Cross-Administration Comparative Analysis of Open Fiscal Data

Fathoni A. Musyaffa
University of Bonn
Endenicher Allee 19a, 53115 Bonn, Germany
musyaffa@cs.uni-bonn.de

Jens Lehmann
University of Bonn & Fraunhofer IAIS
Zwickauer Str. 46, 01069 Dresden, Germany
jens.lehmann@iais.fraunhofer.de

Hajira Jabeen
University of Bonn
Endenicher Allee 19a, 53115 Bonn, Germany
jabeen@iai.uni-bonn.de

ABSTRACT

To improve governance accountability, public administrations are increasingly publishing their open data, which includes budget and spending data. Analyzing these datasets requires both domain and technical expertise. In civil communities, these technical and domain expertise are often not available. Hence, 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 paper, 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.

CCS CONCEPTS

• **Applied computing** → **Computers in other domains** → Computing in government → *E-government*

KEYWORDS

Linked open data, semantic web, comparative analysis, budget and spending, knowledge graph.

ACM Reference format:

Do not remove modify or remove / editorial placeholder for author name. 2019. Do not modify or remove / editorial placeholder for paper title. In *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance (ICEGOV2020), Athens, Greece, March 11-13, 2020*, 00 pages. https://doi.org/00.0000/0000000.0000000

1. INTRODUCTION

Publishing open data is important for public administrations due to several reasons, such as for supporting law enforcement [1], improving government accountability and effectiveness [2], improving democratic control and political participation [1, 2], as well as improving transparency and compliance [2, 3]. There are several domains in open data, and one that is particularly important for transparency is open budget and spending data. Publishing budget and spending data open the possibilities for interesting comparative analysis across datasets and regions [1, 2]. In this paper, we refer to fiscal data as a short term for specifically referring to public budget and spending data as a subpart in the open data domain.

Public administrations have published open fiscal data through their open data portal. Globally, the number of open fiscal data published is increasing. One of the biggest portals in open fiscal data is OpenSpending.org, an open platform backed by the Open Knowledge Foundation (OKF), which stores fiscal data records from public administration, uploaded and maintained by civil communities. As of September 2019, OpenSpending.org stores more than 136 million fiscal records from 81 countries. Additionally, a survey by OKF shows that budget data is one of the most published data with 98 out of 122 surveyed countries publish their budget datasets [4] publicly.

Publishing fiscal datasets are one of the first key steps to be transparent with regards 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 [4], [5], [6], and even beter, if specific quality factors for open fiscal data are considered [7].
2) Providing representations that support semantics for open fiscal data, such as the OpenBudgets.eu (OBEU) ontology

ICEGOV2020, April 1-3, 2020, Athens, Greece

ACM ISBN 000-0-0000-0000-0/00/00…$15.00
https://doi.org/00.0000/0000000.0000000

[8] for Resource Description Framework (RDF) datasets (more on RDF is explained in Section 2).

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 shares a similar topic or classification (more regarding classification is explained in Section 2). 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 rather challenging and demanding. 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.

In this paper, 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 a real-world heterogeneous fiscal datasets.

The remaining paper is organized as follows: preliminaries are provided in Section 2, followed by motivation and in Section 3. A list of related work is elaborated in Section 4, followed by an explanation of the approach in Section 5. Section 6 provides the detail on our experiment and continued with discussion in Section 7. Finally, this paper is concluded in section 8.

2. PRELIMINARIES

Budget and spending data are datasets that are published by public administrations and show how public administrations obtain/allocate their funding. 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, organized as a specific type of *classifications* [9]. There are several types of classifications, in which *functional classification* and *administrative classification* are among the most published classifications along with open fiscal data. Functional classification defines the usage of the money (e.g., `public transportation`, `elementary education`). Administrative classification clarifies the responsible public administration department (e.g., `The Department of Public Infrastructure` and `The Ministry of Education`). In practice, there are a lot more classification types and these types have their own characteristics as detailed in [9]. 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. Some classifications are published by interstate organizations so that it could be reused by different public administrations for publishing relevant datasets, such as Common Procurement Vocabulary (CPV) [10] by the European Union and Classifications of the Function of Government (COFOG) [11] by the United Nations. The use of standardized vocabulary increases open fiscal data reusability and enables the comparative analysis of fiscal datasets. Unfortunately, according to the survey that is done on more than 70 fiscal datasets [7], the number of datasets that uses standardized classifications for open fiscal data is very limited. As mentioned in Section 1, since the use of standardized classifications is very limited, a link sets that map the relation of similar classifications from different public administrations is therefore necessary to enable comparative analysis.

