OpenBudgets.eu: Fighting Corruption with Fiscal Transparency

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

Deliverable 9.1

Data Management Plan

Dissemination Level	Public
Due Date of Deliverable	Month 6, 01.11.2015
Actual Submission Date	06.11.2015
Work Package	WP 9, Project Management
Task	T 9.2
Type	Report
Approval Status	Final
Version	1.0
Number of Pages	14
Filename	D9.1 Data Management Plan.docx

Abstract: This deliverable describes ways that data are managed within the OpenBudgets.eu project. It outlines a Data Management Plan for the datasets used in the OpenBudgets.eu platform. This plan includes the descriptions of dataset lifecycle, stakeholder behaviors, best practices for data management, data management guideline, and templates for data management used in the OpenBudgets.eu project. This plan will be updated at every milestone cycle.

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability.

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

History

Version	Date	Reason	Revised by
0.1	07.10.15	Initial version	Tiansi Dong
0.2	14.10.15	External review	Judie Attard
1.0	31.10.15	Final version	Tiansi Dong

Author List

Organisation	Name	Contact Information
UBONN	Tiansi Dong	tdong@uni-bonn.de
IAIS	Judie Attard	Judie.Attard@iais.fraunhofer.de

Executive Summary

The first version of this deliverable outlines the strategy for data management to be followed throughout the course of the OpenBudgests.eu (OBEU) project.

This deliverable introduces the dataflow cycle of the OBEU project based on five kinds of stakeholders with descriptions about the various ways they manage OBEU datasets, 13 kinds of best practices for data management, data management guidelines, and templates that will be used for data management for all datasets corresponding to project outputs.

Based on Data Management in H2020 [1], Linked Data Life Cycle (LDLC) [2], Data Value Chain of IBM Big Data & Analytics [3], we present the data management guideline for the OBEU and data management templates for OBEU project as follows:

(1) Data Reference Name – a naming policy for datasets;
(2) Dataset Content, Provenance and Value – general descriptions of a dataset, indicating whether it is aggregated or transformed from existing datasets, or original datasets from data publishers;
(3) Standards and Metadata – descriptions about the format and underlying standards, metadata shall be provided to enable machine-processable descriptions of dataset (supporting data transformation of Any2RDF and RDF2Any);
(4) Data Access and Sharing – it is envisaged that all financial datasets in the OBEU project are freely accessed under the Open Data Commons Open Database License (OdbL). Exceptions shall be stated clearly.
(5) Archiving, Maintenance and Preservation – locations of physical repository of datasets shall be listed for each dataset.

This deliverable is updated at every milestone cycle, to take into account of any additional decisions or newly identified best practices.

Abbreviations and Acronyms

LOD	Linked Open Data
OBEU	OpenBudgets.eu project
DMP	Data Management Plan

Table of Contents

List of Tables

1 Introduction

1.1 Purpose and Scope

A Data Management Plan (DMP) is a formal document that specifies ways of managing data throughout a project, as well as after the project is completed. The purpose of DMP is to support the life cycle of data management, for all data that is/will be collected, processed or generated by the project. A DMP is not a fixed document, but evolves during the lifecycle of the project.

The OBEU project aims at providing a generic framework and concrete tools for supporting financial transparency, to enhance accountability of public administrations and to reduce the possibility of corruption. Objectives of the OBEU are as follows:

(1) publish and integrate financial data using Linked Open Data (LOD);

(2) explore, compare, and (visually) demonstrate financial data;

(3) interactively manage budgets, in the sense that stakeholders and citizens can participate through providing with opinions and comments;

(4) develop a comprehensive platform to realise (1)-(3);

(5) test the platform in three applications – journalism, anti-corruption initiatives, and private citizenship engagement;

(6) establish OBEU as a Software-as-a-Service.

The major block of these aims is the heterogenic nature of data formats used by public administrations, which vary extensively. Examples of the most popular formats used include CSV, EXL, XML, PDF, and RDB. By applying DCAT-AP standard for dataset descriptions and making them publicly available, OBEU DMP covers the 5 key aspects (dataset reference name, dataset description, standards and metadata, access, sharing, and re-use, archiving and preservation), following the guidelines on Data Management of H2020 [1].

