Classifying Data Heterogeneity within Budget and Spending Open Data

Fathoni A. Musyaffa
University of Bonn
Römerstr. 164 53117 Bonn
Germany
musyaffa@cs.uni-bonn.de

Fabrizio Orlandi
Fraunhofer IAIS & University of Bonn
Römerstr. 164 53117 Bonn
Germany
orlandi@cs.uni-bonn.de

Hajira Jabeen
University of Bonn
Römerstr. 164 53117 Bonn
Germany
jabeen@iai.uni-bonn.de

Maria Esther Vidal
German National Library of Science
and Technology
Welfengarten 1B 30167 Hannover
Germany
maria.vidal@tib.eu

ABSTRACT

Open data has gained momentum for the past few years, but not much consumption was done 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 budget and spending datasets released by different public administrations from various levels of administrations and locations. We select five datasets, then present and illustrate several types of budget and spending heterogeneities. We compare these heterogeneities 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 hints for both datasets publishers and technical/research communities that deal with open data in budget and spending domain.

CCS CONCEPTS

• **Applied Computing → Computer in other domains → Computing in government** → E-government • **Applied computing → Document preparation**

KEYWORDS

Open data, Fiscal data, Data heterogeneity

ACM Reference format:

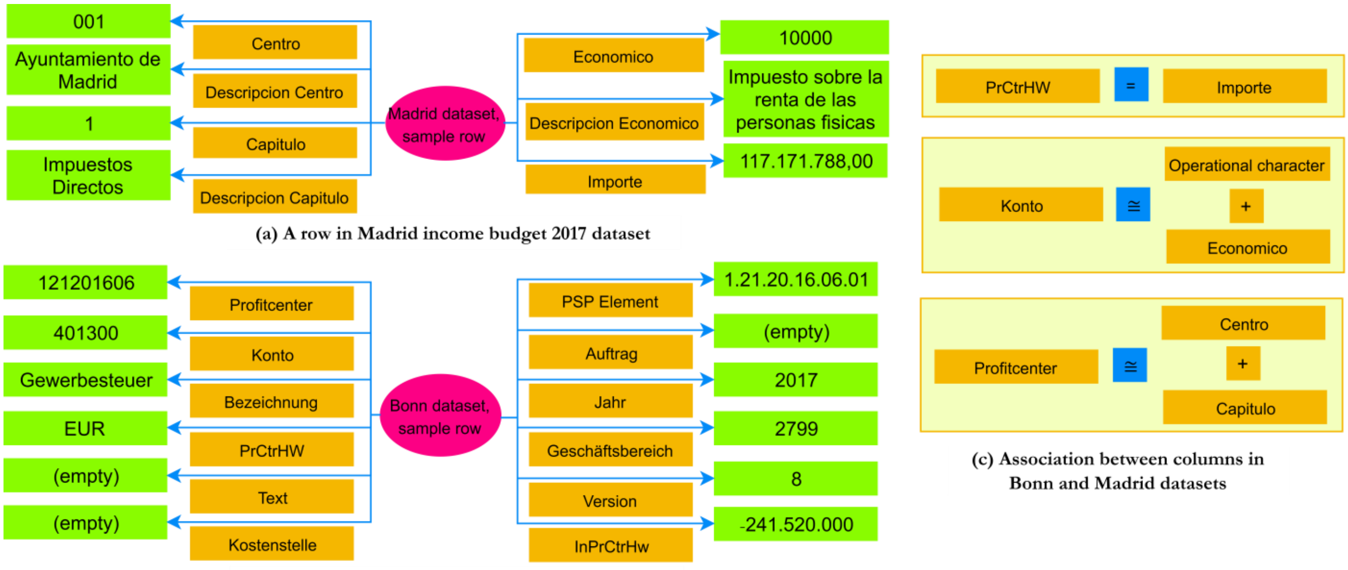
<keep blank / do not remove>. 2018. <keep blank / do not remove>. In *Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance, Galway, Ireland, April 2018 (ICEGOV'18)*, <keep blank / do not remove> pages.
DOI: <keep blank / do not remove>

1 INTRODUCTION

Many public administrators have published budget and spending data as part of their open data program. A survey conducted by Open Knowledge International 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) [1]. 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 [2].