Resource Description Framework (RDF) is a specification by the WWW Consortium to represent and exchange data over the World Wide Web [12]. Data items in RDF are represented as a URI if it represents specific things or objects, or the data items can also be represented as a literal especially for representing values (e.g., amount of spending). The relationship between data items is represented using a Subject-Predicate-Object (SPO) pattern coined as a *triple* in RDF specification, for example, the triple:

```
@prefix obeu-dimension:
<http://data.openbudgets.eu/ontology/dsd/dimension/> .
http://data.openbudgets.eu/resource/dataset/budget-
thessaloniki-expenditure-2017 obeu-dimension:organization
<http://dbpedia.org/resource/Thessaloniki>
```

contains
`http://data.openbudgets.eu/resource/dataset/budget-thessaloniki-expenditure-2017` as a Subject, `obeu-dimension:organization` as a Predicate, and `http://dbpedia.org/resource/Thessaloniki` and as an Object. The triple describes that the dataset of `budget-thessaloniki-expenditure-2017` has an organization (in other words, 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 a URI in the triple. The details and further information of Thessaloniki are published publicly and extra information on it (e.g., area size and population size and much other information) can be traced by following the link to http://dbpedia.org/resource/Thessaloniki that enrich the context of the predicate. The information can also be queried by a specific query language for the RDF datasets, SPARQL (short for *SPARQL Protocol and RDF Query Language*) [13]. By representing data in RDF, the relationship between each granular item in the datasets

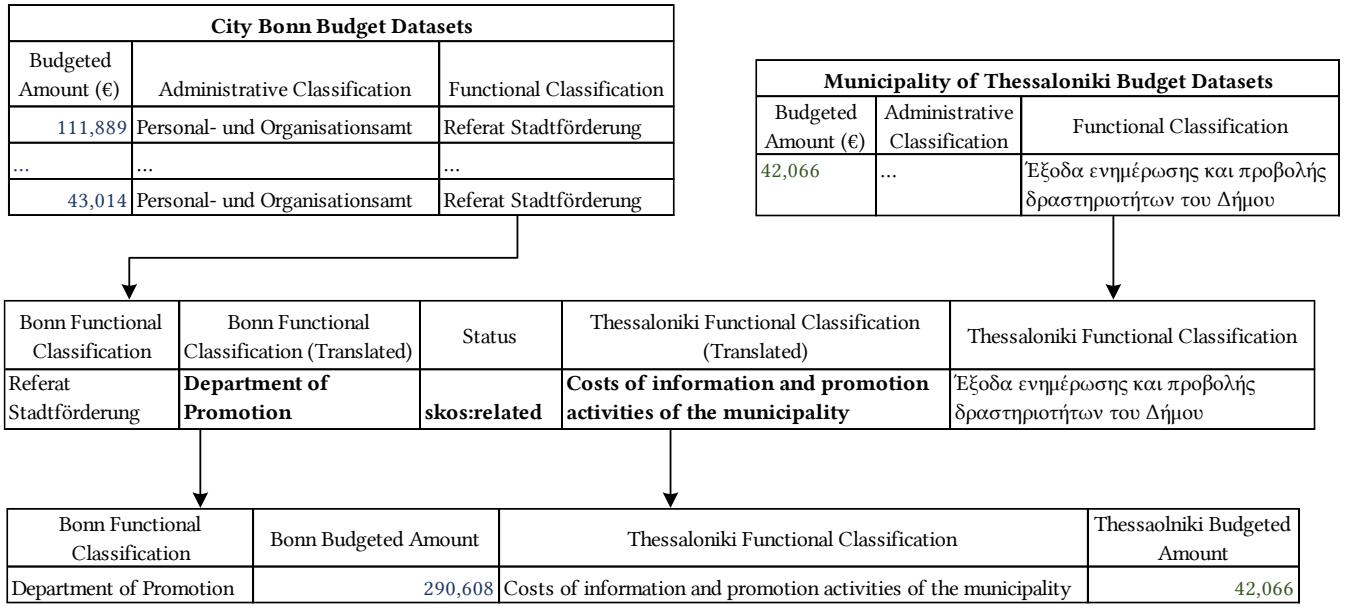

Figure 1. Comparative analysis of open budget data that are represented in different languages.

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 [3]. To publish the data as RDF, several design issues need to be considered, such as (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 [14].

