE*, IEEE, 2014 (cit. on p. 69).

[96] W. Wahlster, "Semantic technologies for mass customization", *Towards the Internet of Services: The THESEUS Research Program*, Springer, 2014 (cit. on p. 69).

[97] S. McIlraith, T. Son and H. Zeng, *Semantic Web services*, Intelligent Systems, IEEE **16** (2001), ISSN: 1541-1672 (cit. on p. 69).

[98] D. Roman, J. de Bruijn, A. Mocan, H. Lausen, J. Domingue, C. Bussler and D. Fensel, "WWW: WSMO, WSML, and WSMX in a Nutshell.", *ASWC*, ed. by R. Mizoguchi, Z. Shi and F. Giunchiglia, vol. 4185, Lecture Notes in Computer Science, Springer, 2006, ISBN: 3-540-38329-8 (cit. on p. 70).

[99] D. Martin, M. Burstein, D. McDermott, S. McIlraith, M. Paolucci, K. Sycara, D. L. McGuinness, E. Sirin and N. Srinivasan, *Bringing Semantics to Web Services with OWL-S*, World Wide Web **10** (2007), ISSN: 1386-145X (cit. on p. 70).

[100] C. Pedrinaci, J. Domingue and A. P. Sheth, "Semantic Web Services.", *Handbook of Semantic Web Technologies*, ed. by J. Domingue, D. Fensel and J. A. Hendler, Springer, 2011, ISBN: 978-3-540-92913-0 (cit. on p. 70).

[101] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel and B. Wiswedel, "KNIME: The Konstanz Information Miner.", *GfKl*, ed. by C. Preisach, H. Burkhardt, L. Schmidt-Thieme and R. Decker, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, 2009, ISBN: 978-3-540-78239-1 (cit. on p. 75).

[102] K. Wolstencroft, R. Haines, D. Fellows, A. R. Williams, D. Withers, S. Owen, S. Soiland-Reyes, I. Dunlop, A. Nenadic, P. Fisher, J. Bhagat, K. Belhajjame, F. Bacall, A. Hardisty, A. N. de la Hidalga, M. P. B. Vargas, S. Sufi and C. A. Goble, *The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud.*, Nucleic Acids Research **41** (2013) (cit. on p. 75).

[103] T. Takase, S. Makino, S. Kawanaka, K. Ueno, C. Ferris and A. Ryman, *Definition languages for RESTful Web services: WADL vs. WSDL 2.0*, IBM Reasearch (2008) (cit. on p. 75).

[104] D. Fensel and C. Bussler, *The Web Service Modeling Framework WSMF.*, Electronic Commerce Research and Applications **1** (2004) (cit. on p. 75).

[105] D. Martin et al., "Bringing Semantics to Web Services: The OWL-S Approach", *Proceedings of the First International Workshop on Semantic Web Services and Web Process Composition (SWSWPC 2004)*, ed. by J. Cardoso and A. P. Sheth, vol. 3387, LNCS, Springer, 2004, ISBN: 3-540-24328-3 (cit. on p. 75).

[106] D. Fensel, H. Lausen, A. Polleres, J. D. Bruijn, M. Stollberg, D. Roman and J. Domingue, eds.,
Enabling Semantic Web Services: The Web Service Modeling Ontology,
Springer-Verlag, 2006 (cit. on p. 76).

[107] R. Akkiraju, J. Farrell, J.Miller, M. Nagarajan, M. Schmidt, A. Sheth and K. Verma, *Web Service Semantics - WSDL-S, " A joint UGA-IBM Technical Note, version 1.0*, tech. rep., 2005 (cit. on p. 76).

[108] J. Kopecky, T. Vitvar, C. Bournez and J. Farrell,
SAWSDL: Semantic Annotations for WSDL and XML Schema,
IEEE Internet Computing **11** (2007), ISSN: 1089-7801 (cit. on p. 76).

[109] T. Vitvar, J. Kopecky, J. Viskova and D. Fensel,
"WSMO-Lite Annotations for Web Services",
Proceedings of the 5th European Semantic Web Conference,
ed. by M. Hauswirth, M. Koubarakis and S. Bechhofer, LNCS,
Springer Verlag, 2008 (cit. on p. 76).

[110] D. Roman, J. Kopecký, T. Vitvar, J. Domingue and D. Fensel, *WSMO-Lite and hRESTS: Lightweight semantic annotations for Web services and RESTful APIs.*, J. Web Sem. **31** (2015) (cit. on p. 76).

[111] M. Maleshkova, C. Pedrinaci and J. Domingue,
Supporting the creation of semantic restful service descriptions, (2009)
(cit. on p. 76).

[112] J. C. Bertot and H. Choi,
"Big data and e-government: issues, policies, and recommendations", *Proceedings of the 14th Annual International Conference on Digital Government Research*, ACM, 2013 1 (cit. on pp. 78, 118).

[113] L. Ding, D. DiFranzo, A. Graves, J. R. Michaelis, X. Li, D. L. McGuinness and J. Hendler, "Data-gov wiki: Towards linking government data",
2010 AAAI Spring Symposium Series, 2010 (cit. on p. 78).

[114] P. Walsh, R. Pollock, T. Björgvinsson, S. Bennett, A. Kariv and D. Fowler,
Fiscal Data Package (v. 0.3.0), 2016,
URL: http://specs.frictionlessdata.io/fiscal-data-package/ (visited on 18/09/2017) (cit. on pp. 78, 101).

[115] U. N. S. Office, *Provisional Central Product Classification*, vol. 77,
United Nations Publications, 1991 (cit. on p. 78).

[116] M. Dudaš, L. Horáková, J. Klímek, J. Kučera, J. Mynarz, L. Sedmihradská, J. Zbranek and T. Dong,
Deliverable 1.4: (OpenBudgets.eu Data Model) User documentation, tech. rep., OpenBudgets.eu Consortium, 2015,
URL: http://openbudgets.eu/assets/deliverables/D1.4.pdf
(cit. on pp. 79, 120).

[117] J. Mynarz, J. Klímek, M. Dudáš, P. Škoda, C. Engels, F. A. Musyaffa and V. Svátek, "Reusable transformations of Data Cube Vocabulary datasets from the fiscal domain", *Proceedings of the 4th International Workshop on Semantic Statistics co-located with 15th International Semantic Web Conference*, 2016 (cit. on pp. 83, 86, 87, 92, 95, 102).

[118] R. Albertoni, D. Brownind, S. Cox, A. G. Beltran, A. Perego and P. Winstansley, *Data Catalog Vocabulary (DCAT) - Version 2*, W3C Recommendation (2020), URL: https://www.w3.org/TR/vocab-dcat-2/ (cit. on p. 86).

[119] J. Klímek, P. Skoda and M. Necaský,
"LinkedPipes ETL: Evolved Linked Data Preparation.", *ESWC (Satellite Events)*, vol. 9989, LNCS, 2016, ISBN: 978-3-319-47601-8, URL:
http://dblp.uni-trier.de/db/conf/esws/eswc2016s.html#KlimekSN16
(cit. on pp. 88, 102).

[120] E. C. I. S. for European Public Administrations (ISA) working group et al., *DCAT application profile for data portals in Europe*, 2015,
URL: https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe/release/11 (cit. on p. 89).

[121] J. Attard, F. Orlandi, S. Scerri and S. Auer,
A systematic review of open government data initiatives.,
Government Information Quarterly **32** (2015) 399, URL:
http://dblp.uni-trier.de/db/journals/giq/giq32.html#AttardOSA15
(cit. on pp. 97, 98, 104, 118, 201).

[122] F. Gökgöz, S. Auer and J. Takis,
D4.2: Analysis of the Required Functionality of OpenBudgets.eu,
http://openbudgets.eu/assets/deliverables/D4.2.pdf,
URL: http://openbudgets.eu/assets/deliverables/D4.2.pdf
(cit. on pp. 98, 104, 201, 203).

[123] L. Ioannidis, C. Bratsas, S. Karabatakis, P. Filippidis and P. Bamidis, "Rudolf: An HTTP API for exposing semantically represented fiscal OLAP cubes.", *SMAP*, IEEE, 2016 177, ISBN: 978-1-5090-5246-2, URL: http://dblp.uni-trier.de/db/conf/smap/smap2016.html#IoannidisBKFB16 (cit. on p. 103).

[124] F. A. Musyaffa, C. Engels, M.-E. Vidal, F. Orlandi and S. Auer,
Experience: Open Fiscal Datasets, Common Issues, and Recommendations.,
J. Data and Information Quality **9** (2018) 19:1, URL: http://dblp.uni-trier.de/db/journals/jdiq/jdiq9.html#MusyaffaEVOA18 (cit. on pp. 104, 201).

[125] F. Orlandi, T. Dong, S. Karampatakis, P. Hernandez, F. Musyaffa and H. Liu, *D7.7 Large-scale Trial Report Including Best Practices*, http://openbudgets.eu/assets/deliverables/D7.7.pdf, URL: http://openbudgets.eu/assets/deliverables/D7.7.pdf (cit. on p. 105).

[126] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes and J. Dean, *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation.*, CoRR **abs/1609.08144** (2016), URL: http://dblp.uni-trier.de/db/journals/corr/corr1609.html#WuSCLNMKCGMKSJL16 (cit. on p. 114).

[127] *py_stringmatching* Documentation, *User Manual for py_stringmatching*, 2016, URL: http://anhaidgroup.github.io/py_stringmatching/v0.4.x/index.html (visited on 29/01/2019) (cit. on pp. 115, 117, 122).

