Project Number: 645833 Start Date of Project: 01.05.2015 Duration: 30 months

Deliverable D1.1

Survey of modelling public spending data & Knowledge elicitation report

Dissemination Level	Public
Due Date of Deliverable	Month #3, 31.07.2015
Actual Submission Date	05.10.2015
Work Package	WP 1, Data Structure Definition for Budgets and Public Spending
Task	T 1.1
Type	Report
Approval Status	Final
Version	1.0
Number of Pages	29
Filename	D1.1.docx

Abstract: Goals of this deliverable are to provide an overview of the existing models, approaches and initiatives aimed at publishing budgetary data and to report on knowledge elicitation with domain experts and prospective users. The deliverable is structured into two respective sections: survey of modelling public spending data and knowledge elicitation report.

Project funded by the European Union's Horizon 2020 Research and Innovation Programme (2014 – 2020)

History

Version	Date	Reason	Revised by
0.1	08.07.2015	Initial version	Jakub Klímek
0.2	30.07.2015	Approved version	Paul Walsh
1.0	31.07.2015	Final version	Jakub Klímek

Author List

Organisation	Name	Contact Information
[UEP]	Jakub Klímek	klimek@opendata.cz
[UEP]	Jan Kučera	jan.kucera@vse.cz
[UEP]	Jindřich Mynarz	mynarzjindrich@gmail.com
[UEP]	Lucie Sedmihradská	sedmihradska@centrum.cz
[UEP]	Jaroslav Zbranek	zbranek.jaroslav@gmail.com

Executive Summary

This deliverable serves two purposes. One is to survey the state of the art in data modelling of budget and spending data on the web and in practice. This will inspire the the following OpenBudget.eu data model definition. The second purpose is to deliver a report on knowledge elicitation performed with domain experts in order to gain additional insight into the domain of budget and spending data.

In the survey part of this deliverable, we identified, analysed, described and compared 9 budget data models, 8 spending data models and one combined data model. The data models were used in several datasets in various data formats such as CSV, XML, JSON and RDF. In addition to the data models, we identified legal requirements on budget and spending data in the context of OpenBudgets.eu use cases. There are the budget of the European Union, the structural funds of the European Union and the budget data of regions and municipalities in Spain.

In the second part of this deliverable, we describe the process and results of the knowledge elicitation with domain experts. We interviewed 9 domain experts in 7 interviews, 5 experts were outsiders to the OpenBudgets.eu project. They included 2 public officials, 2 finance statisticians, a policy officer, a journalist and a civil activist. The results revealed potential communication challenges as well as concrete requirements on the OpenBudgets.eu platform and its data model.

Abbreviations and Acronyms

API	Application Programming Interface
COFOG	Classification of the Functions of Government
CSV	Comma-Separated Values
DCV	Data Cube Vocabulary
ESA	European System of Accounts
GTFS	General Transit Feed Specification
IMF	International Monetary Fund
JSON	JavaScript Object Notation
LOD	Linked Open Data
NAC	National Account Code
NACE	Nomenclature Generale des Activites Economiques dans l`Union Europeenne (General Name for Economic Activities in the European Union)
NUTS	Nomenclature d'unités territoriales statistiques
OBEU	OpenBudgets.eu
OLAP	OnLine Analytical Processing
RDF	Resource Description Framework
SDMX	Statistical Data and Metadata eXchange
SKOS	Simple Knowledge Organization System
SPARQL	SPARQL Protocol and RDF Query Language
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
XML	eXtensible Markup Language

Table of Contents

List of Figures

List of Tables

1 Introduction

Goals of this deliverable are to provide an overview of the existing models, approaches and initiatives aimed at publishing budgetary data and to report on knowledge elicitation with domain experts and prospective users. The deliverable is structured into two respective sections: survey of modelling public budgetary data and knowledge elicitation report.

2 Survey of modelling public spending data

In the survey part of the deliverable we describe the most prominent approaches to modelling budget and spending data. We view budget data as planned spending and revenue and optionally also as information about their execution, which is usually aggregated for past fiscal years and does not contain individual transactions. On the other hand, spending data contains the individual transactions, often including identification of the beneficiary, and may contain aggregated data, but does not contain budget plans. In the OpenBudgets.eu platform we aim to have budget and spending data represented as data cubes in RDF format using the RDF Data Cube Vocabulary (DCV). Therefore, part of our main focus is on surveying approaches that already consider DCV.

2.1 The RDF Data Cube Vocabulary

The target data model for the OpenBudgets.eu platform is RDF and the RDF Data Cube Vocabulary[1]. It is a widely used vocabulary for representing multidimensional statistical data and it is compatible with the well-known SDMX (Statistical Data and Metadata eXchange)

[1] http://www.w3.org/TR/vocab-data-cube/

ISO standard. The key terms of DCV and their relationships are depicted in Figure 1.

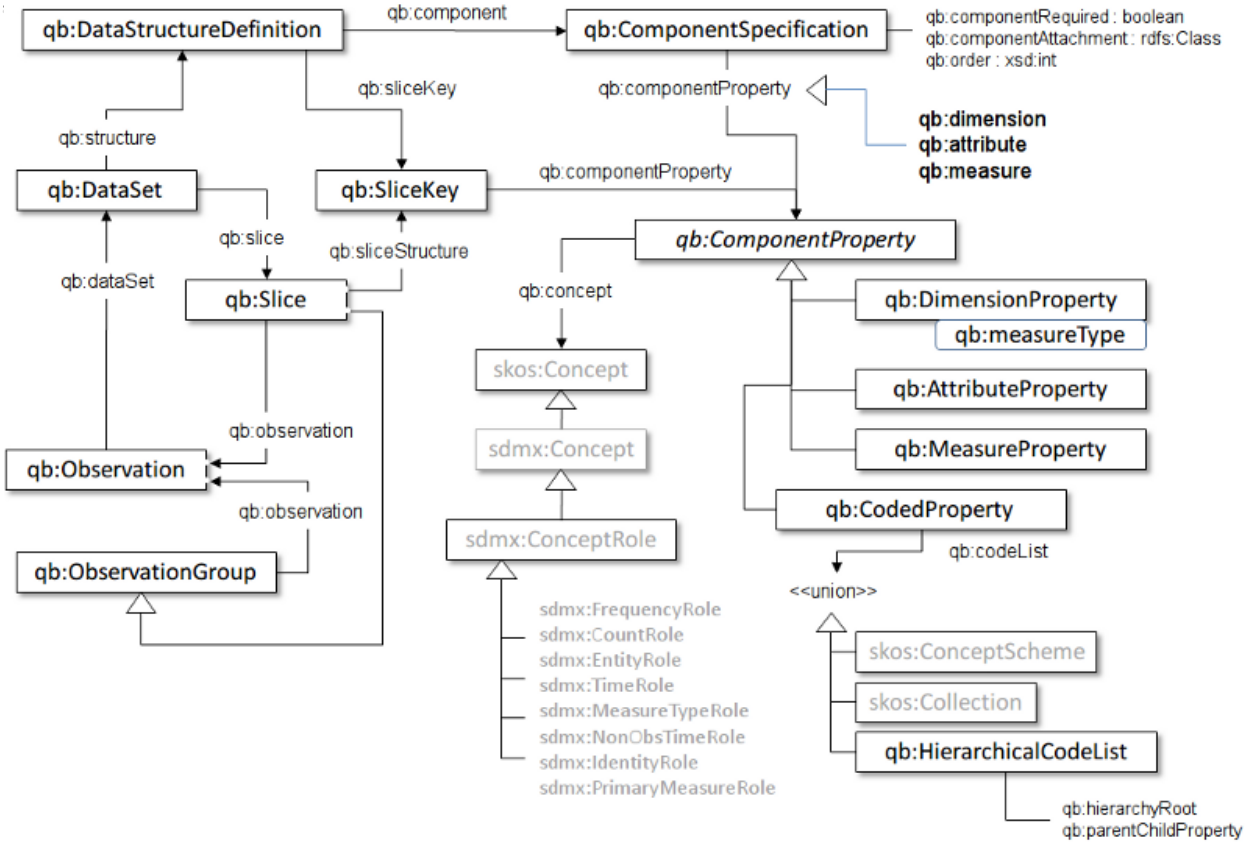

Figure 1 - Key terms and relationships in The RDF Data Cube Vocabulary, source: (Cyganiak & Reynolds, 2014)