1.2 Relation with Work Packages and Deliverables

This deliverable is related to D1.5 “Final release of data definitions for public finance data” [2] and D1.6 “Survey of code lists for the data model’s coded dimensions” [3] which presents existing financial code classifications.

1.3 Structure of the Deliverable

The rest of this deliverable is structured as follows: Section 2 presents the data life-cycle of OBEU, five kinds of stakeholders for the OBEU projects, and 13 best practices for data management. Section 3 describes basic information required for datasets of OBEU project, and guidelines of DMP of OBEU. Section 4 presents DMP templates for data management. Each dataset has a unique reference name. Each data source and each of the transformed form will be described with meta-data, which includes technical descriptions about procedures and tools used for the transformation, and common-sense descriptions for external users to better understand the published data. The Open Data Commons Open Database License (ODBL) is taken as the default data access, sharing, and re-use policies of OBEU datasets. Physical location of datasets shall be provided.

2 Data Lifecycle

The OBEU platform is a Linked Data platform, whose data ingestion and management follow the Linked Data Life Cycle (LDLC) [4]. The LDLC describes the technical process required to create datasets and manage their quality. To ease the process, best practices are described to guide dataset contributors in the OBEU platform.

Formerly, data management was executed by a single person or a working-group, who also took responsibility for data management. With the popularity of the Web and the widely distributed data sources, data management has shifted to a service of a large economic system that has many stakeholders.

2.1 Stakeholders

For OBEU platform, stakeholders are those who have influence on data management, in our case:

(1) Data Source Publisher/Owner refers to organisations those provide financial datasets to the OBEU platform. The communication between OBEU and DSPO is limited to two cases: OBEU downloads financial data from DSPO, and DSPO uploads financial data to OBEU

(2) Data End-User refers to persons and organisations who use the OBEU platform to view financial data, to comment budget policy, and to monitor budget flow. Three end-user examples are entities in the journalism domain, anti-corruption initiatives, and private citizens. All the latter are the key driver for the content of the OBEU platform.

(3) Data Wrangler refers to persons who integrate heterogenic datasets into the OBEU platform. They are able to understand both the terminology used in financial datasets and OBEU data model, and their role is to ensure that the data integration is semantically correct.

(4) Data Analyser refers to persons who provide query results to end-users of OBEU. They may need to use data mining software.

(5) System Administrator and Platform Developer refers to persons responsible for developing and maintaining the OBEU platform.

2.2 The Generic OBEU Data Value Chain

Based on the Data Value Chain of IBM Big Data & Analytics [5], we structure the generic OBEU data value chain as follows:

(1) Discover. An end-user query can require data to be collected from many datasets located within different entities and potentially also distributed in different countries. Datasets hence need to be located and evaluated. For OBEU, the evaluation of datasets results in dataset metadata, which is one of the main best practices in the Linked Data community. DCAT-AP is used as the metadata vocabulary.

(2) Ingest and make the data machine processable. In order to realise the value creation stage (integration, analyse, and enrich), datasets in different formats are transformed into a machine processable format. In the case of OBEU, it is the RDF format. The conversion pipeline from heterogenic datasets into an RDF dataset is fundamental. A Data Wrangler is responsible for the conversion process. For CSV datasets, additional contextual information is required to make the semantics of the dataset explicit.

(3) Persist. Persistence of datasets happens throughout the whole data management process. When a new dataset comes into the OBEU platform, the first data persistence is to backup this dataset and the ingestion result of this dataset. Later data persistence is largely determined by the data analysis process. Two strategies used in data persistence are (a) keeping local copy – copy the dataset from DSPO to the OBEU

platform; (b) caching, to enhance data locality to increase the efficiency of data management.