[128] C. D. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008, URL: http://www-csli.stanford.edu/%5C%7Ehinrich/information-retrieval-book.html (cit. on p. 115).

[129] M. Odell and R. Russell, *Patent Numbers 1261167 (1918) and 1435663 (1922)*, Washington, DC: US Patent Office (1918) (cit. on p. 115).

[130] A. Doan, A. Y. Halevy and Z. G. Ives, *Principles of Data Integration*, Morgan Kaufmann, 2012, ISBN: 978-0-12-416044-6, URL: http://research.cs.wisc.edu/dibook/ (cit. on pp. 115–117, 122, 130).

[131] J. Zobel and P. Dart, "Phonetic String Matching: Lessons from Information Retrieval", *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, ACM, 1996 166, ISBN: 0-89791-792-8, URL: http://doi.acm.org/10.1145/243199.243258 (cit. on p. 115).

[132] I. Bartolini, P. Ciaccia and M. Patella, "String Matching with Metric Trees Using an Approximate Distance.", *SPIRE*, ed. by A. H. F. Laender and A. L. Oliveira, vol. 2476, Lecture Notes in Computer Science, Springer, 2002 271, ISBN: 3-540-44158-1, URL: http://dblp.uni-trier.de/db/conf/spire/spire2002.html#BartoliniCP02 (cit. on p. 116).

[133] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals", *Soviet physics doklady*, vol. 10, 8, 1966 707 (cit. on p. 116).

[134] M. A. Jaro, *UNIMATCH, a Record Linkage System: Users Manual*, Bureau of the Census, 1980 (cit. on p. 116).

[135] W. E. Winkler, *Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage*, tech. rep., Statistical Research Division, U.S. Census Bureau, Washington, DC, 1993 (cit. on p. 116).

[136] A. Cohen, *FuzzyWuzzy: Fuzzy String Matching in Python*, 2011, URL: https://chairnerd.seatgeek.com/fuzzywuzzy-fuzzy-string-matching-in-python/ (visited on 25/10/2018) (cit. on p. 116).

[137] Y. Jiang, G. Li, J. Feng and W.-S. Li, *String similarity joins: An experimental evaluation*, Proceedings of the VLDB Endowment **7** (2014) 625 (cit. on p. 117).

[138] T. Sørensen, *A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons*, Biol. Skr. **5** (1948) 1 (cit. on p. 117).

[139] P. Jaccard, *Étude comparative de la distribution florale dans une portion des Alpes et des Jura*, Bulletin del la Société Vaudoise des Sciences Naturelles **37** (1901) 547 (cit. on p. 117).

[140] M. Vijaymeena and K. Kavitha, *A survey on similarity measures in text mining*, Mach. Learn. Appl. Int. J **3** (2016) 19 (cit. on p. 117).

[141] A. Tversky, *Features of Similarity*, Psychological Review **84** (1977) (cit. on p. 117).

[142] B.-W. On, N. Koudas, D. Lee and D. Srivastava, "Group Linkage.", *ICDE*, ed. by R. Chirkova, A. Dogac, M. T. Özsu and T. K. Sellis, IEEE Computer Society, 2007 496, ISBN: 1-4244-0802-4, URL: http://dblp.uni-trier.de/db/conf/icde/icde2007.html#OnKLS07 (cit. on p. 117).

[143] A. E. Monge, C. Elkan et al., "The Field Matching Problem: Algorithms and Applications.", *KDD*, 1996 267 (cit. on p. 117).

[144] M. Bilenko, R. J. Mooney, W. W. Cohen, P. Ravikumar and S. E. Fienberg, *Adaptive Name Matching in Information Integration.*, IEEE Intelligent Systems **18** (2003) 16, URL: http://dblp.uni-trier.de/db/journals/expert/expert18.html#BilenkoMCRF03 (cit. on p. 117).

[145] J. Firth, *A synopsis of linguistic theory 1930-1955*, Studies in Linguistic Analysis, Philological, Longman, 1957 (cit. on p. 118).

[146] S. Deerwester, S. Dumais, G. Furnas, T. Landauer and R. Harshman, *Indexing by latent semantic analysis*, Journal of the American Society for Information Science **41** (1990) 391 (cit. on p. 118).

[147] K. Lund, "Semantic and associative priming in high-dimensional semantic space", *Proc. of the 17th Annual conferences of the Cognitive Science Society, 1995*, 1995 (cit. on p. 118).

[148] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space", *ICLR (Workshop Poster)*, ed. by Y. Bengio and Y. LeCun, 2013, URL: http://dblp.uni-trier.de/db/conf/iclr/iclr2013w.html#abs-1301-3781 (cit. on p. 119).

[149] A. Joulin, E. Grave, P. Bojanowski and T. Mikolov, *Bag of Tricks for Efficient Text Classification.*, CoRR **abs/1607.01759** (2016), URL: http://dblp.uni-trier.de/db/journals/corr/corr1607.html#JoulinGBM16 (cit. on p. 119).

[150] A. Conneau, G. Lample, M. Ranzato, L. Denoyer and H. Jégou, *Word Translation Without Parallel Data.*, CoRR **abs/1710.04087** (2017), URL: http://dblp.uni-trier.de/db/journals/corr/corr1710.html#abs-1710-04087 (cit. on p. 119).

[151] A. Pappu, R. Blanco, Y. Mehdad, A. Stent and K. Thadani, "Lightweight Multilingual Entity Extraction and Linking.", *WSDM*, ed. by M. de Rijke, M. Shokouhi, A. Tomkins and M. Zhang, ACM, 2017 365, ISBN: 978-1-4503-4675-7, URL: http://dblp.uni-trier.de/db/conf/wsdm/wsdm2017.html#PappuBMST17 (cit. on pp. 119, 120).

[152] D. Moussallem, R. Usbeck, M. Röder and A.-C. N. Ngomo, "MAG: A Multilingual, Knowledge-base Agnostic and Deterministic Entity Linking Approach.", *K-CAP*, ed. by Ò. Corcho, K. Janowicz, G. Rizzo, I. Tiddi and D. Garijo, ACM, 2017 9:1, ISBN: 978-1-4503-5553-7, URL: http://dblp.uni-trier.de/db/conf/kcap/kcap2017.html#MoussallemURN17 (cit. on p. 120).

[153] T. Lesnikova, *Liage de données RDF : évaluation d'approches interlingues. (RDF Data Interlinking : evaluation of Cross-lingual Methods).*, PhD thesis: Grenoble Alpes University, France, 2016 (cit. on p. 120).

[154] H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann and S. Hübner, "Ontology-Based Integration of Information - A Survey of Existing Approaches.", *OIS@IJCAI*, ed. by A. Gómez-Pérez, M. Gruninger, H. Stuckenschmidt and M. Uschold, vol. 47, CEUR Workshop Proceedings, CEUR-WS.org, 2001, URL:

http://dblp.uni-trier.de/db/conf/ijcai/ois2001.html#WacheVVSSNH01 (cit. on p. 120).

[155] J. Volz, C. Bizer, M. Gaedke and G. Kobilarov, "Silk - A Link Discovery Framework for the Web of Data.", *LDOW*, ed. by C. Bizer, T. Heath, T. Berners-Lee and K. Idehen, vol. 538, CEUR Workshop Proceedings, CEUR-WS.org, 2009, URL: http://dblp.uni-trier.de/db/conf/www/ldow2009.html#VolzBGK09 (cit. on p. 121).

[156] S. Karampatakis, C. Bratsas, O. Zamazal, P. M. Filippidis and I. Antoniou, *Alignment: A Hybrid, Interactive and Collaborative Ontology and Entity Matching Service*, Information **9** (2018), ISSN: 2078-2489, URL: http://www.mdpi.com/2078-2489/9/11/281 (cit. on p. 121).

[157] Y. Umaoka, M. Davis and A. Phillips, *BCP 47 Extension T-Transformed Content*, (2012) (cit. on p. 143).

APPENDIX A

JDIQ Questionnaire: Important factors in open fiscal datasets

Fiscal datasets are datasets that provide information about any public government financial matters, for example budget, spending, contracts, and grants beneficiaries. Fiscal datasets are important to support public financial transparency (which will promote political trust and engagement), to persuade citizens to be more proactive in the budget legislation process, and to create stories for journalists who report on fiscal data. Fiscal data publication also plays a role in the reduction of fiscal malpractice.

This survey is a part of works in the OpenBudgets.eu project (http://openbudgets.eu/). OpenBudgets.eu is an European Union's H2020 EU research and innovation programme.

The goal of this survey is to find the important factors that should exist in open fiscal datasets. Our aim is to score the importance of these factors and to subsequently:

- Create an assessment framework for open fiscal datasets.
- Rank open fiscal datasets.
- Compare the ranking result with rankings done by OK-GODI (http://index.okfn.org/) and Open Data Barometer (http://opendatabarometer.org/).
- Establish a guideline for officials to publish open fiscal datasets.

Please check if the following statements apply to you:

- ☐ I am familiar with open data.
- ☐ I am familiar with the linked data concept.
- ☐ I have created fiscal datasets.
- ☐ I am a consumer of open fiscal datasets.
- ☐ I have worked on fiscal datasets.