A data cube consists of dimensions, which describe properties of individual observations such as time period or geographical region. In the context of budget data, these are typically the fiscal year, organization and budget item category. Then there are measures representing the observed values such as height, width, amount, etc. Again, in the context of budget data, this is typically the budgeted amount of money. Finally, there are attributes, which specify additional properties of the measures, such as unit of measurement or multiplicator. In budget data, this is typically the currency of the budget. The data cube can be sliced by grouping of observations with the same values on selected dimensions, e.g., budget items for a selected fiscal year and a specific organization. Using the RDF Data Cube Vocabulary we model the data structure definition using components (dimensions, attributes, measures) and then the defined components are used to classify individual observations.

Some of the surveyed approaches such as LinkedSpending and the Payments Ontology (see below) already use the RDF Data Cube Vocabulary to model budget and spending data. Therefore, we consider them as a base of the future OpenBudgets.eu data model.

2.2 Budget data models

2.2.1 Modelo ontológico da Classificação das Despesas do Orçamento Federal Brasileiro

This is an official Brazilian ontology for modelling budgets of governmental organizations in RDF[2]. The individual expense items have various amounts, among them the amount planned

[2] http://vocab.e.gov.br/2013/09/loa

in the budget, the amount allocated in the budget and the amount actually paid. Each item is then classified by multiple categories including its economic category, program, project, action, activity, function and fiscal year. Due to our limited ability to translate Portuguese, we did not find more precise definitions of these classifications. Nevertheless, the data is still published using this ontology in 2015 as the only format of their open data.

2.2.2 Brazilian revenue and spending data

Several cities and government organizations in Brazil also publish spending and revenue data[3]. The revenue data is published as an aggregate of anticipated, entered and collected revenue. The spending data is more detailed, as it contains classification by budgetary unit, function and sub-function hierarchy, nature of spending, source of funds, type of tender, number of the process, identification of the beneficiary and the good or service provided. However, this data is in majority not accessible as open data. It is available as HTML, PDF, Excel sheets and only in minority as CSV or XML files. However, we did not manage to actually analyse the files due to the language barrier.

2.2.3 Czech Monitor of the State Treasury

Data from the Czech State Treasury contain multiple reports such as the balance sheet, the profit and loss statement, statement of cash flow, statement of changes in equity etc. available in CSV data files. The budget data contains a hierarchy of planned yearly expenses for each organizational unit of government, regional government, municipality and organization ran by the state. The spending data contains a sum of spending of each organization per year. The hierarchies used to classify parts of budget and expenditure sums are based on the Czech legal system. The hierarchies themselves change in time, which makes it almost impossible to create timelines that span across the hierarchy change without large amounts of manual work. In a recent research project[4] this data was transformed to RDF using the Data Cube Vocabulary and it follows the usual pattern where the dimensions represent the organization, the time period, the classification and the amount of money planned or spent. The represented organizations are classified using classifications such as NUTS, NACE and COFOG.

2.2.4 Local government open data schemas: Budget

Budget[5] is one of many schemas of the UK Local Government Association. The data is available in CSV, XML and JSON with a common structure made of the following properties: Payer specification (Publisher label, Publisher URI, Directorate), Classification (Service, Revenue / Capital), Description, Budget Year and Working Budget (amount). This schema is currently in use only in Redbridge, a London Borough.

2.2.5 Combined On-line Information System (COINS) as Linked Data

COINS is used by UK's HM Treasury to collect financial data from the public sector to e.g., support fiscal management. It contains up to 9 years of data, 5 historic years, the current year, and up to 3 planned years. It is a consolidation system and it does not hold individual financial transactions. The budget items have a hypercube structure consisting of 7 dimensions, 33 attributes and a measure. The dimensions include the department responsible (payer), time, counterparty (payee), data type (budgets or actuals), data sub-type (draft, submitted, approved), account (economic classification) and a programme object

[3] http://www.inesc.org.br/biblioteca/publicacoes/textos/pesquisa-dados-abertos-2014/pesquisa-em-ingles/

[4] (in Czech - http://opendata.vse.cz/tacr/mf/index.html)

[5] http://schemas.opendata.esd.org.uk/details?datasetId=15132

(functional classification). The most interesting attributes are COFOG classification, National Account Code (NAC) and links to various documents related to the budget item. In 2010, COINS data was modelled using the RDF Data Cube Vocabulary in a straightforward way as the original data already had a cube format. A single snapshot from June 14, 2010 was transformed and published as Linked Data and made available through a SPARQL endpoint[6].

2.2.6 The Online System for Central Accounting and Reporting (OSCAR)

OSCAR[7] replaces COINS and publishes aggregated spending data of UKs units of government. The data is published quarterly in a form of MS Excel sheets and CSV files and contains a subset of COINS and is using a variety of classifications.

2.2.7 Open Budgets

There is a web application and API under development as a project of HaSadna (the Public Knowledge Workshop), a non-profit organization in Israel dedicated to data transparency in government[8], that aims to store, access, visualize and compare budget data. Budget data with different structure can be mapped using templates, further details are present in the documentation[9]. It is developed for the Israeli environment but not tied to it. It is very relevant to OpenBudgets.eu as it has similar goals. However, the application demo is not available at the time of writing this survey.

2.2.8 City of Boston Open Budget

The City of Boston has a web application for accessing the city budget[10]. In addition, the underlying data is published in Socrata[11] and consists of planned expenditures. Each expenditure has a fiscal year, recommended amount, approved amount and classification by cabinet, department, program, expense type, expense category, account name and fund name and type.

2.2.9 National Accounts and Government Finances in Denmark

Statistics Denmark provides number of datasets on national accounts and government finances[12]. Datasets in the government finance domain contain data about government budget. The budgetary data is available in multiple classification schemes that include expenditure/revenue classification as well as classification according to the functions defined in COFOG. In addition to the data on national accounts and government budget, regional and municipal accounts and budgets are provided as well.

---

[6] http://data.gov.uk/dataset/coins

[7] https://www.gov.uk/government/collections/hmt-oscar-publishing-from-the-database

[8] https://github.com/pwalsh/openbudgets

[9] http://docs.openbudgets.io/en/latest/

[10] http://budget.data.cityofboston.gov/#/

[11] https://data.cityofboston.gov/dataset/Boston-Open-Budget-Operating-Budget/83wv-akpx

[12] http://www.statistikbanken.dk/statbank5a/SelectTable/Omrade0.asp?SubjectCode=14&ShowNews=OFF&PLanguage=1

Data is available for download in several formats including XLS/XLSX, DBF, CSV and TXT. Other formats suitable for statistical data processing applications such as SAS[13] are available as well.