Publishing open data in the domain of budget and spending is often accompanied by different types of *classifications*. For example, during our analysis we found *functional classification* (e.g., elementary education and retirement funds) and *administrative classification* (e.g., Department of Education and Office of Retirement Services). A functional classification lists possible value for items spent/budgeted from the usage perspective. An administrative classification provides the list of offices that manage the budget or spending.

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 of the dataset). Among the factors that contribute to these heterogeneities are the difference of business and budgeting



(a) A row in Madrid income budget 2017 dataset

(b) A row in Bonn budget 2017 dataset

(c) Association between columns in Bonn and Madrid datasets

Figure 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.

process, the coverage level of the administration (e.g., supra-national vs. municipal) or how projects within the public administration are funded.

In this paper, we classify the heterogeneity on budget and spending datasets and correlate these heterogeneities with two state-of-the-art data models designed specifically for budget and spending datasets. We present lessons learned that could be applied to datasets publishers and technical/scientific communities. Currently, we do not cover linguistic and metadata heterogeneity.

This paper is organized as follows: Section 2 provides motivating example, Section 3 briefly presents related work, Section 4 provides analysis of open fiscal data heterogeneity, Section 5 provides design concerns on the available state of the art data model on budget and spending domain, Section 6 defines common terms that are used throughout this paper, while Section 7 links the state of the art data models with enumerated heterogeneities. Section 8 discusses lesson learned from both fiscal data publishers and technical, scientific communities. Finally, Section 9 concludes this paper.

2 MOTIVATING EXAMPLE

Figure 1 (a) illustrates a sample row taken from the City of Madrid's income budget 2017 dataset. Figure 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 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 1 (c). For example, the amount of income is provided in the *PrCtrHw* column in Bonn datasets and in the *Importe* column for 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.

3 RELATED WORK

Current works for modeling heterogeneous fiscal datasets have been done by OpenBudgets.eu (OBEU) with the *OBEU data model* [3] and Open Knowledge International (OKI) with their *Fiscal Data Package (FDP) data model* [4]. The OBEU data model is currently the state of the art data model for fiscal data and was designed based on previous data models. An elaboration of the survey from 14 data models in budget and spending domain is provided in [5].

Open Knowledge International (OKI) has been working on the *OpenSpending* project. By September 2017, OpenSpending has collected 2.238 datasets from 76 countries. OpenSpending provides an open-source technology stack to manage fiscal data, including FDP, which is currently being actively developed by fiscal and transparency communities 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 OpenSpending Viewer tool. Successfully FDP-packaged datasets can also be transformed into OBEU data model. The OBEU data model that is stored in the semantic data server can be queried using a specific API [6].

Many fiscal datasets can be modeled by FDP or OBEU data model, depending on how the fiscal datasets are published. Since there is no binding standard followed by public administrations with regards to fiscal data publishing, datasets can be very heterogeneous. Classification of data heterogeneity on relational databases has been done by Kim and Seo [7]. Their work classified and enumerated general structural heterogeneity of relational databases, including schema and data conflicts. Our work focuses on heterogeneities that occur specifically in the open budget and spending data domain after surveying 77 datasets from different public administrations from various levels.

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)[1] and European Public Sector Accounting Standard (EPSAS)[2]. In this paper, we limit our scope to the structural, contextual and syntactical heterogeneities of fiscal datasets and excluding accounting standards heterogeneities.

4 ANALYSIS OF OPEN FISCAL DATA HETEROGENEITIES

4.1 Example

Two datasets are published by different public administrations from different coverage levels. Both datasets contain different coverage levels and detail, along with different representations, which can be categorized by the *content*, *structure* and *syntax* perspective. Table 1 illustrates the heterogeneities between these two datasets.