Important factors in open fiscal datasets In the perspective of data consumers, please rate the following factors that should be available on fiscal datasets. The ratings are the following:

- 1 = Not at all important
- 2 = Slightly important
- 3 = Moderately important
- 4 = Important
- 5 = Extremely important

Data formats and access

- **Digital form**.
  The dataset is provided as digital file instead of a paper-based document.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important
- **Available online**.
  The dataset is accessible through the internet instead of a local machine or network/intranet.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important
- **Available in Bulk**.
  A dataset is available in bulk if the entire dataset can be downloaded easily and efficiently - instead of having only limited access with the need of many requests to get the entire dataset.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important
- **Free of charge**.
  The dataset is provided free of any fees.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important
- **Non-proprietary format**.
  The dataset is published in a non-proprietary format, e.g., CSV/ODS instead of XLS.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important
- **Structured data**.
  The dataset is published in a structured way, e.g., CSV instead of PDF.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

- **English / multi-lingual dataset**.
  The dataset is provided in multiple languages, especially in a commonly used language (e.g., English).
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

- **Semantic format**.
  The dataset is published using a semantic format, most fundamentally using any RDF serialization (e.g., *.ttl / turtle, *.rdf , JSON-LD).
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

- **Publicly available**.
  The dataset is available through public channels (e.g., website) instead of private channels (e.g., emails, personal cloud service links such as Dropbox).
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

- **Accessible via search mechanism**.
  A search mechanism is provided on the website so that users can easily search for a dataset.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

- **Dataset filtering**.
  A filtering can be applied to the dataset, so that consumers who are only interested in a subset of the dataset can easily query the dataset to get the data they need.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

- **API endpoint**.
  An Application Programming Interface (API) is provided to let programmers easily reuse the dataset for their application.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

Width

- **Multiple coverage**.
  The fiscal datasets are provided in different coverage, for example, the datasets include budget, spending, procurement, contracts, and beneficiaries.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

- **Multiple number of dimensions**.
  The dataset contains several dimensions and classifications, such as economic, functional, administrative, and geographic classifications, and so on.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

- **Granularity (non-aggregated observations)**.
  Transactions are provided in detail instead of aggregations. These details can be in the form of transaction records, or can be in the width of available dimensions/-classifications.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

- **Aggregation**.
  Is it necessary to provide an aggregated version of the dataset in addition to the granular dataset?
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

Assisting documents

- **Availability of documentation**.
  A documentation explaining the terms used in the dataset, how the data was collected, how the data was aggregated, and so on is provided together with the dataset.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

- **Availability of metadata**.
  Metadata that explains the properties of the dataset is available. This metadata can be structural metadata, descriptive metadata or any other type of metadata.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

- **Availability of visualizations**.
  Diagrams, charts, plot or other visualization methods are used to illustrate the dataset.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

- **Availability of stories or data interpretation that summarizes the data**.
  A dataset summary is provided to let readers know the overview of the dataset.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

- **Complete code lists**.
  A code list is a set of limited codes/terms that can be used in the dataset. For example, a code list of possible currency codes (EUR, GBP, and so on). All code lists used in the datasets are either publicly available, or published by the respective officials.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

Publication: originality, provenance, updates

- **Authoritative dataset**.
  The dataset is directly published by the officials.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

- **Primacy**.
  The dataset is source data or published together with the original information collected by the government, details on how the data was collected and the original source documents.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

- **Persistent URI**.
  The URI of the dataset is permanent, i.e. does not change over time.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

- **Timely datasets**.
  A dataset is released as quickly as possible, or before the context is expired (e.g., the dataset about an election should be released soon after all the votes have been counted).
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

- **Regular updates**.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

- **Permanence**.
  If a dataset is updated or changed the previous version remains online, e.g., via appropriate version-tracking.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

- **Sustainable publication**.
  The availability of a web platform or any other tool on the government portal or another trustworthy platform that allows to publish datasets periodically.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

- **Mentioned contributors**.
  The people who contributed on the dataset are mentioned.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

- **Availability of contact points**.
  Some form of communication channel to contact the dataset's maintainer is provided for questions or feedback regarding the dataset (e.g., email, contact from, phone, etc.).
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

License

- **Clearly mentioned license**.
  The dataset has a clearly mentioned license type, for example Creative Commons or Open Government License. Alternatively the data publisher's web page explicitly states the permission of accessing, reusing, and redistributing the dataset.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important
- **Open License**.
  The dataset is openly licensed which grants permission to access, re-use and redistribute with few or no restrictions.
  Not at all important ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 Extremely important

Feedback

- Are there other factors that should be available on fiscal datasets? Please mention them along with the importance, e.g., Accuracy - (5)..
- Do you have any ideas or feedback?

APPENDIX B

List of Surveyed Datasets

The following table only provides a condensed summary of open fiscal datasets survey as mentioned in chapter 4. Some initial entries are removed since the datasets had been deleted from the publisher's website. Due to the limitation of space in this thesis, the complete table is provided online.[1] The complete table provides the following additional information:

1. Dataset URL
2. If the datasets are structured
3. If datasets are published in an open format
4. Datasets' time span
5. Geographical coverage
6. Information covered
7. High-level data model (tabular, taxonomic graph...)
8. Financial data type (budget, spending, transaction, contract, ...)
9. Authoritativeness
10. Administrative level (municipal, regional, state, national supranational)
11. Granularity level
12. Categorization according to Tim Berners Lee's Five Star Data
13. If the dataset is free of charge
14. If the dataset is available in bulk
15. If the dataset is available publicly
16. If the dataset can be easily obtained
17. If data contributors are mentioned
18. The existence of search mechanisms on data portal
19. License name
20. If the dataset is published in an open license
21. if the dataset is updated regularly
22. Language
23. URI Persistence
24. Domain
25. Additional metadata availability
26. The completeness of mentioned code lists
27. Availability of RDF data
28. URL dereferencability, if RDF data is available
29. Whether serialized format is available (if RDF format is available)
30. Documentation availability
31. Contact point availability
32. Visualization availability
33. Version tracking availability
34. If sub-dataset query can be performed

[1] http://bit.ly/jdiq-datasheet-view

35. Further contextual documents and availability of stories or data interpretation that summarizes the data
36. API availability
37. How the OFDP, ODB and GODI scores are calculated
38. Additional concerns and commentary on each datasets

Table B.1: List of datasets inspected for data quality using GODI, OFDP, and OFDP methodology as well as corresponding quality scores.

ID	Dataset	Dataset "Group"	Coverage	Publisher	Level	File Format	Time	Financial Data Type	Granularity	OFDP Score (norm.)	ODB Score	GODI Score
1	Regional Development in United Kingdom (ESIF ERDF)	ERDF UK	Funding beneficiaries, spending	UK Department for Communities and Local Government	Multiple-region	PDF	2007 - 2013 budget, spending up to 2016	Grants (or Spending?)	granular, per project and beneficiary	76	80	85
2	Buxtehude (population of 39 000) in Niedersachsen	Buxtehude town Budget	Budget 2005-2016 in report format.	Hansestadt Buxtehude	Municipal	PDF	2005 - 2016	Budget	granular, per sub-thema	62	65	55
3	[GB] Greater London Authority Spending	London Expenditure	London expenditure	Greater London Authority	Municipal	CSV	2013 - 2016	Spending	granular per vendor and cost element	66	75	70
4	[GB] HMT Public Expenditure Statistical Analyses (PESA)	GB Budget and spending	Public Expenditure Statistical Analyses (PESA) is the yearly publication of information on government spending.	HM Treasury	National	XLS, XLSX, PDF	2010 - 2015	Statistics	aggregated	80	95	100
5	Whole of Government Accounts (WGA)	Consolidated expenditure and revenue	Comprehensive financial statements of UK's public sector	Github: @pudo	National	CSV, XLS, XLSX	2009 - 2011	Financial statements	unclear	66	65	60
6	[EU] Cohesion major project sample maps	Map visualization of EU projects locations	EU major projects and transport projects maps	European Commission	Supra-national	HTML	2007 - 2013	Budget	granular per project, country and fund	64	50	35
7	[EU] Landkreis Kelheim (population of 115 000) in Bayern Budget Data		Budget for year 2914, 2015. Really unstructured	Landratsamt Kelheim	Municipal	HTML	2014, 2015	Spending	unclear	50	60	45
8	Beneficiaries of CAP payments in United Kingdom	CAP GB Beneficiaries	Funding beneficiaries	UK Department for Environment Food & Rural Affairs	National	XLS	10.2013 - 10.2014	Spending	granular per person/beneficiary	77	95	100