2.3 Spending data models

2.3.1 Payments Ontology

The Payments Ontology[14] is an ontology based on the Data Cube Vocabulary that is adjusted for modelling fine grained spending data of organizations in the UK. It adheres to the DCV principles. It introduces its own classes and properties and relates them to the original DCV classes and properties using the subclassing mechanism. It distinguishes between 2 levels of detail of spending data. On the payment level of detail, the smallest block of information (an observation) is a payment, which can be represented e.g., by an invoice. If the source data is even more fine grained and for each invoice, it contains individual expenditure lines, then the line level of detail is used. There the observations are the individual expenditure lines (lines of the invoice) and the payment itself (the invoice) is represented as a data cube slice. Either way, each payment can be categorized using any SKOS-like taxonomy and we can distinguish between gross amount and net amount. In addition, there are some pre-defined attributes e.g., for currency. Thanks to the Linked Data principles, one can also describe each entity using other arbitrary properties such as links to other taxonomies. At the time of writing of this text, the Payments Ontology documentation contains some inconsistencies (e.g., invoice and payment mixup in the worked example) and the latest version is Draft 0.2 from 2010. Nevertheless, it remains the best candidate for OpenBudgets.eu spending representation due to its level of documentation among the RDF based data models and also due to existing approaches based on it (e.g., PSNET - see below).

2.3.2 Schema.org Invoice model

The Schema.org initiative contains a model for invoices[15] mainly used in e-commerce. Among the usual properties there is a link to the customer, the minimum and total amount due, the provider of the service (or the goods producer) and the billing period, it links to the orders related to the invoice and provide support for a broker such as a booking agent. While this model is related and it can be used to model the invoices paid, it is not applicable to modelling spending data itself.

2.3.3 OpenSpending.org

The OpenSpending data model[16] provides a generic model that can be instantiated in various ways in concrete spending datasets. It focuses on tabular CSV data and each dataset has 2 mandatory dimensions - a time dimension and an amount dimension. Another requirement is a specification of a key, which uniquely identifies a so called data point, which can be simplified as a row in a table. The key is then specified as one or more existing dimensions. All other dimensions (columns) that are present in the data imported to OpenSpending can be represented too. One needs to name the column with a human readable label and select its data type. The possible data types include “Dimension” - a compound value, “Attribute” - a simple value, “Date” - a temporal value and “Measure” - a monetary value. For some well

[13] http://www.sas.com

[14] http://data.gov.uk/resources/payments

[15] http://schema.org/Invoice

[16] http://community.openspending.org/help/guide/en/modelling-data/

established dimensions such as “time” and “amount”, preferred labels are suggested for better interoperability among different datasets.

2.3.4 OpenSpending Data Package

There is an ongoing activity that aims to define both the logical and physical data model of OpenSpending to store spending data [17]in packages described by a JSON descriptor. Within the descriptor, JSON Table Schema[18] is used to describe the dataset and then there is a mapping section that maps the schema fields to the actual columns in the packaged CSV files. One dataset can be spread over multiple CSV files. OpenSpending Data Package is a superset of the Budget Data Package, but has some differences, namely it does not require some metadata, such as classification by COFOG, it provides the mapping from a physical to a logical data model whereas Budget Data Package forces users to use predefined column names and it allows to attach metadata to the JSON descriptor rather than the CSV files. This allows the users to have other types of data in the package, such as scripts used to create the data, etc.

2.3.5 Linked Spending

In (Höffner, Martin, & Lehmann, 2014) the authors describe the process of automatic conversion of structured OpenSpending.org data into LOD using the Data Cube Vocabulary and SDMX. They also note some unresolved issues such as dataset language detection and mainly the varying level of granularity of each of the OpenSpending.org datasets, which would require a large amount of work to model properly. Therefore the conversion is fairly basic as even the source data is modelled according to the OLAP Data Cube standard and the conversion to the RDF Data Cube Vocabulary is therefore straightforward. The URIs of dimensions, measures and attributes are generated from their names in OpenSpending.org and their collisions are interpreted as their semantic equality. There are, however, some modelling issues, such as the specification of optional dimensions that are not permitted in the RDF Data Cube Vocabulary.

2.3.6 Publicspending.net - The Public Spending Ontology (PSNET)

The Publicspending.net portal collects spending information from 7 payers, namely the United States federal government, Australia, United Kingdom, Greece, State of Massachussetts, City of Chicago and the State of Alaska. The portal contains data for 2011 and 2012 and is still in beta phase, with some of its parts not working properly. The data itself is modelled in RDF, uses the Public Spending Ontology (PSNET) inspired by the UK Payments Ontology, and is registered on datahub.io[19]. The website also provides a SPARQL endpoint through which one can query the dataset. The ontology models individual Payments grouped into Decisions and classified by Common Procurement Vocabulary (CPV), date, and amount. What is somehow missing is the currency of the amount, which is different for each payer as can be seen in the website but is missing in the RDF data. The default currency is specified as being in EUR[20], but it is not clear whether the amount was converted on import and using which currency exchange rate etc. There is an initial report[21] and also a journal paper (Vafopoulos, et al., 2013) describing how the PSNET Ontology was adjusted for

[17] http://labs.openspending.org/osep/osep-04.html

[18] http://dataprotocols.org/json-table-schema/

[19] http://datahub.io/en/dataset/publicspending-net

[20] https://docs.google.com/document/d/16fxFgtjRZC5AU00RiR0jdzbRFU73cBcOGl8ZZECwl6U/edit?pli=1

[21] http://www.w3.org/2012/06/pmod/pmod2012_submission_32.pdf

Greece, however, it is not clear why the authors chose to change the prefix from psnet to psgr when there are no other substantial changes to the ontology described in the paper.

2.3.7 A data standard for transaction-level spending data

This was intended to be an international standard for transaction-level spending data by OpenSpending, inspired by Google's General Transit Feed Specification (GTFS). However, it was superseded the Budget Data Package specification and remained in an early draft version. Nevertheless, we shortly describe the approach here for completeness. The standard describes 9 types of CSV files, 3 required and 6 optional. The required files contain transactions with their id, amount, date, entity (payer) id, supplier (payee) id, and a variety of optional properties. The next required file contains the suppliers with their id and name and optional properties such as tax identification number, OpenCorporates URI, DUNS number, acronym and address. The last required file contains the entities (payers) with a structure similar to suppliers. In addition to the required files, the standard describes the structure of optional files for description of various classifications of transactions including institutional classifications (projects and programmes), economic classifications (accounts and economic types) and functional classifications (functions) each with their id and name and a few optional properties. Transactions can be then classified using optional id references to the individual types of classifications.

Local government open data schemas: Spending

There are multiple schemas for spending data in the UK Local Government Association. One of them is "Council Spending" by Colchester. The data is available in CSV, XML, and JSON and shares common properties. Those are the identification of the data publisher (name and URI) identification of the payer (name and code), identification of the payee (name and code), the effective date, the payment date, the amount, information about VAT irrecoverability, and a reference to a contract. The expenditure is also classified using various taxonomies such as Service, Service Category, Purpose of spend, Procurement Category, CPV and ProClass[22]. Note that the ProClass classification was also available in RDF but is no more. The other spending schemas are various subsets of this one. There is also a guide for publishing spending and procurement information[23].

2.3.8 Federal Spending Transparency (DATA Act)

There is an ongoing activity in the United States to establish government-wide data standards in conjunction with the Digital Accountability and Transparency Act (DATA Act). The goal is to propose a data exchange standard consisting of standardized data elements. The development happens on GitHub[24] and currently there are only a few of the data elements finalized. For example, they use the D&B DUNS number for identification of companies, the ISO 3166-1 Alpha-3 GENC Profile for Country Codes and NAICS codes for procurement classification.

[22] http://proclass.org.uk/

[23] http://www.local.gov.uk/documents/10180/11655/Transparency+guidance+2014+-+spending+and+procurement++20141201.pdf/b4ef3ce9-7f2a-4e5b-86b2-aa417f803e44

[24] http://fedspendingtransparency.github.io/

2.4 Combined data models

2.4.1 Budget Data Package

Budget Data Package[25] is a data model covering expenditures and revenues in either aggregated (covering a whole category) or transactional level of detail. It supports versions of budget such as proposal, approval, and adjustment and also completed transactions (budget execution version). The data format is CSV where each row represents a budget item and a JSON descriptor explaining the structure of the CSV files. The JSON descriptor is a profile that extends the Tabular Data Package specification, which means that it has to adhere to certain formatting restrictions and it has to contain a JSON Table Schema describing the fields of the CSV files and a description of each of the CSV files. The Budget Data Package's extended CSV file metadata includes currency specification, date of last update, date of publication, fiscal year, granularity (aggregated or transactional) and type (expenditure or revenue). Each budget item has to have at least a name, id and an amount. Additional mandatory fields are specified based on type and granularity of represented data and include COFOG and IMF GFSM (expense[26] and revenue[27]) classifications, supplier specification, date of transaction and the government entity responsible for spending the amount. More fields are recommended to be used.