Table 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

[1] https://www.ipsasb.org/

[2] http://ec.europa.eu/eurostat/web/government-finance-statistics/government-accounting

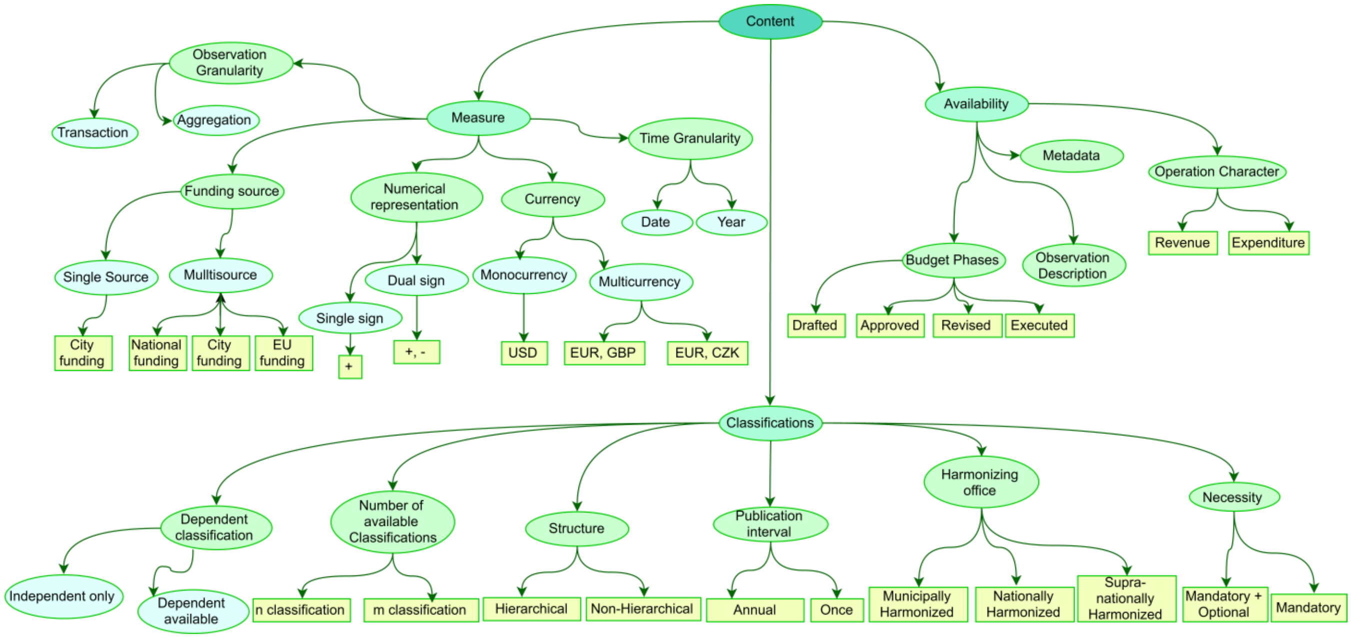


Figure 2: Budget and spending dataset heterogeneity hierarchy from the perspective of content.

No	Heterogeneity	Dataset A	Dataset B
3	*SYNTAX*		
3.1	*File format*	CSV	Excel
3.2	*Character encoding*	ISO-8859-3	UTF-8
3.3	*Metadata*	-	DCAT [7]

4.2 Datasets Example

We have conducted a comprehensive analysis of 77 heterogeneous budget and spending datasets. The spreadsheet of the detailed analysis is available online[3]. These datasets come from different levels (supranational, national, regional and municipalities). Among 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)[4]. 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 that 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*[5]. The Aragon budget datasets contain budget data of Aragon autonomous community from 2006 – 2017.
- *ESIF 2014 – 2022 financing plan datasets*[6]. 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*[7]. This dataset covers the budget from the city of Madrid for the year 2017. The budget covers investment, spending, and income.
- *Swedish national project fund dataset*[8]. This dataset contains project funding in Sweden.

4.3 Heterogeneity Types

This subsection enumerates several types of heterogeneity illustrated with cases from datasets mentioned in section 4.2. Among these datasets, we enumerate several heterogeneities that also likely to occur over other datasets from different public administrations.