Datasets published in RDF acts as a building block for Linked Open Data and enables datasets from different sources to act as a global database [15]. 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 [15]. By publishing datasets in RDF, data from different sources can be combined together to enrich the context of information being analyzed. A guide on publishing Linked Data is summarized by Bauer and Kaltenböck [15].

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 [16] provides huge information extracted from Wikipedia. The English version of DBpedia (version 2016-04) contains 1.3 billion triples. A sister project of Wikimedia Foundation, Wikidata [17], 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 high yet compared to DBpedia.

3. MOTIVATION

The motivation of this pipeline is to compare budgets and spending from two different public administrations with similar properties. For example, DBpedia states that the city of Bonn has a similar population size compared to the municipality of Thessaloniki. From this information, comparing the budget allocation for both cities is interesting, particularly when the labels of that budget item have a similar meaning. This is illustrated in Figure 1. Here we intend to compare the budget for conceptually related items: "`Referat Stadt förderung`" in German and "`Έξοδα ενημέρωσης και προβολής δραστηριοτήτων του Δήμου`" in Greek which according to Google Translate, both are related with *promotions*. Both concepts belong to functional classification. We are interested to see the budget allocation of two public administrations having similar properties. 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.

4. RELATED WORK

There are several works related to open fiscal data, involving open fiscal data publishing standards; heterogeneity analysis; data modeling, transformation, visualization, and analysis; as well as datasets concept interlinking.

To ensure that published open data have a good quality, easy to understand, and reusable, several factors have been discussed previously. These efforts include the Five-Star data rating, coined by Tim Berners Lee [14], which suggested that data should be: (1) Available on the web with an open license, (2) Published as a structured data, (3) Published in a non-proprietary format, (4) Using URI to denote items in the data, and (5) Linked with other data. Open Data Barometer (ODB) publishes documentation on open data publication guidelines, which enumerates nine quality factors for open data [5]. Global Open Data Index (GODI) by OKF suggests ten quality factors for publishing open data [18]. Since both quality factors suggested by GODI and ODB provide a list of

quality factors for generic open data. Open Fiscal Data Publishing (OFDP) framework [7] suggests 29 quality factors as a follow up for fiscal-data-specific quality assessment, in which 12 of them are based on the existing ODB and GODI quality factors.

Since fiscal datasets are published by the public administrations independently, the published datasets are very heterogeneous in nature. An analysis of fiscal data heterogeneity was conducted by Musyaffa et al. [6], suggesting that representing the datasets in a similar structure and format is one of the keys to enable comparative analysis of open fiscal data. At the moment, there are two formats available to represent open fiscal data into a similar structure Fiscal Data Package (FDP) [19] and OBEU Ontology [8].

FDP is a specification that provides additional JSON metadata for CSV-formatted open fiscal datasets. A tool to facilitate the creation of FDP format from CSV datasets is integrated into the OpenSpending.org platform. The FDP-formatted datasets can be created in the platform by annotating the uploaded, compatible CSV file with metadata upon upload into the platform.

The OBEU ontology is a way to represent fiscal datasets into RDF triples. This ontology is designed to specifically represent RDF data based on Data Cube Vocabulary (DCV) Ontology [20], which is a standard to represent multidimensional data in RDF. This fits the characteristics of open fiscal data, which are commonly published in a multidimensional manner (i.e., multiple classifications are available for the published data). The OBEU ontology requires metadata accompanying the published fiscal datasets. The metadata is provided using DCAT-AP specifications[2] (standardized metadata specified by the European Union). As OBEU ontology is a derivation of DCV, it follows the DCV specification in which concepts represented in the OBEU ontology can be grouped into component parts: *dimensions*, *measure*, and *attribute*. Following the DCV specification, every dataset using the OBEU ontology has a *Data Structure Definition* (DSD) that describes available measures, dimensions, attributes available within the dataset. The *measure* shows numerical values on the observed record (e.g., the amount of money spent on that record). *Dimensions* provide additional information about the specified record, and a combination of different dimensions makes a row unique (i.e., index the record). *Attributes* list additional information for the object that does not index the record (e.g., currency units) [21]. Several *observations* can be grouped by certain dimensions to have a single measure value, and this grouping is defined as a *slice*.

A platform for fiscal data storage, ingestion, metadata annotation, and visualization is available as OpenSpending.org platform, developed and maintained by the OKF. However, OpenSpending.org does not facilitate semantics (RDF). The work of [22] proposes an abstracted architecture for the semantic platform of open data. Since it is an abstract architecture, a materialized platform to perform fiscal data analytics, storage, mining, and visualization is done by [23] which transforms fiscal datasets into OBEU ontology using LinkedPipes ETL (Extract, Transform and Load) tool [24]. The resulting transformation is stored in RDF formats within a triple database and can be queried for visualization. Despite supporting semantics, OpenBudgets.eu platform [23] does not support comparative analysis, because there was no linking method to map concepts of labels with similar labels across fiscal datasets.