ID	Dataset	Dataset "Group"	Coverage	Publisher	Level	File Format	Time	Financial Data Type	Granularity	OFDP Score	ODB Score	GODI Score
9	Slovakia Regional Development - Beneficiaries of European Union Cohesion Policy	EU Beneficiaries	List of NSRF beneficiaries in Slovakia, including implementation status	Centrálny koordinačný orgán NSRR	National	XLS	2007 - 2013	Spending	sub functional clasification is available. somewhat granular.	66	80	70
10	Beneficiaries From EAGF and the EAFRD in Slovakia	EAGF / EAFRD Beneficiaries	Information on the beneficiaries of EAGF and EAFRD (Drawing financies from the Funds)	Agricultural Paying Agency	National	HTML	2014	Spending	granular per person	59	50	35
11	Fisheries (Fishery Operational Programme Czech Republic)	Fisheries (Possibly EMFF)	List of approved fisheries project	Ministerstvo zemědělství	National	XLS	2010 - 2015	Spending	granular per name and project	66	80	70
12	Regional Development in Czech Republic	ESIF beneficiaries	List of beneficiaries from the programs funded by EU	Ministerstvo pro místní rozvoj ČR (Ministry of Regional Developmet CZ)	National	XLS, XLSX	2009 - 2016	Spending	granular per name and project	69	80	70
13	Agricultural policy – List of Subsidy Beneficiaries (Czech Republic)	Agricultural funding beneficiaries	Beneficiaries list of programs funded by EU, mostly jointly funded with Czech Government	Státní Zemědělský Intervenční Fond (The State Agricultural Intervention Fund)	National	HTML	2014 - 2015	Spending	granular per name, fund and measure	55	50	35
14	Expenditures in Libya 2011-2014	Post Libya revolution operation by EU	EU expense for operation in Libya	Council of the European Union	National	PDF	2012 - 2013	Spending	aggregated	45	45	40
15	Region of Köln (population of 1.000.000) in Nordrhein-Westfalen	Cologne budget	Income in 2013 and budget plan from 2014 - 2018	Stadt Köln	Municipal	CSV	2013 (Result), 2014-2018 (budget)	Budget	aggregated	64	75	90
16	[Germany] Update dataset of party finances	German Political Party Donations	List of people who donate to German political party.	Unclear	National	XLS,ODS,CSV,XLSX	1994 - 2009	Donation	granular per person and party	34	55	55
17	Open Budget municipality of Thessaloniki	CSV, HTML	-	Ηλεκτρονικής Διακυβέρνησης του Δήμου Θεσσαλονίκης	Municipal	CSV,XLS	2011 - 2015	Spending	aggregated	61	60	60

ID	Dataset	Dataset "Group"	Coverage	Publisher	Level	File Format	Time	Financial Data Type	Granularity	OFDP Score	ODB Score	GODI Score
18	[Italy] Transactions from the European structural funds	? (Language problem, possibly spending for projects funded by EU fund)	Payments under the European structural fund to projects in Italy for the period 2007-2013.	OpenCoesione	National	CSV	2007 - 2013	Spending	granular	83	95	100
19	Open Budget municipality of Athens	Athens Budget	Income, Expense by Departments	Δημαρχείο Λυοσίων (Athens City Hall)	Municipal	HTML, XLSX	2005 - 2016	Spending, Income	granular per department and project	69	60	60
20	[Portugal] Budget data	Portugal Budget Execution, Spending	Public expenditure and revenues, including budget changes	Direção-Geral do Orçamento	National	PDF, XLS	Various	Spending	aggregated	73	80	70
21	Fishery Operational Programme Slovakia	Beneficiaries of EF fisheries in Slovakia	List of Beneficiaries	Pôdohospodárska platobná agentúra (Agricultural Paying Agency)	National	some in PDF, few in XLS	2008 - 2011	Beneficiearies / Spending	granular per project and recipient	55	60	60
22	Nürnberg Budget D	Nürnberg budget	Household	Stadt Nürnberg	Municipal	PDF	2012 - 2016	Budget	unclear	64	65	55
23	Open Budget municipality of Madrid		Income, Expense, Investments	Ayuntamiento de Madrid (Madrid City Council)	Municipal	XLS, CSV	2012 - 2016	Spending, Income, Investments	aggregated budget	79	95	100
24	[Spain] FFE	Unsure, Likely to be spanish govt budget	Expense data from co-funded EU and Spain solidarity/migration spending	Unclear	National	CSV	2007 - 2012	Spending	aggregated	41	55	60
25	Presupuesto del Gobierno de Aragón	Budget	Budget	Gobierno De Aragon	State	CSV (data), RDF (metadata)	2006 - 2015	Budget	granular	84	97.5	100
26	Presupuesto de sociedades públicas pertenecientes a entidades locales de Aragón	Aragon Public companies Budget	Budget of Aragon's public companies	Gobierno De Aragon	State	CSV,JSON,XML	1986 - 2016	Budget	aggregated	69	90	90
27	Presupuestos consolidados de la Comunidad Autónoma de Aragón	Consolidated budget from 2007-2016, grouped per functional/economic classification	Budget classified from functional/economic point of view	Gobierno De Aragon	State	PX (data), RDF (metadata)	2007 - 2016	Budget	aggregated	83	97.5	100

ID	Dataset	Dataset "Group"	Coverage	Publisher	Level	File Format	Time	Financial Data Type	Granularity	OFDP Score	ODB Score	GODI Score
28	EJECUCIÓN PRESUPUESTARIA de la Comunidad Autónoma de Aragón	Spending	Budget classified from functional/economic point of view	Gobierno De Aragon	State	CSV (data), RDF (metadata)	2006 - 2015	Budget	semi aggregated (detailed but nit up to transaction level)	86	97.5	100
29	Proyecto de Presupuesto del Gobierno de Aragón	Budget draft	Budget draft	Gobierno De Aragon	State	CSV (data), RDF (metadata)	2014 - 2016	Budget	semi aggregated (detailed but nit up to transaction level)	84	97.5	100
30	Presupuesto y ejecución presupuestaria de Comarcas de Aragón	unclear (needs spanish speaker to interpret)	Execution and Budgeting from Council of Municipalities (Aragon, Spain)	Gobierno De Aragon	State	CSV,JSON,TXT,XML, XLS	2010 - 2013, 2014+	Budget	unclear, seems to be detailed but complicated to understand	83	95	100
31	FTS OCHA - International Aid	Humanitarian aid funding	Humanitarian aid funding report	Financial Tracking Service (FTS)	Supranational	XLS, PDF, CSV (beta portal)	2000 - 2016	Spending and Budget (status included)	granular per donor/recipient country and project	79	80	70
32	Breakdown of The Available Funds By Theme for 2007-2013	ESIF Budget	Breakdown of the available funds by theme and member states	European Commission	Supranational	CSV, JSON, PDF, RDF, RSS, XLS, XLSX, XML	2007 - 2013	Spending	granular per member states and theme	97	100	100
33	Breakdown Of The Available Funds By Thematic Objective By MS For 2014-2020	ESIF Budget	Budget by country, program and thematic objective	European Commission	Supranational	CSV, JSON, PDF, RDF, RSS, XLS, XLSX, XML	2014 - 2020	Budget	granular per funding program, member states and thema	97	100	100
34	Total EU allocations per MS - Transported (2014-2020)	EU Budget	EU Budget by country and fund	European Commission	Supranational	CSV, JSON, PDF, RDF, RSS, XLS, XLSX, XML	2014 - 2020	Budget	aggregated per member state and theme	89	85	70
35	US Federal Budget	US Budget	Budget (receipt/outlays) including the authorities responsible for the proposed outlays.	The Office of Management and Budget, US	National	CSV,PDF,XLS, XLSX	1962 - 2021 (Outlays/Receipt), 1976 - 2021 (Authority)	Budget	Granular per year, function, category, title, account etc	80	95	100
36	[Russia] Government Budgets		Budget execution categorized by year and functionality	The Ministry of Finance of the Russian Federation	National	XML, CSV, XLS, PDF, DOC	2007 - 2016	Spending, Budget	aggregated	58	60	60

ID	Dataset	Dataset "Group"	Coverage	Publisher	Level	File Format	Time	Financial Data Type	Granularity	OFDP Score	ODB Score	GODI Score
37	Budget of the European Union for the year 2014	EU Budget	Budget of the European Union 2014 adopted by the European Parliament	European Union Data Portal Publishing Office	Supranational	CSV,XML	2014	Budget	unclear	80	95	100
38	EU - Financial Transparency System	EU Grant (and contracts) beneficiaries	Beneficiaries of EU grants & support	EU Directorate-General for the Budget	Supranational	XLS, CSV	2007 - 2014	Spending	granular per recipient	75	65	70
39	[Moldova] Upload BOOST data for 2012	Moldova Spending as composed by BOOST project	Moldova Expenditure composed by the World Bank BOOST project.	?	National	CSV	2005 - 2010	Spending	granular to some extent	69	75	90
40	Big Lottery Fund data	Lottery Fund Beneficiaries	List of funding applicants for social causes.	Big Lottery Fund UK	National	CSV	2004 - 2015	Grant	granular per recipient (not explicit)	69	80	90
41	Tanzania Budget	Tanzanian Budget	Budget estimation for each ministry, institution, or region	The United Republic of Tanzania: Ministry of Finance	National	CSV, scraped from PDF, PDF originally available from	2011 - 2016	Budget	aggregated per ministry	76	65	55
42	[Catalonia, ES] Budget 2012 (current, no budget 2013)	Catalan Budgets	Catalonia Budget (past)	Generalitat de Catalunya	State	XLSX	2010 - 2012, 2014- 2015	Budget	granular	66	65	60
43	Poland Budget BOOST	World Bank	Poland's National Budget	BOOST at World Bank, originally Poland's Ministry of Finance's website	national, regional and municipal	XLSX, (originally PDF)	2004-2013	Budget	Budget is aggregated according to detailed classifications	79	80	60
44	Mexico Budget BOOST	World Bank	Mexico's National Budget	BOOST at World Bank, originally several different government sources	national, regional and municipal	XLSX, CSV	1989-2014	Budget	Rather aggregated versions, detailed one for 2014	80	65	60
45	Armenia Budget BOOST	World Bank	Armenia's National Budget	BOOST at World Bank, originally from Armenias's Ministry of Finance	national and regional	XLSX	2006-2015	Budget	Rather aggregated data	79	80	60
46	Moldova Budget BOOST	World Bank	Moldova's National Budget	BOOST at World Bank, originally from Fintehinform (together with Moldova's ministry of finance)	national and regional	XLSX, CSV	2005-2014	Budget	Budget is aggregated according to detailed classifications	79	65	60