1. Content. The hierarchical heterogeneities regarding content are illustrated in Figure 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 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

[3] http://bit.ly/jdiq-datasheet-view
[4] https://goo.gl/BTxmNp
[5] https://goo.gl/LwVazu
[6] https://goo.gl/6kzrir
[7] https://goo.gl/SEMhrL
[8] https://goo.gl/1q5U8D

administrations. For example, ESIF planned funding have several measure columns which separate amount funded by the European Union or 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, negative sign interpreted as revenue, while positive sign indicates expenditure. In case a dataset has both positive and negative sign, interpreting the meaning of these signs should be done carefully by consulting domain expert 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.

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 (i.e., hierarchical but not necessarily having a sub-class relationship). 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, Aragon budget dataset's functional classification has a four-level hierarchy. On the other hand, the classification within 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 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 data cube-based data model.

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 published with metadata.

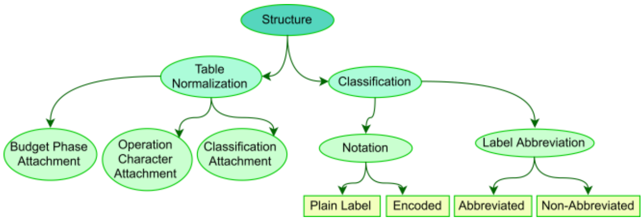


Figure 3: Budget and spending dataset heterogeneity hierarchy from the perspective of the structure.

2. Structure. The heterogeneous hierarchy with regards to the structure is illustrated in Figure 3.

2.1. Table *Normalization*

2.1.1. *Budget phase attachment.* Madrid executed budget datasets[9], for example, provides drafted and approved amount 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 of the file.

2.2. *Classification structure*

2.2.1. *Classification notation.* Some datasets encode their classification in a unique notation, 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 limitation on their datasets. The abbreviation can be a problem if a further effort to analyze the datasets involves techniques such as machine translation or natural language processing. Fortunately, other datasets mentioned above do not contain abbreviated labels.

3. Syntax. The heterogeneous syntax hierarchy is illustrated in Figure 4.

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 HTML page for 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 source public administration by inferring to the ISO 8859 standard [9].

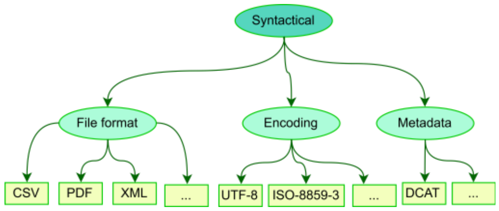


Figure 4: Budget and spending dataset heterogeneity hierarchy from the perspective of the syntax.

3.3. *Metadata.* Different public administrations may provide a different type of metadata. For example, Aragon datasets are provided with DCAT metadata [8], while Bonn dataset is not provided with any metadata.

5 OPENBUDGETS.EU DATA MODEL AND FISCAL DATA PACKAGE

The OBEU data model is modeled after *Data Cube Vocabulary* (DCV) [10]. In the context of the semantic web, *vocabulary* is a set of defined concepts that can be used to annotate information published on the datasets over the web. DCV is a vocabulary that recommended by W3C to publish multi-dimensional data on the web. Multidimensional data include statistics, as well as budget and spending data. Publishing datasets in DCV allows linking to related concepts and datasets. The OBEU data model considers the following modeling patterns, which are extracted from [11]:

1. *Data Structure Definition* (DSD). A DSD is an additional file that provides detailed information regarding every dimension, measure, and attribute that are available in the datasets. Within OBEU data model, a DSD is required.
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 a 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) [12] vocabulary. SKOS supports hierarchical relationship linking among concepts within the datasets by using relationship properties such as *narrower*, *broader*, *narrowMatch*, *broadMatch*, etc.
4. *Integrity Constraints.* To avoid inconsistencies in data modeling, several constraints introduced in the OBEU data

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

model, 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 original datasets source.
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 regarding *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 reduce 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.* OBEU data model recognizes the existence of optional properties in the fiscal domain, even though DCV is strict regarding the cardinality dimension. 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 case may violate the data cube integrity constraints on data cube-based model, which include the OBEU data model.
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.