Both FDP format and OBEU ontology provide a unified way to unify open fiscal datasets into a unified format. However, publishing datasets that are ready to be transformed into FDP is not a common practice, since datasets published are often not in good quality, as stated in a survey [7]. Publishing datasets into RDF using OBEU ontology is also not a feasible option for public administrations, due to the steep learning curve on RDF semantic technology stack. The learning curve requires public administrations to invest technical resources to publish the datasets. Putting more resources for public administrations may not be convincing enough to make the administrations keen on publishing fiscal datasets both in good quality and in a semantic format. This work serves to persuade public administrations to publish such datasets in both manners, showing that a proof of concept laid out in this paper could be an additional use case which started by publishing high-quality open fiscal data. In this paper, we are focusing on enabling comparative analysis across fiscal datasets published by different public administrations and using information available in the open knowledge bases as a comparison point for those datasets.

5. PIPELINE

Our pipeline is illustrated in Figure 3. 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. After the links have been found, we evaluate the links. 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.[3] Stored data are then queried for comparative analysis. A more detailed approach is provided in the following sub-sections.

5.1. Datasets, Analysis, and Transformation

There are two datasets that we use for the experiment: 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 is compatible with the OBEU ontology. LinkedPipes ETL tool [24] is used to perform the

[2] https://joinup.ec.europa.eu/release/dcat-ap/11

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

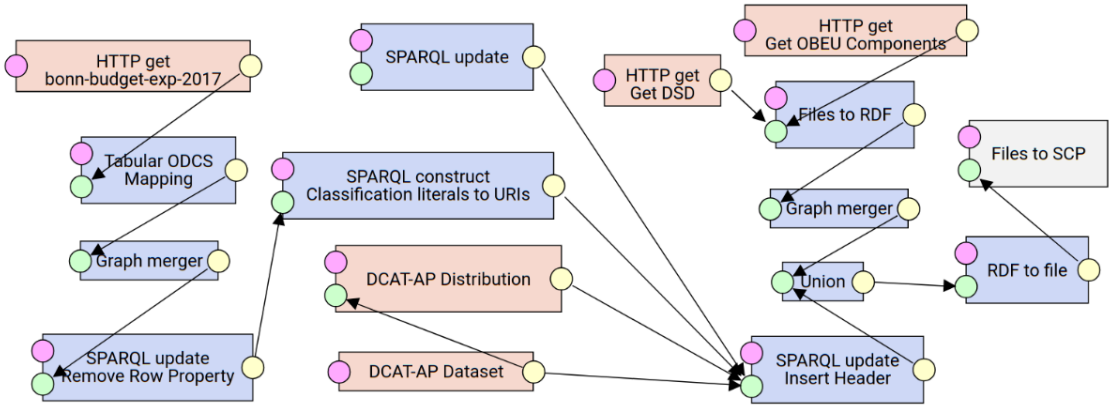

Figure 2. A transformation ETL pipeline from Bonn expenditure dataset 2017 using LinkedPipes ETL. The raw data is formatted in CSV, in which their column is mapped to OBEU ontology properties. Later, the mapped properties are transformed and enriched by using SPARQL statements to fit the remaining OBEU ontology requirements and constraints.

transformation, which allows loading the datasets from tabular formats, adding metadata over the datasets that 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 2. Each box in the Figure 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. These transformation pipelines can be found in a GitHub repository[4] and can be inspected and executed online using the LinkedPipes Demo website.[5] The Thessaloniki expenditure datasets[6] 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.[7]

5.2. Concept Mapping

Concept mapping among the two datasets was done utilizing Apache Spark [25] and *py_stringmatching*,[8] a string-matching library. Initially, 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. For the initial experiment, we use the European Union's Common Procurement Vocabulary (CPV) classification [10] 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 (in this case, we use German, French and Spanish labels of CPV datasets labels) into English, and label the translation based on RFC 6497 – BCP 47 Extension T [26]. 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 measures have the best-performance, and (3) how robust these similarity measures against changes in similarity thresholds. From our experiment [27], we know that the TF-IDF similarity measure provides the best F-Measure performance. We reuse the conclusion from this mapping experiment for this paper, 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.

5.3. Data Storage

The result 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 have used Apache Jena Fuseki[9] as the triple store.