ID	Dataset	Dataset "Group"	Coverage	Publisher	Level	File Format	Time	Financial Data Type	Granularity	OFDP Score	ODB Score	GODI Score
47	Paraguay Budget BOOST	World Bank	Paraguay's National Budget	BOOST at World Bank, constructed at the intitiative of World Bank country economist as part of the Public Expenditure Review.	unclear	XLSX	2003-2014	Budget	Rather aggregated data	79	65	60
48	Peru Budget BOOST	World Bank	Peru's National Budget	BOOST at World Bank, some data are provided by Ministry of Finance	national, regional and municipal	XLSX, CSV	1999-2015 (Central), 2004-2015 (Regional), 2007-2015 (Local)	Budget	granular	76	80	60
49	ESF Beneficiaries Austria	EU Funds	ESF Beneficiaries Austria	Bundesministerium für Arbeit, Soziales und Konsumentenschutz	national	PDF	2014	Funds	Data aggregated per beneficiary and program/-function.	60	45	55
50	Croatian Funds Beneficiaries	EU Funds	Croatian Funds Beneficiaries	Ministarstvo regionalnoga razvoja i fondova Europske unije	national	XLS	2008-2016 (?)	Funds	Data aggregated per project.	62	75	60
51	Cyprus Funds Beneficiaries	EU Funds	Cyprus Funds Beneficiaries	Γενική Διεύθυνση Ευρωπαϊκών Προγραμμάτων, Συντονισμού και Ανάπτυξης	national	XLSX	2008-2015 (?)	Funds	Data aggregated per project.	55	60	60
52	Czech Funds Beneficiaries	EU Funds	Czech Funds Beneficiaries	Ministry of Regional Development CZ	national	XLSX	2009-2015	Funds	Transactions (date of interim payment)	66	80	60
53	Danish Funds Beneficiaries	EU Funds	Danish Funds Beneficiaries	Regional Udvikling	national	CSV	2014-2020	Funds	Data aggregated per project.	64	75	60
54	Estonian Funds Beneficiaries	EU Funds	Estonian Funds Beneficiaries	Ministry of Finance of the Republic of Estonia	national	Web	2004-2016 (?)	Funds	Data aggregated per project/beneficiary.	48	30	35
55	Finnish Funds Beneficiaries	EU Funds	Finnish Funds Beneficiaries	Työ- ja elinkeinoministeriö (Ministry of Employment and Economy)	national	CSV	2014-2019	Funds	Data aggregated per project (?).	62	75	60
56	German Funds Beneficiaries	EU Funds	German Funds Beneficiaries	ESF: Bundesministerium für Arbeit und Soziales (BMAS)	national	PDF	2009-2014 (?)	Funds	Detailed data (transactions?).	50	45	45
57	Greek Funds Beneficiaries	EU Funds	Greek Funds Beneficiaries	Ministry of Economy, Development and Tourism	national	CSV	2009-2023 (?)	Funds	Somehow aggregated data.	66	80	70

ID	Dataset	Dataset "Group"	Coverage	Publisher	Level	File Format	Time	Financial Data Type	Granularity	OFDP Score	ODB Score	GODI Score
58	Italian Funds Beneficiaries	EU Funds	Italian Funds Beneficiaries	Dipartimento per le Politiche di Coesione, Presidenza del Consiglio dei Ministri (Department for Cohesion Policy)	national and regional	CSV + XLS	2007-2013	Funds	Granular, Very detailed data.	76	95	100
59	Latvian Funds Beneficiaries	EU Funds	Latvian Funds Beneficiaries	Latvijas Republikas Finanšu ministrija (Latvian Republic Ministry of Finance)	national	XLS (dataset), XLSX (code list)	2007-2013	Funds	Data aggregated per project.	66	80	60
60	Lithuanian Funds Beneficiaries	EU Funds	Lithuanian Funds Beneficiaries	Ministry of Finance of the Republic of Lithuania	national	XLS	2007-2013 (?)	Funds	Data aggregated per project (?), applicant mentioned	69	75	70
61	Luxembourg Funds Beneficiaries	EU Funds	Luxembourg Funds Beneficiaries	Centre des Technologies de l'Information de l'État/Gouvernement du Grand-Duché de Luxembourg (National Information Technology Center/Government of Luxembourg)	national	Web	2007-2013 & 2014-2020	Funds	Only web data for (selected?) projects available.	59	45	35
62	Maltese Funds Beneficiaries	EU Funds	Maltese Funds Beneficiaries	Planning and Priorities Co-ordination Division	national	Web	2014-2020	Funds	Only web data is available for the projects.	44	30	35
63	Dutch Funds Beneficiaries	EU Funds	Dutch Funds Beneficiaries	Sterc webpage / North Netherland Cooperation	national	CSV	2001-2016 (Project start date)	Funds	Data aggregated per project.	55	75	60
64	Polish Funds Beneficiaries	EU Funds	Polish Funds Beneficiaries	Centralny Punkt Informacyjny Funduszy Europejskich (Central European Funds Information Point), Ministry of Development	national	CSV	2004-2020	Funds	Data aggregated per project.	62	75	60
65	Slovene Funds Beneficiaries	EU Funds	Slovene Funds Beneficiaries	Republika Slovenija, služba vlade rs za razvoj in evropsko kohezijsko politiko (Republic of Slovenia, Government Office for Development and European cohesion policy)	national	XLSX	2007-2013	Funds	Data aggregated per project (?).	59	75	70
66	Spanish Funds Beneficiaries	EU Funds	Spanish Funds Beneficiaries	ministerio de hacienda y administraciones públicas	national	PDF	2007-2013	Funds	Rather granular	59	65	45

ID	Dataset	Dataset "Group"	Coverage	Publisher	Level	File Format	Time	Financial Data Type	Granularity	OFDP Score	ODB Score	GODI Score
67	Swedish Funds Beneficiaries	EU Funds	Swedish Funds Beneficiaries	Swedish Agency for Economic and Regional Growth	national	Web		Funds	Rather granular	65	65	55
68	Bonn Budget	City Budgets	Bonn Budget Draft	Stadt Bonn	Municipal	XLSX	2017-2024	Budget Draft	?	77	95	90
69	Austria National Budget	National Budgets	Austria National Budget	Bundesministerium für Finanzen	national	XLSX	2014-2015	Budget	Only different very aggregated tables are available.	74	80	70
70	Austria National Budget	National Budgets	Austria National Budget	Bundesministerium für Finanzen	national	XLSX	2016	Budget	Only different very aggregated tables are available.	74	80	70
71	Belgium National Budget	National Budgets	Belgium National Budget	Belgischer Federale Overheidsdiensten	national	XLS	2016	Budget	Seems to be quite detailed, several classifications	63	80	60
72	Bulgaria National Budget	National Budgets	Bulgaria National Budget	Министерство на финансите на Република България (Bulgarian Ministry of Finance)	national	XLS	2016	Budget	Highly aggregated	63	80	60
73	Bulgaria National Budget	National Budgets	Bulgaria National Budget	Министерство на финансите на Република България (Bulgarian Ministry of Finance)	national	XLS	2015	Budget	Highly aggregated	63	80	60
74	[Philippines] Procurement data	Government contract	Open government contract and awarded contract	Republic of the Philippines	National	HTML	2016	Contract		69	65	75
75	Projects funded from European Social Fund	ESF Projects	ESF-funded projects	Eurepean Commission	Supranational	HTML	2011 - 2016	Budget		68	50	45
76	Russian spending data (Clearspending.ru)	Russian gov customers, contracts, suppliers	Government Contracts, Supplier and Customer	Goszatraty project	National	HTML, JSON	unclear	contracts, supplier, consumer	granular per awarded contracts	71	75	70
77	UN: Budgetary Central Government - Expense	Worldwide Budget	Government Budgets (with Functional Classification)	IMF	Supranational	CSV	1991-2009	Budget	highly aggregated	69	65	60

APPENDIX C

Semantic Description of OpenAPI

```yaml
paths:
  /FindByMunicipalityYear:
    get:
      tags:
        - municipality, year
      summary: Finds data by year and municipality.
      description: Find published governmental budget or spending datasets based on municipality
      ↪  and year. <http://www.openbudgets.eu/ontology/obeu/Dataset>
      operationId: FindByMunicipalityYear
      produces:
        - application/json
      parameters:
        - in: query
          name: municipality
          description: Municipality name of corelated budget/spending data.
          ↪  <http://dbpedia.org/ontology/Municipality>
          required: true
          type: string
        - in: query
          name: year
          description: Year of budget/spending data.  <http://dbpedia.org/ontology/year>
          required: true
          type: number
      responses:
        "200":
          description: successful operation. <http://www.openbudgets.eu/ontology/obeu/Dataset>
          schema:
            type: array
            items:
              $ref: "#/definitions/Dataset"
        "400":
          description: Invalid status value.
      security:
        - datasets_auth:
            - write_datasets
            - read_datasets
```