Fiscal Data Package is another, state of the art, evolving data model. FDP consists the original data in CSV format, accompanied by a JSON file to describe the CSV file. FDP is designed based the following modeling patterns, which are summarized from [4]:

1. Consisted of main dataset/resource and metadata 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 or 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 which 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 [13] by the United Nations and GFSM [14] by International Monetary Fund. In FDP, non-harmonized classifications could be modeled as well.

6 GLOSSARY

Since some of the terms discussed in this paper are rather technical, this section attempt to provide common understanding regarding the definitions used throughout this paper.

- *Dimension, classification,* and *code list. A dimension* defines the qualitative element of a budgeting line [3]. 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 a classification is coined as a *code list.*
- *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 value that is available in a particular observation. The measure is a concept in DCV which specifies the fact being observed. In the context of budget and spending, a measure typically represents the amount of money being budgeted or spent within an observation.
- Spending. *Spending* here defines the actual value that is spent on an item. In this paper, the *executed* budget is considered similar to spending.
- Budget. *The budget* contains a list of planned values to be spent with regards to specified dimensions and attribute. 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 a budget that has been changed

with regard to approved budget, and *executed* budget for the actual value paid after the budget has been spent.

- Expenditure. *Expenditure* refers to the amount of money budgeted to be spent on an item. In this paper, while *expenditure* refers to the budget that may has been or has not been spent, *spending* refers specifically to actual budgeted money that has *already* been spent.
- Income. *Income* refers to the amount of budgeted money which would flow in as revenue for the corresponding public administration.
- Fiscal datasets. *Fiscal datasets* refer to any datasets that elaborate the financial management of public administration. Fiscal datasets may include datasets of the budget, spending, procurement, contracts, beneficiaries and so on.
- Resource Description Framework (RDF). RDF is a standard data model published by the World Wide Web Consortium (W3C). Data in RDF are represented as *triples*. A triple connects two items with a *property* or *predicate* which would facilitate data merging despite schema difference. The first item is called as a *subject*. The second item is known as an *object*[10]. A subject consists of a *Uniform Resource Identifier* (URI) that provides an identifier as a web link to further information regarding the item. An object can be in the form of a URI or a literal. When an RDF datasets are linked to another dataset, they form a *linked data*. Linked open data represented as RDF enables linking across datasets from different sources.

7 LINKING OBEU DATA MODEL AND FDP

The following Table 2 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 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 respective data model as well as the included tools accompanying the data model. For example, 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[11].

Table 2: Support of heterogeneities on the state of the art fiscal data models.

			Data Model	
			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	***Classification***			
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	+	-
		Every time datasets published	*	+
1.2.5	*Classification Harmonization*	Harmonized	*	*
		Non-harmonized	*	*
1.2.6	*Classification Necessity*	Mandatory only	+	+
		Mandatory and optional	*	+
1.3	***Availability***			
1.3.1	*Budget phases*	Drafted	+	+
		Revised	+	+
		Approved	+	+
		Executed	+	+

[10] https://www.w3.org/TR/rdf-concepts/

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

			Data Model *OBEU DM Stack*	*FDP Stack*
No	Heterogeneity	Subheterogeneity	1	2
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	+	+
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	***SYNTAX***			
3.1	*File format*	CSV	+	+
		Excel	+	*
		XML	+	-
		HTML	-	-
		PDF	-	-
		RDF	+	-
3.2	*Character encoding*	Encoding supported	Any	Unknown
3.3	*Metadata*	Metadata type	DCAT	Data Package

8 LESSONS LEARNED

8.1 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 2 above, it can be seen that some of the 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 currency 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 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 are mostly interested in publishing their data to be compatible with FDP stack. FDP-packaged datasets can be transformed semi-automatically using an ETL pipeline [15].

8.2 For Technical and Scientific Communities

Table 2 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 currency (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 are published once - and therefore normalized (in FDP stack), classification which are 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 currently evolving budget and spending data model, as well as the technology stacks to process the budget and spending datasets.