2.5 Comparison of data models

In the attached table we can see a comparison of identified data models. The properties compared are the license under which the model or data is published, the year the model was introduced, the intended data format, the country of origin, and its focus (spending, budget, or both). Note that OGL stands for Open Government License28 and there are two proprietary licenses identified, the one of City of Boston Open Budget[29] and the one of National Accounts and Government Finances[30].

Table 1 - Data models description

Title	Year	Budget-Spending	Countries	Data format	Creator	Level	License
Monitor SP	2014	Budget	CZ	RDF (DCV), CSV	University - CUNI, UEP	Units of government	CC-BY
Payments Ontology	2010	Spending	UK	RDF (DCV-Compatible)	Epimorphics Ltd.	Any organization	OGL
Budget Data Package	2014	Both	Multiple	CSV	Open Knowledge	Any organization	CC-BY-SA 4.0
OpenSpending Data Package	2015	Both	Multiple	CSV, JSON	Open Knowledge	Any organization	CC-BY
LinkedSpending	2015	Spending	Multiple	RDF (DCV)	University - AKSW	Any organization	PDDL 1.0
Modelo ontológico da Classificação das Despesas do Orçamento Federal Brasileiro	2013	Budget	Brazil	RDF	Secretary of Federal Budget (Brasil)	Units of government	CC-BY 3.0

[25] https://github.com/openspending/budget-data-package/blob/master/specification.md#budget-specific-metadata

[26] http://www.imf.org/external/np/sta/gfsm/pdf/text14.pdf, p. 115

[27] http://www.imf.org/external/np/sta/gfsm/pdf/text14.pdf, p. 88

[28] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/

[29] http://www.cityofboston.gov/doit/databoston/data_disclaimer.asp

[30] http://www.dst.dk/en/OmDS/omweb.aspx

The Public Spending Ontology (PSNET)	2013	Spending	EU	RDF		Any organization	No license
A data standard for transaction-level spending data	2012	Spending	Multiple	CSV (GTFS)	Open Knowledge	Any organization	
Local government open data schemas: Spend	2013	Spending	UK	CSV, XML, JSON	Colchester	Units of government	OGL
Local government open data schemas: Budget	2013	Budget	UK	CSV, XML, JSON	Redbridge	Units of government	OGL
Combined On-line Information System (COINS) as Linked Data	2010	Budget	UK	RDF (DCV)		Units of government	OGL
City of Boston Open Budget	2015	Budget	USA	CSV	City of Boston	Units of government	Proprietary
The Online System for Central Accounting and Reporting (OSCAR)	2013	Budget	UK	CSV		Units of government	OGL
National Accounts and Government Finances	Unknown	Budget	Denmark	XLS/XLSX, DBF, SAS, CSV, TXT, TSD, ASB	Statistics Denmark	General government, Central government, Social security funds, Regions, Municipalities	Proprietary

In the second table there is a comparison of the identified data models according to their support of common properties. For each data model and property, the value answers the question “Is the data model able to capture the given dimension?”. Note that the “Payee” dimension applies only to the combined models and models for spending data. Also note that some of the models support arbitrary properties, but only selected properties are understood as “common” and therefore support comparability. When a support for a property in a data model is through this dynamic support of everything, we mark it as Yes*.

Table 2 - Data models properties support

ID	Title	Payer	Payee	Amount	Date	Currency	Tax considered	Transaction (item) ID
MSP	Monitor SP	Yes	No	Yes	Yes	No	No	No
PAYMENT	Payments Ontology	Yes	Yes	Yes	Yes	Yes	Net/Gross	Yes
BDP	Budget Data Package	Yes	Yes	Yes	Yes	Yes	Yes*	Yes*
OSDP	OpenSpending Data Package	Yes	Yes	Yes	Yes	Yes*	Yes*	Yes
LS	LinkedSpending	Yes*	Yes*	Yes	Yes	Yes*	Yes*	Yes
BRAZIL	Modelo ontológico da Classificação das Despesas do Orçamento Federal Brasileiro	Yes	No	Yes	Yes	No	No	No
PSNET	The Public Spending Ontology (PSNET)	Yes	Yes	Yes	Yes	No	No	Yes
TLSD	A data standard for transaction-level spending data	Yes*	Yes	Yes	Yes	No	No	Yes

ESD	Local government open data schemas: Spend	Yes	Yes	Yes	Yes	No	Yes	No
EBD	Local government open data schemas: Budget	Yes	No	Yes	Yes	No	No	No
COINS	Combined On-line Information System (COINS) as Linked Data	Yes	Yes	Yes	Yes	No	No	Yes
BOSTON	City of Boston Open Budget	Yes	No	Yes	Yes	No	No	No
OSCAR	The Online System for Central Accounting and Reporting (OSCAR)	Yes	No	Yes	Yes	No	No	No
DKNAGF	National Accounts and Government Finances	Yes	No	Yes	Yes	No	No	No

2.6 Legal requirements on budget and spending data in the context of OpenBudgets.eu use cases

Benefits of the OpenBudgets.eu platform will be demonstrated by three use case applications of the project outcomes. These use cases will be aimed at:

1. Journalism: this use case shall empower journalists when they report on spending items and it will provide journalists throughout Europe with a tool that makes it easy to understand and communicate budget and spending decisions.
2. Transparency: this use case is aimed at EU policy makers and involves collecting and analysing the EU's budget and the structural and cohesion funds data.
3. Participatory budgeting: The objective of this use case is to facilitate and promote engagement of citizens and other stakeholders in the pre- and post-budget decision-making process. To do so, stakeholders will be given means and tools to give feedback on budget allocations and specific expenditure transactions.

The first use case is mostly focused on tailoring the developed solutions according to the requirements and needs of journalist. Datasets that will be involved in implementation of this use case will be selected based on the discussion with the relevant stakeholders.

In the second use case data about the EU budget and structural and cohesion funds will be used to demonstrate the value of the developed platform and of the open data principles in general to the EU policy makers. Spanish municipalities will be involved in the third use case which is aimed at the participatory budgeting. In order to be able to implement these use cases, legal context of the EU budget, EU structural funds and the Spanish municipal level budgets need to be understood.

2.6.1 Budget of the European Union

The European Union's financial system is based on 3 types of legal instruments (European Commission, 2014, pp. 118-122):

- the provisions of the Treaties, which set basic budget principles and budgetary procedures,
- secondary legislation, e.g., Financial regulation[31], which sets the own resources system, principles, establishment, structure, implementation and auditing of the general budget and principles of budgetary discipline and

[31] http://ec.europa.eu/smart-regulation/evaluation/docs/syn_pub_rf_mode_en.pdf

- provisions adopted by agreement between the institutions, which overcome risks of conflict in the budget procedures.

There are nine principles governing the EU budget (European Commission, 2014, pp. 148-178):

1. The principle of unity
2. The principle of accuracy
3. The principle of universality
4. The principle of annuality
5. The principle of equilibrium
6. The principle of specification
7. The principle of the unit of account
8. The principle of transparency
9. The principle of sound financial management

From the perspective of the modelling of the budgetary data, the structure of the EU budget is one of its most important elements. Structure of the EU budget is determined by the principle of specification which sets both horizontal and vertical structure of the budget.