Listing C.1: OpenAPI description using YAML.

```
{
   "swagger":"2.0",
   "info":{
```

```
    "version":"1.0.0",
    "title":"EU Government Budget Data API.",
    "description":"An API to provide access EU government budget datasets",
    "termsOfService":"http://openbudgets.eu/terms/",
    "contact":{
       "name":"OBEU API team",
       "email":"api@openbudgets.eu",
       "url":"http://openbudgets.eu"
    },
    "license":{
       "name":"MIT",
       "url":"http://opensource.org/licenses/MIT"
    }
 },
 "paths":{
    "/FindByMunicipalityYear":{
       "get":{
          "tags":[
             "municipality, year"
          ],
          "summary":"Finds data by year and municipality.",
          "description":"Find published governmental budget or spending datasets based on
          ↪  municipality and year. <http://www.openbudgets.eu/ontology/obeu/Dataset>",
          "operationId":"FindByMunicipalityYear",
          "produces":[
             "application/json"
          ],
          "parameters":[
             {
                "in":"query",
                "name":"municipality",
                "description":"Municipality name of corelated budget/spending data.
                ↪  <http://dbpedia.org/ontology/Municipality>",
                "required":true,
                "type":"string"
             },
             {
                "in":"query",
                "name":"year",
                "description":"Year of budget/spending data.
                ↪  <http://dbpedia.org/ontology/year>",
                "required":true,
                "type":"number"
             }
          ],
          "responses":{
             "200":{
                "description":"successful operation.
                ↪  <http://www.openbudgets.eu/ontology/obeu/Dataset>",
                "schema":{
                   "type":"array",
                   "items":{
                      "$ref":"#/definitions/Dataset"
                   }
                }
             },
             "400":{
                "description":"Invalid status value."
             }
          },
          "security":[
             {
```

```
                "datasets_auth":[
                   "write_datasets",
                   "read_datasets"
                ]
             }
          ]
       }
    }
 }
}
```

Listing C.2: OpenAPI description using JSON.

```
{
  "@context": {
    "dcterms": "http://purl.org/dc/terms/",
    "owl": "http://www.w3.org/2002/07/owl#",
    "rdf": "http://www.w3.org/1999/02/22-rdf-syntax-ns#",
    "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
    "schema": "http://schema.org/",
    "swg": "http://eis.iai.uni-bonn.de/vocab/swagger/swagger#",
    "vcard": "http://www.w3.org/2006/vcard/ns#",
    "xsd": "http://www.w3.org/2001/XMLSchema#",
    "obeu": "http://www.openbudgets.eu/description#"
  },
  "@graph": [
    {
      "swg:swagger": "2.0"
    },
    {
      "swg:_apiUrl":"http://www.openbudgets.eu/api"
    },
    {
      "@id": "obeu:OBEUInfo",
      "@type": [
        "swg:Info",
        "owl:NamedIndividual"
      ],
      "dcterms:description": "An API to provide access to EU governments financial datasets.",
      "dcterms:title": "EU Government Budget Data API.",
      "swg:contact": {
        "@id": "obeu:OBEUcontact"
      },
      "swg:license": {
        "@id": "obeu:OBEUlicense"
      },
      "swg:termsOfService": "http://openbudgets.eu/terms/",
      "swg:version": "1.0.0"
    },
    {
      "@id": "obeu:OBEUcontact",
      "@type": [
        "swg:Contact",
        "owl:NamedIndividual"
      ],
      "foaf:name": "OBEU API Team",
      "vcard:email": "api@openbudgets.eu",
      "foaf:url": "http://openbudgets.eu"
    },
    {
```

```
    "@id": "obeu:OBEUlicense",
    "@type": [
      "swg:License",
      "owl:NamedIndividual"
    ],
    "schema:url": "http://opensource.org/licenses/MIT",
    "swg:name": "MIT"
  },
  {
    "@id": "obeu:OBEUpaths",
    "@type": [
      "swg:Paths",
      "owl:NamedIndividual"
    ],
    "swg:path": {
      "@id":"obeu:FindByMunicipalityYear"
    }
  },
  {
    "@id": "obeu:FindByMunicipalityYear",
    "@type": [
      "swg:PathItem",
      "owl:NamedIndividual"
    ],
    "swg:_pathItemName":"/FindByMunicipalityYear",
    "swg:_operation":{
      "@id":"obeu:Get_FindByMunicipalityYear"
    }
  },
  {
    "@id": "obeu:Get_FindByMunicipalityYear",
    "@type": [
      "swg:Operation",
      "owl:NamedIndividual"],
    "swg:_operationType": "get",
    "swg:operationTags": [
      "municipality",
      "year"
    ],
    "swg:summary": "Finds data by year and municipality.",
    "dcterms:description": "Find published governmental budget or spending datasets based on
    ↪  municipality and year.  Output: <http://www.openbudgets.eu/ontology/obeu/Dataset> ",
    "swg:operationId": "FindByMunicipalityYear",
    "swg:produces": "application/json",
    "swg:parameters": [
      {"@id":"obeu:Param_Municipality"},
      {"@id":"obeu:Param_Year"}
    ],
    "swg:operationResponses": [
      {"@id":"obeu:Get_200_FindByMunicipalityYear"},
      {"@id":"obeu:Get_400_FindByMunicipalityYear"}
    ]
  },
  {
    "@id": "obeu:Param_Municipality",
    "@type": [
      "swg:Parameter",
      "owl:NamedIndividual"
    ],
    "swg:in": "query",
    "dcterms:description": "Municipality name of corelated budget/spending
    ↪  data.<http://dbpedia.org/ontology/Municipality>",
```

```
      "swg:required": true,
      "swg:type": "string"
    },
    {
      "@id": "obeu:Param_Year",
      "@type": [
        "swg:Parameter",
        "owl:NamedIndividual"
      ],
      "swg:in": "query",
      "description": "Year of budget/spending data.  <http://dbpedia.org/ontology/year>",
      "swg:required": true,
      "swg:type": "number"
    },
    {
      "@id": "obeu:Resp_Get_200_FindByMunicipalityYear",
      "@type": [
        "swg:Response",
        "owl:NamedIndividual"
      ],
      "dcterms:description": "successful operation.
      ↪  <http://www.openbudgets.eu/ontology/obeu/Dataset>",
      "swg:schema": {
        "@id":"obeu:Schema_Resp_Get_200_FindByMunicipalityYear"
      }
    },
    {
      "@id": "obeu:Resp_Get_400_FindByMunicipalityYear",
      "@type": [
        "swg:Response",
        "owl:NamedIndividual"
      ],
      "dcterms:description": "Invalid status value."
    },
    {
      "@id": "obeu:Schema_Resp_Get_200_FindByMunicipalityYear",
      "@type": [
        "swg:Schema",
        "owl:NamedIndividual"
      ],
      "swg:type": "array",
      "swg:itemsType": "#/definitions/Dataset"
    }
  ]
}
```

Listing C.3: OpenAPI description using JSON-LD.

APPENDIX D

Platform Life-cycle, Significance, Requirements and Related Quality Factors

Table D.1: Aligning requirements from Open Data Life Cycle [121], open fiscal data platform functionality requirements [122], and matching data quality factors [124]

Open Data Life Cycle [121])				**Required Functionality [122]**	**Data Quality Factors [124]**
Level	**Title**	**Description**	**Significance**	**Functionality**	**Related OFDP Quality Factors**
1	Data Creation	Creating the datasets in public administrations is usually part of daily procedures. Among the requirements within data creation are documentation, provenance, and authoritative.	Very Important/Critical since it is related to the existence of the datasets as well as ensuring a valid source and understandability for the created fiscal datasets.	Proper documentation, Provenance	Authoritative, Data Existence, Documentation
2	Data Selection	Data selection involves the removal of existing private and personal data, as well as identification of conditions for publishing the data. Curating the list of available classifications (i.e.,code lists), checking for missing data, and enlisting available investments alternatives are part of the requirements.	Highly Important, the availability of privacy concerns hinders the analysis of the data and incomplete code lists prevents the datasets from easily described.	Curation for Code lists, Red Flag upon missing data, List of available investment alternatives	Complete Code List