5.4. 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.

[4] https://git.io/JejR1
[5] https://demo.etl.linkedpipes.com/#/pipelines
[6] https://gaiacrmkea.c-gaia.gr/city_thessaloniki/index.php
[7] https://git.io/JejRM
[8] https://git.io/JejRy
[9] https://jena.apache.org/documentation/fuseki2/

Figure 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.

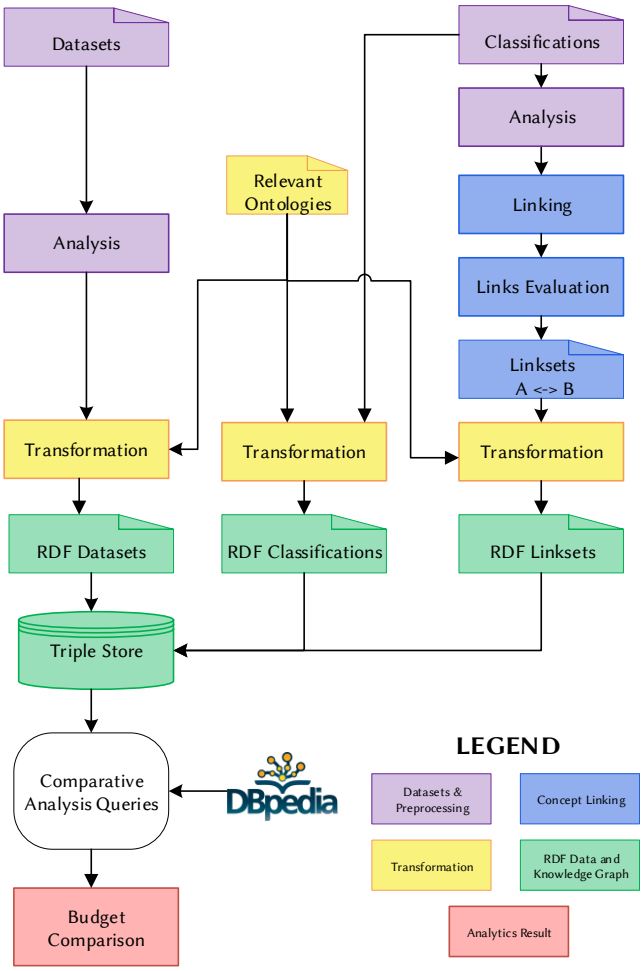

Figure 3: Flow of data and operations to analyze, map, transform, store and query open fiscal datasets.

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.

6. 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, datasets from the municipality of Thessaloniki comprises of administrative classification as well as 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.

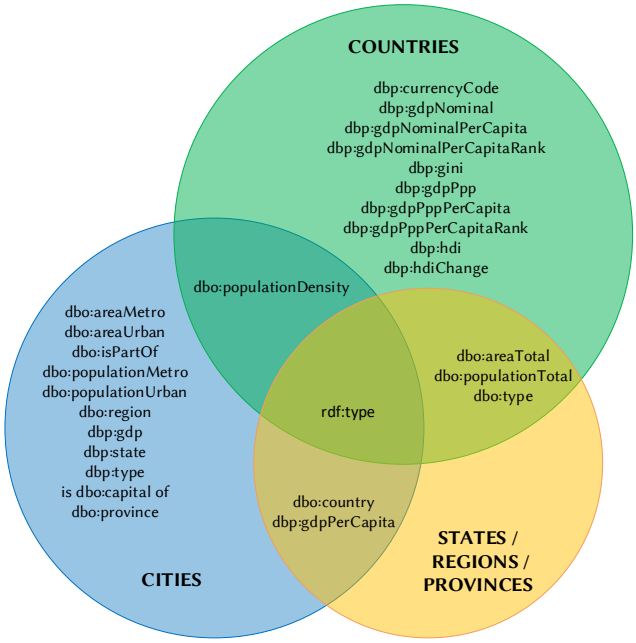

Figure 4. Relevant DBpedia properties that can enrich open fiscal datasets for further comparative analysis.

These data are also different in terms of t*ransformation 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 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

Table 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

Bonn and Thessaloniki datasets have a 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 1. In Table 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 less quality compared to its Google Translate web version (see the English-translated label from Table 1). Bonn functional classification consists of 183 concepts. The functional classification of Bonn is also provided in a hierarchical manner as well (see Table 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 go for *related* links instead.