Horizontal structure divides the EU budget into (European Commission, 2014):

- a general statement of revenue;
- sections that are subdivided into statements of revenue and of expenditure. There are ten sections, one for each European institution;[32]
- section III - Commission is further divided in 32 titles that correspond to the policy areas of the European Commission. Each of the titles is further subdivided into chapters.

Vertical structure of the EU budget is represented by the budget nomenclature. Activity Based Budgeting nomenclature is used to classify revenue and expenditure (European Commission, 2014). According to (European Commission, 2014) the nomenclature is determined during the budgetary procedure.

Titles are further divided into chapters. There is one chapter per activity of the Activity Based Budgeting nomenclature. Slots that accommodate revenue and expenditure are represented by articles (European Commission, 2014). Articles might be further broken down into items.

For each individual item, article, chapter and title the following information are shown:

- appropriations for year t
- appropriations for year t-1
- actual expenditures in year t-2
- explanations about the nature and purpose of the appropriation and references

So called token entries are used in case there is no legal basis for an appropriation or it is difficult to cost new operations or in case of a temporarily stopped operation. A dash is entered to indicate headings (budget lines) which are no longer operational.

It is important to note that the budget nomenclature changes regularly and significantly, e.g., changes[33] between 2013 and 2014 budget.

[32] See the (European Commission, 2014, pp. 162) for more details.

In order to get a deeper insight into the structure of the EU budget we recommend studying the 2015 EU Budget[34] as an example. Please note that the classification of section III-Commission differs from the other sections and therefore it is shown separately.

2.6.2 Structural funds of the European Union

The European Commission has the overall responsibility for implementing the EU budget. According to the Article 58 of the Financial Regulation (European Commission, 2013, pp. 84-87) there are three way the Commission shall implement the budget:

- directly by the Commission (direct management);
- under a shared management with the EU member states (shared management);
- indirectly by entrusting the budget implementation to a defined set of institutions, bodies or persons (indirect management, see Article 58 (1c) of the Financial Regulation).

Within the system of the shared management there are five so called “big funds” - the Structural and Investment funds (European Union, 2009):

- European Regional Development Fund[35] (ERDF),
- European Social Fund[36] (ESF),
- Cohesion Fund[37] (CF),
- European Agricultural Fund for Rural Development[38] (EAFRD),
- European Maritime and Fisheries Fund[39] (EMFF).

Article 35 of the Financial Regulation sets the basis for publication of information on recipients and other information regarding the measures financed from the EU budget. According to the Rules of application of the Financial Regulation (see European Union, 2012) the following information should be published about the recipients, unless specified otherwise:

- the name of the recipient;
- the locality of the recipient
    - the address of the recipient when the latter is a legal person;
    - the Region on NUTS 2 level when the recipient is a natural person;
- the amount awarded;
- the nature and purpose of the measure.

Information about the beneficiaries are available through various web portals depending on the nature of the regime under which they received the funding (European Commission, 2015):

---

[33] http://ec.europa.eu/budget/library/biblio/documents/2014/SEC_2013_370_final_III_en.pdf

[34] http://eur-lex.europa.eu/budget/www/index-en.htm

[35] http://ec.europa.eu/regional_policy/en/funding/erdf/

[36] http://ec.europa.eu/regional_policy/en/funding/social-fund/

[37] http://ec.europa.eu/regional_policy/en/funding/cohesion-fund/

[38] http://ec.europa.eu/agriculture/rural-development-2014-2020/index_en.htm

[39] http://ec.europa.eu/fisheries/cfp/index_en.htm

- direct management - information on the beneficiaries of funds directly managed by the European Commission between 2007 and 2013 and about the beneficiaries of the European Development Fund between 2010 and 2013 are available via the Financial Transparency System[40];
- shared management - each EU member state is responsible for publication of data about the beneficiaries of funds it administers. The funds could be managed by national governments or regional managing authorities. European Commission maintains the following websites that provide access to the national or regional portals providing the data about the beneficiaries:
    - Agricultural policy[41] (direct payments & market-support measures, European Agricultural Fund for Rural Development);
    - Regional development[42] (European Regional Development Fund, Cohesion Fund);
    - Employment[43] (European Social Fund);
    - Fisheries[44] (European Maritime & Fisheries Fund);
- indirect management - data about the beneficiaries funded within the programmes managed by various EU partners can be accessed through the websites of the respective agencies and other EU bodies[45] and the EU institutions and other bodies[46].

2.6.3 Budget data of regions and municipalities in Spain

Nomenclature for classification of the revenue and expenditure in municipal budgets in Spain is regulated at the national level by the Ministry of Finance and Public Administrations[47] via Order EHA/3565/2008[48]. It was later extended to be more precise by Order HAP/419/2014 in 2014[49]. The resulting consolidated text is now available[50].

2.6.3.1 The municipal budget structure

The law specifies the content of the budget i.e. the names of the four different levels of the economic classification (chapter, article, concept, subconcept) and the four levels of the functional one (area, policy, group of programmes, programmes). The law specifies the items used in the first two levels of both the economic and functional categories (i.e. chapter/articles, and area/policies), which municipalities are not allowed to change. It also specifies some "common" elements for the lower two levels (i.e. concept/subconcept, and programmes/ group of programmes), which municipalities should use if possible, but they are

[40] http://ec.europa.eu/budget/fts/index_en.htm

[41] http://ec.europa.eu/agriculture/cap-funding/beneficiaries/shared/index_en.htm

[42] http://ec.europa.eu/regional_policy/country/commu/beneficiaries/index.cfm?lan=en

[43] http://ec.europa.eu/esf/main.jsp?catId=46&langId=en&list=0

[44] http://ec.europa.eu/fisheries/contracts_and_funding/the_european_transparency_initiative/index_en.htm

[45] http://europa.eu/about-eu/agencies/index_en.htm

[46] http://europa.eu/about-eu/institutions-bodies/index_en.htm

[47] Formerly Ministry of Finance (Ministerio de Economía y Hacienda), http://www.minhap.gob.es/es-ES/El%20Ministerio/Historia%20del%20Ministerio/Paginas/Historia.aspx

[48] http://www.boe.es/diario_boe/txt.php?id=BOE-A-2008-19916

[49] http://www.boe.es/diario_boe/txt.php?id=BOE-A-2014-2922

[50] http://www.boe.es/buscar/act.php?id=BOE-A-2008-19916&p=20140319&tn=1#ani

free to add their own elements (programmes, for example) if they need to. The administrative classification is not specified by the law, each municipality can break it down as they prefer.

The actual format of the budget is not specified, so usually a user gets different PDFs for each public body.

2.6.3.2 The municipal budget – Torrelodones

Torrelodones is a city in the province of Madrid which will participate in the OpenBudgets.eu use case scenario. The budget data of this city for years from 2011 to 2015 are available[51], see for example "Presupuesto Inicial de Gastos" (Initial Spending Budget). This budget data of Torrelodones is available also in a form of a visualization[52], yet only until 2014.

2.6.3.3 The municipal budget – Rubí

Another example of Spanish municipal budget concerns Rubí, near Barcelona. Its budget data in Catalan is presented in a bunch of PDFs[53]. The budget data of this city are available for years 2004–2015. This municipality also has its budget data available in form of visualization[54] with the starting year 2011. Such visualization makes the data easier to understand.

The basic structure for both city budgets, for Torrelodones and for Rubí, is the same since both must follow the same legislation. Each of the cities presents budget data in different way and goes down to different levels of detail.

2.6.3.4 Regional budgets

Aragón and Basque Country are also considered by OpenBudgets.eu for a use case scenario. However, they are not municipalities, they are regions. On the level of regions the legal background differs. Regions can be more flexible in regard of budget structure. For example, the code for the Healthy policy is different across regions and they may decide to join or split policies if they want to. Nonetheless the structure of budgets is very similar in case of classifications like functional, administrative or economic. The budget data is broken down in the same levels sharing the same names for these levels (chapters, articles...). The budget data of Aragón are visualized[55]. The budgets of the regions are bigger than those of municipalities, concerning original budget data of Aragón in PDFs[56].