9 CONCLUSION AND FUTURE WORK

In this paper, we present a list of heterogeneities that appear in open fiscal datasets. The heterogeneities are collected after analyzing different datasets from different public administration. A comparison has been made between these heterogeneities and support within state of the art data model. Lessons learned are provided for both datasets publishers and scientific/technical communities.

In the future, we would like to extend the work by analyzing heterogenous, multilingual datasets from different public administration and proposing an approach to map related concepts from the multilingual budget and spending dataset classifications. From the accounting perspective, considering accounting standard heterogeneities would also be a useful contribution to open budget and spending communities.

ACKNOWLEDGEMENTS

The first author is supported by DAAD. EU H2020 Funding no. 645833 funds OpenBudgets.eu (OBEU) project.

REFERENCES

[1] Open Knowledge International, "Government Budget," [Online]. Available: https://index.okfn.org/dataset/budget/.

[2] M. Ballard, "Poor data quality hindering government open data programme," [Online]. Available: http://www.computerweekly.com/news/2240227682/Poor-data-quality-hindering-government-open-data-transparency-programme.

[3] J. Mynarz, V. Svátek, S. Karampatakis, J. Klímek and C. Bratsas, "Modeling Fiscal Data with the Data Cube Vocabulary," in *SEMANTiCS*, Leipzig, 2016.

[4] Open Knowledge International, "Fiscal Data Package (v. 0.3.0)," 27 September 2016. [Online]. Available: http://specs.frictionlessdata.io/fiscal-data-package/. [Accessed 18 09 2017].

[5] M. Dudas, J. Klimek, J. Kucera, J. Mynarz, L. Sedmihradska, J. Zbranek and B. Seeger, "The OpenBudgets Data Model and The Surrounding Landscape," OpenBudgets.eu, Prague, 2016.

[6] L. Ioannidis, C. Bratsas, S. Karabatakis, P. Filippidis and P. Bamidis, "Rudolf: An HTTP API for exposing semantically represented fiscal OLAP cubes," in *International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, Thessaloniki, 2016.

[7] W. Kim and J. Seo, "Classifying schematic and data heterogeneity in multidatabase systems," *Computer,* vol. 24, no. 12, pp. 12-18, 1991.

[8] F. Maali, J. Erickson and P. Archer, "Data catalog vocabulary (DCAT), W3C Recommendation," 16 January 2014. [Online]. Available: https://www.w3.org/TR/vocab-dcat/. [Accessed 19 September 2017].

[9] R. Czyborra, "The ISO 8859 Alphabet Soup," 1 December 1998. [Online]. Available: http://czyborra.com/charsets/iso8859.html. [Accessed 19 September 2017].

[10] R. Cyaganiak, D. Reynolds and J. Tennison, "The RDF Data Cube Vocabulary," 16 January 2014. [Online]. Available: https://www.w3.org/TR/vocab-data-cube/#h2_normalize. [Accessed 19 September 2017].

[11] M. Dudáš, J. Klímek, J. Kučera, J. Mynarz, L. Sedmihradská and J. Zbranek, "Deliverable 1.5: Final release of data definitions for public finance data," 26 October 2016. [Online]. Available: http://openbudgets.eu/assets/deliverables/D1.5.pdf. [Accessed 19 September 2017].

[12] A. Miles and S. Bechhofer, "SKOS simple knowledge organization system reference," World Wide Web Consortium, 2009. [Online]. Available: http://www. w3. org/TR/skos-reference/.

[13] United Nations, "United Nations Statistics Division - Classification registry: Detailed structure and explanatory notes," [Online]. Available: https://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=4. [Accessed 19 September 2017].

[14] International Monetary Fund, "Governance Finance Statistics Manual," IMF, Washington, D.C., 2014.

[15] J. Mynarz, J. Klimek, M. Dudas, P. Skoda, C. Engels, F. A. Musyaffa and V. Svatek, "Reusable transformations of Data Cube Vocabulary datasets from the fiscal domain," in *Proceedings of the 4th International Workshop on Semantic Statistics co-located with 15th International Semantic Web Conference (ISWC 2016)*, Aachen, 2016.