Level	Title	Description	Significance	Functionality	Related OFDP Quality Factors
3	Data Harmonization	Making the datasets conform with open data publication standards is the focus of data harmonization. Several requirements within data harmonization includes: creation of RDF data model that supports budgets, revenues, incomes, transactions, classifications, amount, payer, payee and currency; acquisition of metadata; clarification of data usage license; semantic mapping of CSV; mapping of OpenSpending data model to RDF; association of targeted amount to spending; and the linking of data items. Published datasets should ideally provided as structured data in an open format using open license.	Highly Important, properly modeled and well integrated fiscal datasets allow more straightforward analysis that attracts open data / civic / academic / research enthusiasts.	RDF data structure for budgets, RDF data structure for transactions, Mapping OpenSpending Data Package to RDF, Modeling of code lists in RDF, Data structure for modeling revenues/incomes, Ability to model payer, payee, amount, date, currency; Ability to attach concrete targets to spending, Link ability, Semantic Mapping of CSV, Acquisition of metadata, Clear licensing information	Open Format, Open License, Structured Data
4	Data Publishing	The main data publishing stage consists of different requirements, such as data loading from CSV format or an API; providing kiosk mode on the data web page, as well as fully customization continuous integration, download button and links to Freedom of information act /Access to Documents. Ideally the published datasets should be easily and publicly accessible online through API as well as bulk download, with license, contributors, contact point and English information provided. The datasets should also ideally be free of charge and published in a sustainable manner.	Very Important, there is no open data without data publishing stage The way data are published determines data consumers engaged in further open data life cycle stage.	Loading of CSV, Loading from an API, Kiosk mode, Fully-customizable CI, Download button all the way, Links to FoI/ATD Tools	API Availability, Available Online, Contact Point, Easily Available, English Info Available, Free of Charge, In Bulk, In Digital Form, Mentioned Contributors, Mentioned License, Public, Sustainable Publication
5	Data Interlinking	Data interlinking connects datasets and items within the datasets to other resouces, which makes the datasets have richer contexts. One of the requirements for datasets interlinking is that there is a mapping between related classifications from different datasets. Datasets should also be published in RDF and hence have a dereferenceable URI.	Slightly important to important. In theory interlinked fiscal data would allow fiscal data enrichment which would provide extra context for comparative analysis acoss heterogeneous fiscal datasets, however most datasets are not interlinked due to technical barriers as well as no such comparative analysis 'killer apps'.	Code lists' mappings support	Dereferenceable LD URI, RDF Availability
6	Data Discovery	The existence of open data should be discovered by data consumers. From the requirements perspective, data discovery can be enhanced by the availability of free-text search, the availability of semantic search, processed datasets that can be explored, availability of metadata, feature to perform different levels of query, implementation of a user-friendly user interface.	Highly Important. The ease of discovering fiscal data should help fiscal data enthusiasts to collect all the necessary fiscal data effectively and efficiently.	Free-text search, Semantic search, Relevance ranking, Exploration of processed datasets, Metadata, Different levels of difficulty in queries, User-Friendliness of the UI, Export and share.	Dataset Filtering, Search Mechanism
7	Data Exploration	In the data exploration stage, simple ways to consume the data are performed. The related requirements for this stages are: exporting and sharing high quality and indisputable visualization, previewing the visualization, availability of geographical visualization, explained flow of budgeting process, visual exploration of both RDF and non-RDF data, search result relevance ranking, availability of data exploration samples, availability of visualization suggestions, tracking of user data processing workflow and cache processing data, tracking of budget version, budget comparison with using different dimensions (public administrations, time, and function), filtering (by spending or administration type), availability of top-level aggregation, availability of localized or translated data, querying by administrative regions or institutions, and attach participatory budgeting result.	Highly Important. Data exploration allows the datasets to be explored in a generic way for common use cases. This stage results an easy to understand fiscal data analytics which should be suitable for the dissemination of fiscal data analytics for most people.	Good quality visualizations, Indisputable visualizations, Provide geographical visualizations, Preview, Budget process model, Visual exploration (RDF), Non Semantic exploration,Provide samples, Suggest first, Do not repeat, Version tracking of budgets, Entity comparison, Temporal comparison, Functional comparison, Filter by spend, Filter by administration type, Get top-level aggregates, Localized data, Query by institution, administrative regions	Visualization

Level	Title	Description	Significance	Functionality	Related OFDP Quality Factors
8	Data Exploitation	The next level of data cycle is exploiting the data, which is a more advanced step in consuming the data and allows users to provide analysis, mashup or some other innovations by using, reusing or distributing the data. The requirements involved in the data exploitation stage include building custom visualization, performing exploit analysis, filtering commensurable objects, detecting outliers, extrapolating the data, aggregating the data (by time interval, temporal trend of planned vs actual spent money, normalizing by key metrics, differentiating between real vs nominal value (e.g., inflation adjustments), providing contextual information, breaking down the budget and spending items, and attaching spending to participatory budgeting result.	Important. The advanced level of data exploitation would typically provide unique actionable insights that can be taken for interested parties, however, the trade off between technical barrier and the possible collected insight may not seem to be paid off.	Build custom visualizations, Exploit analysis, Filtering commensurable objects, Outlier detection, Extrapolations on data, Aggregation by time interval, Temporal trend of the difference between planned and actual spending, Normalize by key metrics, Real vs. Nominal, Contextual information, Break Down functionality, Displaying results, Attach Targets to spending	
9	Data Curation	Data curation is important to ensure data sustainability. The requirements within data duration include pointing missing data, indexing both tabular and RDF graph data structure, and gathering budget votes in terms of participatory budgeting. Datasets should ideally published with metadata, updated regularly, and timely. A version tracking for datasets would also be desirable.	Very Important. This stage determines whether the fiscal datasets will always be available, authoritative, and sustainable. This stage will make the fiscal data enthusiasts to be more confidence that their efforts in analyzing some particular data will always be reusable for further fiscal data publication.	Point missing data, Indexing data w.r.t. tabular vs. graph structures, Gathering votes	Metadata, Regular Update, Up to date, Version Tracking

Note:
1. Categorization may not always binary.
2. One requirement/quality factors may belong to several levels. In this case, the requirement is mapped to the level with strongest association sense.

Table D.2: OpenBudgets.eu platform functional requirements and the associated tools and open data life cycle.

OpenBudgets.eu Platform Functional Requirements [122]				Tools, Satisfiability, ODLC Level		
No	**Features**	**Functionality**	**Description**	**Utilized/Integrated Tools**	**Satisfied?**	**ODLC Level**
F001	**Data Model**	RDF data structure for budgets	Budgeting information representation by utilizing Data Cube Vocabulary DCV.	OBEU Ontology	Yes	3
F002	**Data Model**	RDF data structure for transactions	Transaction information representation by using DCV.	OBEU Ontology	Yes	3
F003	**Data Model**	Mapping OpenSpending Data Package to RDF	Providing a mapping from OpenSpending Fiscal Data Model to OpenBudgets Data Model	FDP2RDF Pipeline	Yes	3
F004	**Data Model**	Curation for Code lists	System for managing code lists.	GitHub (Not Integrated) + RDF Browser + Virtuoso (and SPARQL Endpoint)	Yes	2
F005	**Data Model**	Modeling of code lists in RDF	Utilizing/extending available RDF vocabulary to model the code lists.	OBEU Ontology	Yes	3
F006	**Data Model**	Data structure for modeling revenues/incomes	Providing data structure definition for modeling revenue/income data.	OBEU Ontology	Yes	3
F007	**Data Model**	Budget process model	Modeling the flow of budget from the very beginning (planned budget) until the very end (transaction and achieved result).	OBEU Ontology (through it's budget-phase property)	Yes	7
F008	**Data Model**	Code lists' mappings support	Provide localization and external mappings.	UI Alignment	Yes	5

No	Features	Functionality	Description	Utilized/Integrated Tools	Satisfied?	ODLC Level
F009	**Data Model**	Ability to model payer, payee, amount, date, currency	Provide support for modelling transactions and its properties.	OBEU Ontology	Yes	3
F010	**Data Model**	Ability to attach concrete targets to spending	Provide a way for community to associate a specific budget with concrete targets, and the community could evaluate the resulting target in the end of the budgeted project.	OBEU Ontology	Yes	3
F011	**Data Model**	Link ability	Every data items must have a link (URI).	OBEU Ontology. Also depending on the datasets, whether the datasets provide both approved/drafted and executed budget phase.	Yes	3
F012	**Data acquisition, loading, semantic lifting**	Loading of CSV	Provide support for loading CSV data, as well as selecting specific column in the CSV data.	OS Packager, LInkedPipes ETL	Yes	4
F013	**Data acquisition, loading, semantic lifting**	Semantic Mapping of CSV	Provide support for mapping CSV columns to corresponding RDF properties.	LinkedPipes ETL	Yes	3
F014	**Data acquisition, loading, semantic lifting**	Acquisition of metadata	Provide support for capturing metadata (public administrations, year, data uploader) for each loaded data package.	OS Packager, LinkedPipes ETL	Yes	3
F015	**Data acquisition, loading, semantic lifting**	Loading from an API	Provide support for acquiring datasets from API, if the datasets are not available as bulk download.	NA	No	4
F016	**Exploration, search**	Visual exploration (RDF)	Provide support for graph browsing, to find relationship between data items.	RDF Browser	Yes	7
F017	**Exploration, search**	Non Semantic exploration	Provide support for datasets faceted browsing or tabulated view.	Indigo, OS Viewer	Yes	7
F018	**Exploration, search**	Free-text search	Search trough a keyword for: datasets, attributes, field names.	Indigo, Virtuoso (through SPARQL endpoints - only for experts)	Partial	6
F019	**Exploration, search**	Semantic search	Provide SPARQL endpoints for advanced users, provide pre-stored script for common queries, and user friendly interface for users without SPARQL expertise.	Indigo, Virtuoso (through SPARQL endpoints - only for experts)	Yes	6
F020	**Exploration, search**	Exploration of processed datasets	Provide an aggregate API for searching analysis results previously done.	OS Viewer and Babbage/Rudolf API by caching in the backend.	Yes	6
F021	**Exploration, search**	Metadata	Provide a way to search additional metadata within the datasets.	Virtuoso (through SPARQL Endpoint), Indigo - Data Search	Yes	6
F022	**Exploration, search**	Different levels of difficulty in queries	Provide different levels of query difficulties, for example: 1) lowest level e.g., drop down menu; basic questions; off-the-shelf visuals. 2) intermediary level e.g., access / filter data, pivot tables. 3) SPARQL queries.	OS Viewer through visualization browsing, Virtuoso through SPARQL queries.	Yes	6
F023	**Exploration, search**	User-Friendliness of the UI	Provide user friendly interface.	Indigo	Yes	6
F024	**Exploration, search**	Relevance ranking	Provide search relevance ranking.	No	No	6
F025	**Visualization**	Build custom visualizations	Provide visualization customization, by column / relation selection in one or more datasets.	OS Viewer through visualization browsing, KPI	Yes	8
F026	**Visualization**	Exploit analysis	Provide an interface for using features offered in the Analysis Tools, allowing users to flexibly combine data and hide complex queries.	OS Viewer, Indigo, KPI, and Babbage / Rudolf API	Yes	8