The Basque budget data are visualized[57]. The original PDFs with Basque budget data are published in Basque only[58]. The budget is split across number of files covering different aspects.

---

[51] http://www.torrelodones.es/presupuestos-municipales

[52] http://torrelodones.dondevanmisimpuestos.es

[53] http://www.rubi.cat/fitxers/seu/informacio-financera/pressupost-municipal/pressupost-2014

[54] http://pressupostos.rubi.cat

[55] http://presupuesto.aragon.es

[56] http://aragon.es/DepartamentosOrganismosPublicos/Departamentos/HaciendaAdministracionPublica/AreasTematicas/Presupuestos/PresupuestosAnuales/Presupuesto2015?channelSelected=eed9a4ef3173a210VgnVCM100000450a15acRCRD

[57] http://aurrekontuak.irekia.euskadi.eus

[58] http://www.euskadi.net/k28aVisWar/k28aPrin.jsp

2.6.3.5 Reporting obligation of the municipalities to the higher authority

The municipalities must prepare annual accounts (Cuenta General) at the end of the year. These annual accounts include the actual revenues and expenditures but only at the top economic level. Chapters, areas, balance sheet, a profit and loss statement and a written report summarising what happened during the year are parts of the annual report. These documents include dependent bodies (i.e. public companies owned by the municipality).

These accounts are sent to the Court of Auditors (Tribunal de Cuentas), a national body that checks whether the legislation is being followed correctly. In some regions the Court of Auditors delegate the powers to a regional body. The Court of Auditors has a website explaining some of this process. Part of this information is available in English[59]. This webpage can also be used to access the past accounts (starting from the financial year 2012) for municipalities which submitted them[60]. Unfortunately quite a few municipalities do not submit their accounts, which is illegal but rarely sanctioned.

Some municipalities publish these accounts, like Madrid[61]. These published accounts of Madrid are in very detailed format. Similarly Móstoles, relativelly large city in Madrid, which publishes some PDFs with all this data, i.e. accounts[62]. However the data might be difficult to understand. Often just the final revenues and expenditures ("liquidación") are published, as for example in the case of Torrelodones[63].

On top of this, municipalities have to send their data to the Ministry of the Finance and Public Administrations, but this process i.e. level of detail or data format is not public. In this case the level of detail of data should fulfill the requirements of the European Commission for the calculation of EDP statistics according to Council Regulation (EC) No 479/2009 of 25 May 2009[64]. Currently this process should be fully electronic.

As a consequence of the economic crisis, there has been a closer control of local budgets by the Ministry of Finance and Public Administrations in the past few years. The Ministry of Finance and Public Administrations now has the authority to freeze tax transfers to municipalities if they do not fulfil certain conditions, e.g. paying invoices in time or avoiding overspending. It seems, this is all quite opaque. Nonetheless national authorities responsible for the compilation of EDP statistics are obliged not to provide individual data about the economy of public units in compliance with legislation. This kind of information is classified as sensitive.

Some of the budget and actual spending information sent by the municipalities is then published by the Ministry of Finance and Public Administrations[65] via downloadable Excel spreadsheets or an Access database. Published data sets are not as detailed as the budgets published by municipalities themselves but they are at in a common format.

---

[59] http://www.rendiciondecuentas.es/en/informaciongeneral/cuentageneral/index.html#1

[60] http://www.rendiciondecuentas.es/en/consultadeentidadesycuentas/

[61] http://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Hacienda/Informacion-financiera-y-presupuestaria/Presupuestos/Ejecucion-presupuestaria/Cuentas-anuales/Cuentas-Anuales-del-Ayuntamiento?vgnextfmt=detNavegacion&vgnextoid=4145ac8ccc5c1210VgnVCM2000000c205a0aRCRD&vgnextchannel=2d4bc1258a2f8210VgnVCM2000000c205a0aRCRD

[62] http://www.mostoles.es/es/ayuntamiento/ayuntamiento/estructura-gobierno/concejalia-hacienda-patrimonio-regimen-interior-contratacio/organo-gestion-presupuestaria-contabilidad/4-cuentas-anuales-ayuntamiento

[63] http://www.torrelodones.es/presupuestos-municipales

[64] http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32009R0479

[65] http://serviciosweb.meh.es/apps/EntidadesLocales/

2.7 Survey Conclusions

We have surveyed various approaches to collecting and modelling budget and spending data from the current decade. Quite a few of the approaches already use RDF and some of them even DCV (e.g., Payments Ontology, LinkedSpending, COINS). However, many of those approaches are short-lived - the data was published once or was being published for a short period of time and the schema or ontology froze in a draft stage.

Specifically, we have identified core properties for budget and spending items that in some way appear in majority of data models and data sources and that form an intersection that needs to be unified so that the data can be integrated and comparable. The core properties identified for budget data are:

- payer identification - usually structured into departments, units, etc. This needs to be realized as a URI further described using appropriate models such as The Organization Ontology[66].
- fiscal year
- various versions of budget - drafted, submitted, accepted, actual
- various classifications of budget items
- currency
- amount

The core properties identified for spending data are:

- payer identification (as further described URI)
- payee identification (as further described URI)
- date
- various classifications of spending items
- currency
- amount

Many of the approaches that use RDF and DCV create the data cubes by straightforward mapping of source properties, usually CSV columns, to DCV dimensions, attributes and measures. In many cases, the goal is integration of data from various sources. However, in each data source, properties (columns, dimensions) are named differently when they represent the same thing and sometimes they are named the same when they represent different things. For example, OpenSpending deals with heterogeneity of property names by mapping each data source (physical model) to a logical model, in which the core properties of spending items have standardized names (e.g., time, amount).

Finally, it is clear that while majority of the identified properties in both budget and spending domains such as payer id, date, currency and amount are quite easy to be modelled, used and compared. The real challenge are the classifications, which hold a crucial piece of information for interpretation and aggregation of the individual spending and budget items and which, at the same time, differ among data sources and countries and their mappings are frequently missing.

3 Knowledge elicitation report

Apart from the survey of relevant resources we used knowledge elicitation as a complementary source of understanding of the domain of budgets and as a way to assess user needs from which requirements on the OBEU data model may be derived. We elicited knowledge from domain experts and prospective users of the OBEU platform. By conducting interviews we gathered qualitative observational data, from which we extracted key findings and attempted to translate them to requirements on the developed data model.

[66] http://www.w3.org/TR/vocab-org

Requirements gathered from these interviews can compensate for those derived from the survey of literature, data models, and datasets. In this way, development of the data model for budget data can become more demand-driven in contrast to development driven by the supply of datasets. This way we aim to address a previously described shortcoming:

> *"Too often, standardization in this context appears to be supply-driven: every publisher wants to express the full range of data they hold and are willing to release. Necessarily, such an approach leads to a standard that is the superset of all the systems that feed into it."*[67]

3.1 Knowledge elicitation protocol

The selected knowledge elicitation approach was inspired by the methods for creating ontology requirements specification (Suárez-Figueroa, Gómez-Pérez, Motta, & Gangemi, 2012) and the method for designing a vocabulary for budget data presented by (Brusa, Caliusco, & Chiotti, 2006). We did not commit to a particular methodology but instead hand-picked methods that we deemed appropriate for the kind of data model that we create for the OpenBudgets.eu project. Consequently, heavy-weight ontology engineering methodologies were out of the picture, but instead more informal techniques, such as eliciting competency questions to approximate functional ontology requirements, were adopted.