Since there are multiple observations or slice with the same functional classification spanned over different values, the aggregation operation needs to be performed. For example, a functional classification concept *transport* could be distributed among different administrative offices. In this case, all budget or spending items are summed from different administration offices. This enables one to one comparison of related labels from different municipalities.

The transformation is done using the latest LinkedPipes version,[10] with Apache Jena Fuseki v 3.12.0 as the triple store. Each transformation pipelines are available on GitHub (Bonn[11] and Thessaloniki[12]). 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 beyond the scope of this paper and is discussed in [27].

7. RESULT AND DISCUSSION

Querying available datasets that have similar contextual properties (e.g., as seen in Figure 4) can be done using DBpedia's SPARQL service, as illustrated in Listing 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 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.

Table 3. The resulting query of available datasets that fulfil certain population numbers in DBpedia.

dataset	organization	PopulationTotal
http://data.openbudgets.eu/resource/dataset/bonn-budget-exp-2017	http://dbpedia.org/resource/Bonn	311287
http://data.openbudgets.eu/resource/dataset/bonn-budget-exp-2018	http://dbpedia.org/resource/Bonn	311287
http://data.openbudgets.eu/resource/dataset/bonn-budget-exp-2019	http://dbpedia.org/resource/Bonn	311287
http://data.openbudgets.eu/resource/dataset/budget-thessaloniki-expenditure-2017	http://dbpedia.org/resource/Thessaloniki	385406
http://data.openbudgets.eu/resource/dataset/budget-thessaloniki-expenditure-2018	http://dbpedia.org/resource/Thessaloniki	385406
http://data.openbudgets.eu/resource/dataset/budget-thessaloniki-expenditure-2019	http://dbpedia.org/resource/Thessaloniki	385406

[10] https://git.io/JejRS
[11] https://git.io/JejR1
[12] https://git.io/JejR7

```
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  ?populationT
otal } }
```

Listing 1. Querying available datasets based on specific values (e.g., population size) available in DBpedia.