(4) Integrate, analyse, enrich. One of the data management tasks is to combine a variety of datasets and find out new insights. Data integration needs both domain knowledge and technical knowhow. This is achieved by using a Linked Data approach enriched with a shared ontology. The life cycle of Linked Data ETL process starts from the **extraction** of RDF triples from heterogenic datasets, and storing the extracted RDF data into a storage, that is available for SPARQL querying. The RDF storage can be manually updated. Then, the interlinking and data fusion is carried out, which use ontologies in several public Linked Data sources and creates the Web of Data. In contrast to a relational data warehouse, the Web of Data is a distributed knowledge graph. Based on Linked Data technologies, new RDF triples can be derived, and new enrichment is possible. Evaluation is necessary to control the quality of new knowledge, which further results in searching more data sources, and performing data **extraction**.

(5) Expose. The result of data analysis will be exposed to end-users in a clear, salient, and simple way. The OBEU platform is a Linked Data platform, whose outcomes include (a) meta-data description about the results; (b) a SPARQL endpoint for the meta-data; (c) a SPARQL endpoint for the resulting datasets; (d) a user-friendly interface for the above results.

2.3 Best Practices

The OBEU platform is a Linked Data platform. The best practices for publishing Linked Data are described in [5]. 13 stages are recommended to publish a standalone dataset, 6 of them are vital (marked as **must**).

(1) Provide descriptive metadata with locale parameters

Metadata ***must*** be provided for both human users and computer applications. Metadata provides DEU with information to better understand the meaning of data. Providing metadata is a fundamental requirement when publishing data on the Web, because DSPO and DEU may be unknown to each other. Then, it is essential to provide information that helps DEU – both human users and software systems, to understand the data, as well as other aspects of the dataset.

Metadata should include the following overall features of a dataset: The **title** and a **description** of the dataset; the **keywords** describing the dataset; the **date of publication** of the dataset.; the **entity responsible (publisher)** for making the dataset available; the **contact point** of the dataset; the **spatial coverage** of the dataset; the **temporal period** that the dataset covers; the **themes/categories** covered by a dataset.

Locale parameters metadata should include the following information: the language of the dataset; the formats used for numeric values, dates and time.

(2) Provide structural metadata

Information about the internal structure of a distribution ***must*** be described as metadata, for this information is necessary for understanding the meaning of the data and for querying the dataset.

(3) Provide data license information

License information is essential for DEU to assess data. Data re-use is more likely to happen, if the dataset has a clear open data license.

(4) Provide data provenance information

Data provenance describes data origin and history. Provenance becomes particularly important when data is shared between collaborators who might not have direct contact with one another.

(5) Provide data quality information

Data quality is commonly defined as "fitness for use" for a specific application or use case. The machine readable version of the dataset quality metadata may be provided according to the vocabulary that is being developed by the DWBP working group, i.e., the Data Quality and Granularity vocabulary.

(6) Provide versioning information

Version information makes a dataset uniquely identifiable. The uniqueness enables data consumers to determine how data has changed over time and to identify specifically which version of a dataset they are working with.

(7) Use persistent URIs as identifiers

Datasets ***must*** be identified by a persistent URI. Adopting a common identification system enables basic data identification and comparison processes by any stakeholder in a reliable way. They are an essential pre-condition for proper data management and re-use.

(8) Use machine-readable standardised data formats

Data ***must*** be available in a machine-readable standardised data format that is adequate for its intended or potential use.

(9) Data Vocabulary

Standardised terms *should* be used to provide metadata, Vocabularies *should* be clearly documented, shared in an open way, and include versioning information. Existing reference vocabularies *should* be re-used where possible

(10) Data Access

Providing easy access to data on the Web enables both humans and machines to take advantage of the benefits of sharing data using the Web infrastructure. Data *should* be available for bulk download. APIs for accessing data *should* follow REST (REpresentational State Transfer) architectural approaches. When data is produced in real-time, it *should* be available on the Web in real-time. Data ***must*** be available in an up-to-date manner and the update frequency made explicit. If data is made available through an API, the API itself *should* be versioned separately from the data. Old versions *should* continue to be available.

(11) Data Preservation

Data depositors willing to send a data dump for long term preservation ***must*** use a well established serialisation. Preserved datasets *should* be linked with their "live" counterparts.