We decided to carry out knowledge elicitation in a series of interviews. The interviews were semi-structured and each lasted 1 hour. Audio from the interviews was recorded for further transcription and analysis. The interviewees were made aware that they were recorded and recording was done with their prior consent. Results from the interviews were anonymized. Therefore, in the following we refer to the interviewees using their own provided self-identification. Even though we had not adopted an explicit script for the interviews, they revolved around pre-defined topics including:

- **Terminology:** definition of the scope of budget data
- **Linking data:** linking planned and executed expenditures and linking versions of a single budget
- **Data analysis:** comparison of spending items, aggregating budget data, and trend discovery
- **Data quality:** error detection and consistent use of classifications

For each topic we devised several questions and scenarios that we discussed with the interviewees. The open semi-structured format of the interviews was chosen because of its ability to explore ideas brought up by the interviewees while following a few pre-defined concerns. For example, during the interviews we explored the competency questions *"What data do you need to be able to compare 2 monetary amounts?"* or *"What do you need to know in order to be able to associate a payment to a budget line?"*. Regarding the terminology we focused on finding out what the interviewees understood budget data to include (e.g., planned expenditures, actual expenditures, accounting data).

Only a few of the requirements we identified in these interviews address the data model directly. Direct users of the data model are those who are either producing or consuming data described with it. For the most part, the consulted interviewees were end users who interact with budget data primarily through applications. Accordingly, a large share of what they mentioned applied to the application level rather than the level of data. Nevertheless, the requirements on applications may indirectly translate to requirements on data models the applications use. It was up to us to see if the points raised by the interviewees can be translated into concrete requirements on the data model. Therefore, parts of what the interviewees conveyed may be lost in translation and thus our interpretation should be read

[67] http://community.openspending.org/research/standard/introduction

only as approximate requirements. Nevertheless, many of the interviewed persons identified themselves as data analysts and reported interacting with data directly. Few of their concerns were thus related directly to the data model. As a side effect, points addressing the application level gathered during the interviews were fed into the preparation of deliverable D4.2 Analysis of the required functionality of OpenBudgets.eu.

Relevance of the interviews is undermined by the small sample that consisted of 9 interviewees. In total, we conducted 7 interviews with 9 interviewees; meaning that in two occasions we interviewed two persons in one sitting. 5 interviewees were outsiders to the OpenBudgets.eu project, while 4 of them were involved with the project either as use case partners or directly employed domain experts. The interviewed persons included 2 public officials, 2 finance statisticians, a policy officer, a journalist, and a civil activist. 4 of the interviews were done in person, while the remaining 3 interviews were conducted via a teleconference. A shortcoming of the selected sample of interviewees may be a bias towards the Czech environment because 6 out of the 9 interviewed persons were from the Czech Republic.

3.2 Summary of findings

Even though the sample of interviewees was small, recurrent themes and issues emerged. In the following we try to summarize the findings we identified in the interviews. In general, when finding a common language with the interviewees we struggled the most with public officials. We learnt that if we want to reach them as a target user group, we need to be aware of a communication challenge.

3.2.1 Scope of budget

One of the questions we started the interviews with was about what budget is. We asked this question in order to clearly delimit the scope of budget data covered by OBEU and to provide exact definition of budget for the OBEU's data model. In doing so we discovered there is a terminological confusion over the definition of budget. The divergence stems in part from countries' legislations that define budgets in diverse ways. Consequently, there is no exact and shared pan-European understanding of what budget is and as it varies among the EU member states. Moreover, the budget-related law changes frequently and so the definition of budget evolves with it. The interviewees mentioned other authoritative sources of terminology as well, including the European System of Accounts 2010[68] and the Open Budget Survey's methodology by the International Budget Partnership[69]. Nevertheless, we came upon several aspects of budgets that the interviewees agreed on.

In most discourses related to budgets 2 terms are used to distinguish plans and reality (e.g., appropriations and payments). Moreover two bases of accounting might be involved when reporting about the actual revenues and expenditures:

1. **Accrual basis:** accounts for when an expense is incurred
2. **Cash basis:** accounts for when an expense is paid

The key difference between accrual and cash basis is in the period of time for which revenues/expenditures are reported. Usually, there is a delay between the time when an expense is incurred and the time when it is paid. For example, this delay is apparent in case of investments that are typically split into multiple payments paid over an extended period of time. One of the interviewees remarked that spending in accounting may be vastly different from the actual spending and noted that it is important to know this distinction because people can be manipulated into mistaking one for the other and become subject to

[68] http://ec.europa.eu/eurostat/web/esa-2010

[69] http://internationalbudget.org/opening-budgets/open-budget-initiative/open-budget-survey/research-resources/methodology/

accounting tricks. The interviewee also added that in most cases it is difficult to get access to the actual cash flow (public accounting) data of a public body.

We learnt that budget data are typically classified on greater level of detail and better standardized than accounting data.

Based on these concerns we decided to agree within the OBEU consortium to adopt a pragmatic definition of budget data that includes both planned and actual expenditures and revenues, but excludes accounting data. What this implies for the data model is that we will have only revenues and expenditures at an aggregate level according to the classification used in particular budget data. The decision to adopt the above-mentioned scope of budget data will be explicitly documented in the OBEU data model, so that we prevent confusion as much as possible, since we are aware of the issues it may raise. For example, one interviewee marked such understanding of budget as clearly wrong and asserted that budgets contain only the planned expenditures and revenues. However another interviewee pointed out that budget goes through a cycle of phases including planning, execution, and evaluation. Definition of the budget data in OBEU allows us to cover not only the planning phase of the life cycle but other phases as well. This would allow analysing planned vs. actual expenditures/revenues. Interestingly, “cash-flow” was used in the interviews to mean both spending (statistician’s perspective) and public accounting (journalist’s perspective). Therefore, definitions of the used terms will be provided and we will be careful about the used terminology.

3.2.2 Self-describing data

A principal issue of budget data is that it is far from being self-descriptive. Analysis of budget data yielding valid interpretations typically requires not only the data but also a thorough understanding of how budgets work and of the analysed domain. This goes contrary to the principle of self-description proposed for data on the Web (Mendelsohn, 2009), which we plan to pursue in the OBEU data model.

The interviewees mentioned repeatedly that it is difficult to tell what budget data is about. Budget classifications are often too vague, imprecise, or confusing to help determine the subjects of payments. In some cases, even public officials revealed uncertainty when working with budget data. As a result, understanding of budget data remains mostly elusive for the public.

Understanding of the legal context is typically a prerequisite to attempts at correct interpretation of budget data (e.g., knowing which ministry is responsible for the agenda in question). Nowadays, budget data is usually provided in a way that fits public accounting methodologies, so it targets accountants rather than regular citizens. Moreover, users of budget data need to have a solid understanding of the inner workings of the domain where the money is spent. Insider information is especially needed to be able to discover stories in budget data. A story creator needs to know the history of how a budget was made. For example, it is necessary to recognize the political pressures that influenced a budget when it was made. To sum up, while access to budget data is often easy, understanding it is difficult.

In order to address this shortcoming of budget data in the OBEU project we will follow the principles of the semantic web to make budget data as self-describing as possible. However, instead of pursuing detailed ontological modelling grounded in description logic or enforcing elaborate classifications, we will try to achieve this goal by linking external data, such as standards, to provide shared context. In effect, having access to budget data should be a sufficient prerequisite for most analyses.

3.2.3 Data quality

Quality of budget data was usually reported by the interviewees as satisfactory. Experience with the most in-depth quality checks was shared by the interviewed statisticians. They mentioned using logical tests that validate if budget data conforms to the expected rules. These rules can be based on invariants applicable to all budgets. For example, every

municipality in the Czech Republic must have revenue from property tax and failing to report it constitutes an error. Similarly, logical rules may test relations between values in budget data. For instance, rate of interest must correspond to the status of interest-bearing assets. Some errors are revealed when budget data is aggregated (e.g., negative balance usually indicates an error). In fact, different methodologies for aggregation are commonly the root cause of contradicting values found when comparing multiple datasets. Finally, the interviewed statisticians reported using outlier detection in distribution of costs to discover errors (e.g., exceedingly large amounts).