No	Features	Functionality	Description	Utilized/Integrated Tools	Satisfied?	ODLC Level
F027	**Visualization**	Provide samples	Give beginners some templates to start playing, without requiring them to read any documentation. Give helpful hints and maybe a tutorial., e.g., ideal sample would ideally be real, successful news stories.	OS Viewer	Yes	7
F028	**Visualization**	Suggest first	Provide suggestions regarding visualize-able datasets, instead of overwhelmingly provide all visualizations.	OS Viewer, Indigo, KPI	Yes	7
F029	**Visualization**	Do not repeat	Provide tracking of users' movement towards user interface and allow the users to repeat the interaction with updated data.	NA	No	7
F030	**Visualization**	Export and share	Let the information shareable in other platform, as well as printing.	OS Viewer, iframe export is supported but not for print.	Partial	7
F031	**Visualization**	Good quality visualizations	Provide good quality visualizations for printing.	OS Viewer provide an intuitive visualization, but it is not for print	Partial	7
F032	**Visualization**	Indisputable visualizations	Provide a way to verify the source of the data that is used for the visualization.	OS Viewer (Uploader name and partial email is provided, but without uploader's email domain)	Partial	7
F033	**Visualization**	Provide geographical visualizations	Provide maps that shows datasets geographical availability - if the datasets have geographical information.	NA	No	7
F034	**Visualization**	Point missing data	Based on the patterns of previously available datasets, provide a way to point if a dataset is missing.	NA	No	9
F035	**Visualization**	Preview	Provide a way to preview the datasets.	NA	No	6
F036	**Analytics**	Filtering commensurable objects	Provide aggregate analytics that can be used for commensurable objects (objects with the comparable size in any terms).	KPI	Partial	8
F037	**Analytics**	Version tracking of budgets	Provide budget evolution analysis for tracking budgets over its preparation phase.	NA/Available specifically for Bonn (proof of concept) but not for other public administrators as it is very data-structure specific.	No	7
F038	**Analytics**	Indexing data w.r.t. tabular vs. graph structures	Optimize indexing for both tabular and graph data structures.	Data Analytics and Mining component	Yes	9
F039	**Analytics**	Outlier detection	Provide a way to find disproportionately used categories based on outlier detection algorithms.	Data Analytics and Mining component	Yes	8
F040	**Analytics**	Extrapolations on data	Outline trends for predicting / recommending future budget allocations.	Time Series	Yes	8
F041	**Analytics**	Aggregation by time interval	Provide aggregation feature for e.g., sum or average of budget / spending amounts over a defined period of time (e.g., quarter, year).	NA, very dependent on datasets	No	8
F042	**Analytics**	Temporal trend of the difference between planned and actual spending	Analyze the differences between planned and actual expenditure during the course of a time, then get insight with regards to the differences.	NA	No	8
F043	**API**	Entity comparison	Provide a way to query datasets from multiple entities.	NA	No	7
F044	**API**	Temporal comparison	Provide comparison over time, e.g., budgeted / spent over time.	NA/Partial, available specifically for Bonn datasets (as a proof of concept) but not for other public administrators as it is very data-structure specific.	No	7

No	Features	Functionality	Description	Utilized/Integrated Tools	Satisfied?	ODLC Level
F045	**API**	Functional comparison	Provide a comparison over functional classification (e.g.: education), also along different entities and time.	Virtuoso, OS Viewer (manual navigation through classification), Babbage/Rudolf API	Partial	7
F046	**API**	Filter by spending amount	Provide a way to filter spending amount by queries on a desired public entity.	Virtuoso, Indigo	Yes	7
F047	**API**	Filter by administration type	Filter by administrative classifications (managing offices).	Virtuoso	Yes	7
F048	**API**	Get top-level aggregates	Provide a way to query for total allocated/spent amount across all (e.g.: countries) in (e.g.: years 2010 - 2018).	OS Viewer (partial, since multi-years aggregation is not supported)	Partial	7
F049	**API**	Normalize by key metrics	Provide normalization by population (also breakdown population by gender, age, etc.)	KPI	Yes	8
F050	**API**	Real vs. Nominal	Provide necessary adjustments, e.g., inflation adjustments.	LinkedPipes ETL Pipeline Fragment	Yes	8
F051	**API**	Localized data	Provide a feature for data localization and translation (e.g.: entities titles, budget lines).	OBEU Ontology, LinkedPipes ETL (also data dependent)	Yes	7
F052	**SAAS Interface**	Kiosk mode	Provide activity report and software management.	Microsite	Yes	4
F053	**SAAS Interface**	Fully-customize-able CI	Provide continuous, customize-able integration for ready-to-deploy working copy.	Microsite	Yes	4
F054	**Journalism Use**	Download button all the way	Provide a way for journalist to download and store the data at every step of the analysis.	OS Packager, Virtuoso	Yes	4
F055	**Journalism Use**	Contextual information	Provide additional contextual information (e.g., the budget-holder responsible for, on and-off budget items, data with regards to population, relation with Eurostat data).	NA	No	8
F056	**Journalism Use**	Proper documentation	Provide information on methodology, data sources, and how the mapping has been done. The information should be done on a dataset level.	OS Packager, LinkedPipes ETL	Yes	1
F057	**Journalism Use**	Provenance	Provide provenance information.	OS Packager, LinkedPipes ETL, OS Viewer	Yes	1
F058	**Journalism Use**	Red Flag	Provide 'red flag' feature to report missing data.	NA	No	2
F059	**Transparency Use**	Links to FoI/ATD Tools	Provide link to ask the AsktheEU.org/fragdenstaat/Freedom of information act/Access to Documents	NA	No	4
F060	**Transparency Use**	Break Down functionality	provide a way to break down the budget items into major categories, institutions, etc.	OS Viewer	Yes	8
F061	**Transparency Use**	Clear licensing information	Provide information licensing information to encourage reuse of visualizations or data.	LinkedPipes ETL	Yes	3
F062	**Transparency Use**	Query by institution, administrative regions	Provide a way for filtering per dataset as well as aggregates of all data that refers to the institution.	OS Viewer, depending on the availability of administrative classification on the datasets	Yes	7
F063	**Participation Use**	List of available investment alternatives	Provide a way for municipalities to create and update a list of potential investment options for users to pick.	Participatory Budgeting Tool	Yes	2
F064	**Participation Use**	Gathering votes	Store votes from citizens to the proposed investment alternatives and potentially provide a way to comments/feedback on these items.	Participatory Budgeting Tool	Yes	9

No	Features	Functionality	Description	Utilized/Integrated Tools	Satisfied?	ODLC Level
F065	**Participation Use**	Displaying results	Count votes according to the agreed process and then display the vote to other users and municipality.	Participatory Budgeting Tool	Yes	7
F066	**Participation Use**	Attach Targets to spending	Provide a way to show the spending result, e.g., hospitals built or disease index reduced. Provide a way to attach spending to concrete results so that the spending can be later scrutinized.	NA	No	8

APPENDIX E

List of Publications

- *Journal Articles and Book Chapters*:
  1. **Fathoni A. Musyaffa**, Christiane Engels, Maria-Esther Vidal, Fabrizio Orlandi, and Sören Auer. 2018. *Experience: Open Fiscal Datasets, Common Issues, and Recommendations.* J. Data and Information Quality 9, 4, Article 19 (April 2018), 10 pages. DOI:https://doi.org/10.1145/3190576
  2. **Fathoni A. Musyaffa**, Maria-Esther Vidal, Fabrizio Orlandi, Jens Lehmann, Hajira Jabeen, *IOTA: Interlinking of heterogeneous multilingual open fiscal DaTA*, Expert Systems with Applications, Volume 147, 2020, 113135, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2019.113135
- *Conference Papers*:
  3. **Fathoni A. Musyaffa**, Fabrizio Orlandi, Hajira Jabeen, and Maria Esther Vidal. 2018. *Classifying Data Heterogeneity within Budget and Spending Open Data.* In Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance (ICEGOV '18). Association for Computing Machinery, New York, NY, USA, 81–90. DOI:https://doi.org/10.1145/3209415.3209482
  4. **Fathoni A. Musyaffa**, Jens Lehmann, and Hajira Jabeen. 2020. *Cross-Administration Comparative Analysis of Open Fiscal Data.* In Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance (ICEGOV '20). Association for Computing Machinery, New York, NY, USA
  5. **Fathoni A Musyaffa**, Lavdim Halilaj, Yakun Li, Fabrizio Orlandi, Hajira Jabeen, Sören Auer. *OpenBudgets.eu: A Platform for Semantically Representing and Analyzing Open Fiscal Data.* In 18th International Conference on Web Engineering (ICWE) 2018 Proceedings, Springer
  6. **Fathoni A Musyaffa**, Lavdim Halilaj, Ronald Siebes, Fabrizio Orlandi, Sören Auer. *Minimally Invasive Semantification of Lightweight Service Descriptions.* In IEEE International Conference on Web Services (ICWS) 2016 Proceedings, 672–677, IEEE

- *Workshops and Demos*:
  7. **Fathoni A. Musyaffa**, Fabrizio Orlandi, Tiansi Dong, Lavdim Halilaj. *Open-Budgets.eu: A Distributed Open-Platform for Managing Heterogeneous Budget Data*. In 13th International Conference on Semantic Systems (Semantics) - Posters and Demo Track, 2017
  8. Jindřich Mynarz, Jakub Klímek, Marek Dudáš, Petr Škoda, Christiane Engels, **Fathoni A. Musyaffa**, Vojtěch Svátek: *Reusable transformations of Data Cube Vocabulary datasets from the fiscal domain*, Proceedings of the 4th International Workshop on Semantic Statistics (SemStats 2016), Kobe, Japan, October 2016. CEUR Workshop Proceedings, Volume 1654, 2016.

List of Figures

List of Tables