(12) Feedback

Data publishers *should* provide a means for consumers to offer feedback.

(13) Data Enrichment

Data *should* be enriched whenever possible, generating richer metadata to represent and describe it.

3 Data Management Plan Guidelines

In this section, we describe guidelines of DMP of OBEU.

3.1 Dataset Content, Provenance and Value

(1) What dataset will be collected or created?

Financial data in any file format from EU members are used as input data to the OBEU platform. They shall be transformed into RDF triple formats.

(2) What is its value for others?

Using the OBEU platform, different stakeholders can easily scrutinise financial data and express their comments on financial policies.

3.2 Standards and Metadata

(3) Which data standards will the data conform to?

Following the Linked Data approach, raw input datasets will be semantically enriched to comply with the RDF standards. The OBEU project will re-use and extend a number of tools of the LinDA project, such as RDF2Any and Any2RDF, and other data transform tools that will be used/developed.

(4) What documentation and metadata will accompany the data?

Following the best practices for data on the web, all ***must*** information described in section 2.3 will be accompanied. The use of W3C standards such as PROV-O for provenance, and DCAT for data catalogue description will be followed.

3.3 Data Access and Sharing

(5) Which data is open, re-usable and what licenses are applicable?

The OBEU project aims at reducing the possibility of corruption through increasing financial transparency. It is envisaged that all financial datasets in the OBEU project should be freely accessed. In particular, the Open Data Commons Open Database License (OdbL) to open datasets is adopted as a project's best practice. Since we only cater for financial datasets within the OBEU project, we do not envisage to have any data of a private or personal nature.

(6) How will open data be accessible and how will such access be maintained?

Data *should* be available for bulk download. APIs for accessing data *should* follow REST architectural approaches. Real-time data *should* be available on the Web in real-time. Data ***must*** be available in an up-to-date manner, with explicitly demonstrated update frequency. For data available through an API, the API itself *should* be versioned separately from the data. Old versions *should* continue to be available. See Section 2.3 10 for detail.

3.4 Data Archiving, Maintenance and Preservation

(7) Where will each dataset be physically stored?

Datasets will be initially stored in a repository hosted by OBEU server, or one of participating consortium partners. Depending on its nature, a dataset may be moved to an external repository, e.g. European Open Data Portal, or the LOD2 project's PublicData.eu.

(8) Where will the data be processed?

Datasets will be processed locally at the project partners. Later, datasets will be processed on the OBEU server, using cloud services.

(9) What physical resources are required to carry out the plan?

Hosting, persistence, and access will be managed by the OBEU project partners. They will identify virtual machines, cloud services for long term maintenance of the datasets and data processing clusters.

(10) What are the physical security protection features?

For open accessible financial datasets, security will be taken to ensure that the datasets are protected from any unwanted tempering, to guarantee the validity.

(11) How will each dataset be preserved to ensure long-term value?

Since the OBEU datasets will follow Linked Data principles, the consortium will follow the best practices for supporting the life cycle of Linked Data, as defined in the EU-FP7 LOD2 project. This includes curation, reparation, and evolution.

(12) Who is responsible for the delivery of the plan?

Members of each WP should enrich this plan from her/his own aspect.

4 Data Management Plan Template

The following template will be used to establish plans for each dataset aggregated or produced during the project.

4.1 Data Reference Name

A data reference name is an identifier for the data set to be produced [1].

Description	A dataset should have a standard name within OBEU, which can reveal its content, provenance, format, related stakeholders, etc.
Metadata	Interpretation, guideline, and software tools shall be given, provided, or indicated for generating, interpreting data reference names.

Table 1 - Template for Data Reference Name

4.2 Dataset Content, Provenance and Value

When completing this section, please refer to questions and answers 1-2 in Section 3.1

Description	A general description of the dataset, indicating whether it has been: ☑ aggregated from existing source(s) ☑ created from scratch ☑ transformed from existing data in other formats ☑ generated via (a series of) other operations on existing dataset The description should include reasons leading to the dataset, information about its nature and size and links to scientific reports or publications that refer to the dataset.
Provenance	Links and credits to original data sources
Operations performed	If the dataset is a result of transformation or other operations (including queries, inference, etc.) over existing datasets, this information will be retained.