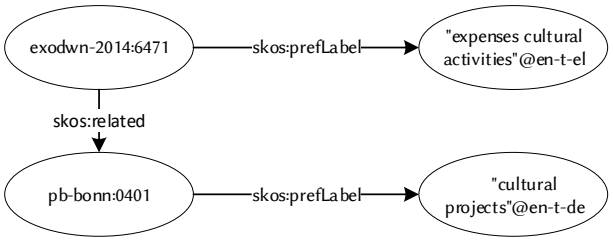

Figure 5. An illustration of a relation between concepts from the city of Bonn and the municipality of Thessaloniki.

```
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  thess-be2015-dimension: <http://data.openbudgets.eu/ontology/dsd/budget-thessaloniki-expenditure-2015/dimension/>
PREFIX  xsd:  <http://www.w3.org/2001/XMLSchema#>
PREFIX  skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX  obeu-dimension: <http://data.openbudgets.eu/ontology/dsd/dimension/>
PREFIX obeu: <http://data.openbudgets.eu/ontology/>

SELECT  ?bnFC ?bnLabel ?thSliceFC ?thLabel (xsd:decimal(?bnAmountTotal) AS ?bnAmountTotalDec) (SUM(?thAmount) AS ?thAmountTotalDec)
WHERE
  { ?thDataset a qb:DataSet;
               obeu-dimension:fiscalYear <http://reference.data.gov.uk/id/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 <http://reference.data.gov.uk/id/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 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.

Table 4. An example of comparative analysis query result.

Bonn Concepts URI	Concept Labels - English Translation	Thessaloniki Concepts URI	Concept Labels - English Translation	Amount Approved: Bonn	Amount Appoved: Thessaloniki
<http://data.openbudgets.eu/resource/codelist/produktuebersicht_bonn/0802>	"sports promotion"@en-t-de	<http://data.openbudgets.eu/resource/codelist/kae-ota-exodwn-2014/6472>	"expenses sports"@en-t-el	"2,198,574.75"^^xsd:decimal	"15,054.9"^^xsd:decimal
<http://data.openbudgets.eu/resource/codelist/produktuebersicht_bonn/0119>	"department of promotion"@en-t-de	<http://data.openbudgets.eu/resource/codelist/kae-ota-exodwn-2014/6431>	"costs of information and promotion activities of the municipality"@en-t-el	"290,608.375"^^xsd:decimal	"42,065.87"^^xsd:decimal
<http://data.openbudgets.eu/resource/codelist/produktuebersicht_bonn/0124>	"administrative organization and it applications"@en-t-de	<http://data.openbudgets.eu/resource/codelist/kae-ota-exodwn-2014/6266>	"maintenance of software applications"@en-t-el	"5,007,714.5"^^xsd:decimal	"63,819.04"^^xsd:decimal
<http://data.openbudgets.eu/resource/codelist/produktuebersicht_bonn/0401>	"cultural projects"@en-t-de	<http://data.openbudgets.eu/resource/codelist/kae-ota-exodwn-2014/6471>	"expenses cultural activities"@en-t-el	"842,904.75"^^xsd:decimal	"392,930.09"^^xsd:decimal

7.1. Result

The mapping experiment results in 87 related links. The links are provided with SKOS ontology[13], specifically using `skos:related` property. Figure 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. The result is that there are 219.220 triples that we have on our experiment.

Listing 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 the Listing 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.

The result of the query is sampled in Table 4 with the following columns: Bonn functional concept URI, translated concept labels from Bonn datasets, Thessaloniki functional 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 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 costs 392,930 €. This is visually represented in Figure 6. The comparative analysis experiment results in 47 related to 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. For those interested in seeing the mapping and query result, the whole resulting experiment is provided in our GitHub repository[14] .

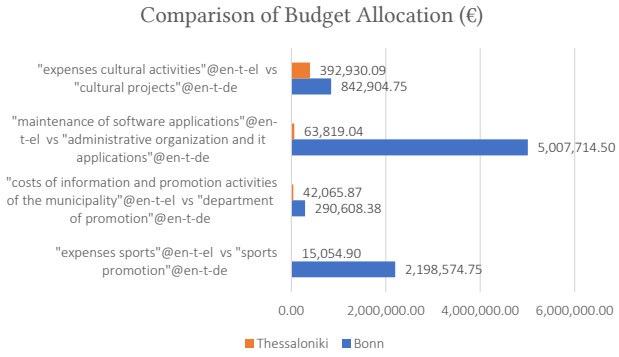

Figure 6. A visualized comparison of related and aggregated budgets from both public administrations.

7.2. Lesson Learned

This paper presents efforts that have enabled comparative analysis of open fiscal data. This is the first work providing proof of concept that there a lot of potential in analyzing open fiscal data by exploiting the ever-increasing linked open data knowledge

[13] https://www.w3.org/2004/02/skos/

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

base such as DBpedia and hence more efforts are needed. 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 have learned:

- Different public administrations have different legislation, business process, and data flow. Therefore, each dataset is most probably complex. We suggest that a careful simplification process should be performed if such datasets are initially complicated (e.g., contain positive/negative values, composite classification items). The datasets should be documented. The details of what each column in the datasets contains and the available classification type 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 [4]–[7]) 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 in an easier data integration process.
- With the rise of AI, the need for a structured knowledge base in the form of linked data is getting more visible and therefore the size of the information available in initiatives such as DBpedia and Wikidata is expanding. Publishing Linked Open Data enables data consumers to get more context from these structured knowledge bases.

8. CONCLUSION AND FUTURE WORK