A grave problem of budget data is that in general it is not possible to tell errors from misclassifications. Budget classifications allow some leeway in the ways in which they are applied. This is known as the problem of inter-indexer consistency. Inter-indexer consistency is a *"quantitative measure of the degree to which two or more indexers perceive the important information concepts contained in a document and represent these concepts using identical codes and/or terms"* (Leonard, 1977). In other words, if we apply it to the context of budget data, inter-indexer consistency measures the degree to which multiple public officials agree on classification categories for the same or similar expenditures. For example, a commonly used category may be assigned zero spending, but related spending is classified into a different category. The least consistently used categories turn into classification "black holes". Categories such as "miscellaneous" may account for significant parts of budgets and thus severely limit validity of data analyses. For example, at some point, 95 % of the Brazilian budget was classified as miscellaneous, but it was corrected since. An example of a similar issue was reported for the Czech Republic, where it was discovered that the "Other services" budget line contains mostly expenses on IT services.

The interviewed public officials see harmonization of methodologies as the solution of this issue. They expect that classification methodologies can be made precise enough to make misclassification an error. We agree that more precise methodologies could possibly mitigate the misclassification issue. However it might not be always possible to fully avoid the problem of inter-indexer consistency. We believe that alongside the methodology an improvement in classification can be achieved by network effect fed by public availability of budget data and public officials' desire to conform. It is a challenge we plan to address especially in our work on classifications and code lists used by the OBEU data model.

3.2.4 Data comparison

A common approach to data analysis is comparison. In the context of budget data, undermined by the previously mentioned issues, taking this approach is difficult. The interviewees suggested to treat budget data as incomparable by default. Incomparability may be ascribed to several causes. Perhaps the main one is that budget classifications, methodologies to apply them, and people who do so are different. In some cases, the employed classification methodologies may even be completely unknown. Inconsistent use of classifications makes budget data effectively incomparable.

Valid comparisons usually require having background knowledge about the structure of the compared budgets. For example, an interviewee brought to our attention that there are expenditures, such as fines, that are mostly out of control of the spenders. Such payments can skew the aggregated amounts and so well-founded comparisons should exclude them.

Additionally, as is usually the case for endeavours spanning the EU, another obstacle in comparing budget data is multilinguality. For example, the recipients of EU structural and cohesion funds are required to publish data on the received funds at least in 1 official EU language. To save their effort the authorities will presumably publish the data only in their native tongue. If important data is disclosed in natural language it poses a challenge for cross-country comparison. We expect the linguistic barrier to be a less of an issue for OBEU, since its data model will be based on RDF, which is immune to most of the problems associated with multilinguality, and it will prefer machine-readable data to natural language descriptions.

In the interviews we learnt about two kinds of approaches to making budget data better comparable. One of the approaches is to make data comparable by designing classification crosswalks. In this way, categories from one classification can be mapped to categories from another classification; effectively making the amounts classified with the mapped categories commensurate. For instance, one interviewee reported using an internal classification onto which classifications from the compared datasets were mapped. Similarly, Eurostat ensures comparability by enforcing a single classification for budget data defined in the European System of Accounts (ESA). Consequently, national statistical offices in the EU member states are responsible for devising crosswalks from their local classifications to ESA. In the context of OBEU we will adopt this method by establishing links between classifications and external reference datasets.

A complementary approach to improving comparability is to compare expenditures in relation to contextual data. Rather than comparing absolute values, comparison of relative values is usually more telling. To do so the interviewees reported using macro-economical indicators including gross domestic product, inflation, or average salary. We plan to pay extra attention to incorporating these indicators since comparison of budget data in relation to values drawn from external datasets is fundamental for the data analyses planned in the course of OBEU.

3.2.5 Missing data

Budget data is often not collected on the level of detail that would enable to perform desired analyses. In many cases, data is available only in an aggregated form. It is frequently aggregated in such a way that it is not possible to split it according to the distinction of interest. In such situation, multiple kinds of spending are reported into a single category, while only one kind of spending is of interest. In that situation, the interviewed statisticians reported resorting to qualified estimates supported by additional statistical surveys at times.

To avoid these stumbling blocks the data model developed for OBEU should allow to describe both disaggregated and aggregated data. A preference will be given to disaggregated values, since aggregates can be derived automatically from disaggregated data, whereas the inverse is not the case.

3.2.6 Linking data

Links created for budget data will constitute a key value added by OBEU. Links to contextual data, such as the above-mentioned macro-economical indicators, will be a principal device to enable more intelligent analyses of budget data. This is why we asked our interviewees about what links budget data already contains and what added links would bring the most value.

When it comes to the links budget data is required to have, the interviewed domain experts mentioned that each expenditure must be linked to a single budget line that justifies its existence. Since the presence of these links is a subject of regular audits data consumers can rely on these links being available. In order to allow following money further back to its sources, it is important to establish a connection between budget lines and taxes, so that it is possible to see where taxpayers' money goes. In most countries, there are specific taxes earmarked to be spent for pre-designed purposes (e.g., in the Czech Republic the road tax flows into the budget of the State Fund for Transport Infrastructure). Explicit links between these taxes and respective budgets allows to follow the money and deliver visualizations like Where Does My Money Go?[70].

The interviewees remarked that having the links between expenditures and the budget of the European Union is of particular importance. Currently, it is difficult to distinguish funding originating from the EU's budget and from the national budget. It becomes confusing especially in the case of pre-financing projects that are planned to be subsidized from the EU

[70] http://wheredoesmymoneygo.org

but eventually are not. Knowing the funds drawn from the EU's budget is also vital from the statistical perspective, because these funds must be excluded from the national deficit.

We plan to address these concerns in the OBEU data model. Since all entities in the data model will be identified via URIs, linking them will be enabled by default. We plan to spend ample time on linking the data model and its supporting classifications and code lists as there are 2 forthcoming deliverables devoted to this task: deliverable D1.8 on linking of data structure definitions to vocabularies and deliverable D1.9 on linking code lists to external datasets.

3.3 Knowledge Elicitation Report Conclusions

The purpose of the described work was to elicit input from domain experts and prospective users of the OBEU's outcomes in order to complement our findings based on literature survey. In a series of interviews we covered a range of topics including terminological definitions, requirements for data analysis, data quality issues, and opportunities for linking data. Our next step is to put the gathered input to use in the development of the OBEU data model. We will see how we can sufficiently address the concrete requirements extracted from the elicited findings. Moreover, because of their broader scope the findings will not only feed into the data model design, but inform the general OBEU platform as a whole.

4 References

Brusa, G., Caliusco, M. L., & Chiotti, O. (2006). A process for building a domain ontology: an experience in developing a government budgetary ontology. *Proceedings of the second Australasian workshop on Advances in ontologies* (pp. 7-15). Darlinghurst: Australian Computer Society.

Höffner, K., Martin, M., & Lehmann, J. (2014). LinkedSpending: OpenSpending becomes Linked Open Data. *Semantic Web Journal*.

Leonard, L. E. (1977). Inter-indexer consistency studies. *Occasional papers*. Allikas: http://hdl.handle.net/2142/3885

Mendelsohn, N. (2009). The self-describing web. Retrieved 07 20, 2015, from http://www.w3.org/2001/tag/doc/selfDescribingDocuments

Suárez-Figueroa, M., Gómez-Pérez, A., Motta, E., & Gangemi, A. (2012). Ontology Engineering in a Networked World. Heidelberg, Berlin: Springer.

Vafopoulos, M., Meimaris, M., Anagnostopoulos, I., Papantoniou, A., Xidias, I., Alexiou, G., . . . Loumos, V. (2013). Public spending as LOD: the case of Greece. *Semantic Web Journal*.