In this paper, we demonstrate a proof of concept that enables comparative analysis of open budget and spending data. This involves the usage of a specific ontology to enable a unified representation of open fiscal data, using information available on public knowledge bases to enrich the context of the datasets and create relation links between similar concepts across datasets. The analysis can be severely hindered by data quality and missing data. AI methods that could improve the data quality or data linking, could assist in the efficient cross-comparison analysis of heterogeneous data. These methods are suggested to be developed further as future works.

ACKNOWLEDGMENTS

The first author is grateful for the research funding provided by the German Academic Exchange Service/*Deutscher Akademischer Austauschdienst* (DAAD). This work was partly supported by the EU Horizon2020 project LAMBDA (GA No. 809965).

REFERENCES

[1] N. Huijboom and T. Van den Broek, "Open data: an international comparison of strategies," *Eur. J. EPractice*, vol. 12, no. 1, pp. 4–16, 2011.
[2] 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," in *Proceedings of the 9th ICEGOV*, 2016, pp. 276–286.
[3] N. Shadbolt *et al.*, "Linked open government data: Lessons from data. gov. uk," *IEEE Intell. Syst.*, vol. 27, no. 3, pp. 16–24, 2012.
[4] Open Knowledge International, "Global Open Data Index: Place overview." [Online]. Available: https://index.okfn.org/place/. [Accessed: 29-Aug-2017].
[5] T. Davies, "Open Data Barometer, 2013 Global Report." World Wide Web Foundation and Open Data Institute, 2013.
[6] Sunlight Foundation, "Open Data Policy Guidelines," *Sunlight Foundation*. [Online]. Available: https://sunlightfoundation.com/opendataguidelines/. [Accessed: 11-Oct-2019].
[7] F. A. Musyaffa, C. Engels, M.-E. Vidal, F. Orlandi, and S. Auer, "Experience: Open Fiscal Datasets, Common Issues, and Recommendations.," *J Data Inf. Qual.*, vol. 9, no. 4, pp. 19:1-19:10, 2018.
[8] M. DUDÁŠ *et al.*, "The OpenBudgets Data Model and The Surrounding Landscape." .
[9] F. A. Musyaffa, F. Orlandi, H. Jabeen, and M.-E. Vidal, "Classifying Data Heterogeneity within Budget and Spending Open Data.," in *ICEGOV*, 2018, pp. 81–90.
[10] European Commission, "Information System for European Public Procurement: Common Procurement Vocabulary," 2008. [Online]. Available: https://simap.ted.europa.eu/cpv. [Accessed: 27-May-2019].
[11] United Nations Statistics Division (UNSD), "Classification of the Functions of Government (COFOG)," 1999. [Online]. Available: https://unstats.un.org/unsd/iiss/Classification-of-the-Functions-of-Government-COFOG.ashx. [Accessed: 27-May-2019].
[12] F. Manola and E. Miller, "RDF Primer." [Online]. Available: https://www.w3.org/TR/rdf-primer/. [Accessed: 11-Oct-2019].
[13] E. Prud'hommeaux and A. Seaborne, "SPARQL Query Language for RDF." [Online]. Available: https://www.w3.org/TR/rdf-sparql-query/. [Accessed: 11-Oct-2019].
[14] T. Berners-Lee, "Linked Data - Design Issues." [Online]. Available: https://www.w3.org/DesignIssues/LinkedData.html. [Accessed: 30-Sep-2019].
[15] F. Bauer and M. Kaltenböck, *Linked open data: the essentials: a quick start guide for decision makers*. Wien: ed. mono/monochrom, 2012.
[16] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," *Semantic Web*, pp. 722–735, 2007.
[17] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," *Commun. ACM*, vol. 57, no. 10, pp. 78–85, 2014.
[18] OKF, "Global Open Data Index - Methodology," *Open Knowledge Foundation*, Aug-2014. [Online]. Available: http://index.okfn.org/methodology/.
[19] P. Walsh, R. Pollock, T. Björgvinsson, S. Bennett, A. Kariv, and D. Fowler, "Fiscal Data Package." [Online]. Available: https://frictionlessdata.io/specs/fiscal-data-package/. [Accessed: 27-Sep-2019].
[20] R. Cyganiak, D. Reynolds, and J. Tennison, "The RDF Data Cube Vocabulary," *W3C Recommendation*. [Online]. Available: https://www.w3.org/TR/vocab-data-cube/. [Accessed: 07-Oct-2019].
[21] M. Dudaš *et al.*, "Deliverable 1.4: (OpenBudgets.eu Data Model) User documentation," OpenBudgets.eu Consortium, Prague, Czech Republic, 2015.
[22] A. L. Machado and J. M. Parente de Oliveira, "DIGO: An Open Data Architecture for e-Government," 2011, pp. 448–456, doi: 10.1109/EDOCW.2011.34.
[23] F. A. Musyaffa *et al.*, "OpenBudgets.eu: A Platform for Semantically Representing and Analyzing Open Fiscal Data," in *Web Engineering - 18th International Conference, ICWE 2018, Cáceres, Spain, June 5-8, 2018, Proceedings*, 2018, pp. 433–447, doi: 10.1007/978-3-319-91662-0_35.
[24] J. Klímek, P. Škoda, and M. Nečaskỳ, "LinkedPipes ETL: Evolved linked data preparation," in *European Semantic Web Conference*, 2016, pp. 95–100.
[25] M. Zaharia *et al.*, "Apache Spark: A Unified Engine for Big Data Processing," *Commun ACM*, vol. 59, no. 11, pp. 56–65, Oct. 2016, doi: 10.1145/2934664.
[26] M. Davis, A. Phillips, Y. Umaoka, and C. Falk, "BCP 47 Extension T - Transformed Content.," *RFC*, vol. 6497, pp. 1–15, Feb. 2012.
[27] F. A. Musyaffa, M.-E. Vidal, F. Orlandi, J. Lehmann, and H. Jabeen, "IOTA: Interlinking of Heterogeneous Multilingual Open Fiscal DaTA," *Expert Syst. Appl.*, p. 113135, 2019, doi: https://doi.org/10.1016/j.eswa.2019.113135.