Value in Reuse	Information about the perceived value and potential candidates for exploiting and reusing the dataset. Including references to datasets that can be integrated for added value.

Table 2 - Template for Dataset Content, Provenance and Value

4.3 Standards and Metadata

When completing this section, please refer to questions and answers 3-4 in section 3.2

Format	Identification of the format used and underlying standards. In case the DMP refers to a collection of related datasets, indicate all of them.
Metadata	Specify what metadata has been provided to enable machine-processable descriptions of dataset. Include a link if a DCAT-AP representation for the dataset has been published.

Table 3 - Template for Standards and Metadata

4.4 Data Access and Sharing

When completing this section, please refer to questions and answers 5-6 in section 2.3

Data Access and Sharing Policy	It is envisaged that all financial datasets in the OBEU project should be freely accessed, in particular, under the Open Data Commons Open Database License (OdbL). When an access is restricted, justifications will be cited (ethical, personal data, intellectual property, commercial, privacy-related, security-related)
Copyright and IPR	Where relevant, specific information regarding copyrights and intellectual property should be provided.
Access Procedures	To specify how and in which manner can the data be accessed, retrieved, queried, visualised, etc.
Dissemination and reuse Procedures	To outline technical mechanisms for dissemination and re-use, including special software, services, APIs, or other tools.

Table 4 - Template for Data Access and Sharing

4.5 Archiving, Maintenance and Preservation

When completing this section, please refer to questions and answers 6-12 in section 3.4

Storage	Physical repository where data will be stored and made available for access (if relevant) and indication of type: ☑ OpenBudgets partner owned ☑ societal challenge domain repository ☑ open repository ☑ other
Preservation	Procedures for guaranteed long-term data preservation and backup. Target length of preservation.

Physical Resources	Resources and infrastructures required to carry out the plan, especially regarding long-term access and persistence. Information about access mechanism including physical security features.
Expected Costs	Approximate hosting, access, maintenance costs for the expected end volume, and a strategy to cover them.
Responsibilities	Individual and/or entities are responsible for ensuring that the DMP is adhered to the data resource.

Table 5 - Template for Archiving, Maintenance and Preservation

5 Conclusion

This deliverable outlines the guidelines and strategies for data management of OBEU, which will be fine-tuned and extended throughout the course of the project. Following the guideline on Data Management in H2020 [1], we described the purpose and scope of datasets of OBEU, and specified the datasets management for the OBEU project. Five kinds of stakeholders related to OBEU are described: original data producer, data wrangler, data analyser, system administrator/developer, and data end-user; generic data flow chain of OBEU is listed and explained: data discover, data ingest, data persist, data analyse, and data expose. Following the best practices of Linked Data Publishing, we specified the 13 steps of best practices for OBEU dataset management. Based on the above, we present DMP guidelines for OBEU, and DMP templates for data management process during the lifetime of OBEU projects

6 References

[1] European Commission, [Online]. Available: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

[2] Linked Data Stack, [Online]. Available: http://stack.linkeddata.org.

[3] OpenBudgets.eu, “Final Release of Data Definitions for Public Finance Data, Deliverable D1.5,” [Online]. Available: http://openbudgets.eu/assets/deliverables/D1.5.pdf.

[4] OpenBudgets.eu, “Survey of modelling public spending data & Knowledge elicitation report, OpenBudgests.eu Project Deliverable D1.1,” [Online]. Available: http://openbudgets.eu/assets/deliverables/D1.1.pdf.

[5] OpenBudgets.eu, “Survey of Code Lists for the Data Model’s Coded Dimensions, Deliverable 1.6,” [Online]. Available: http://openbudgets.eu/assets/deliverables/D1.6.pdf.

[6] World Wide Web Consortium, “Data on the Web Best Practices,” [Online]. Available: http://www.w3.org/TR/2015/WD-dwbp-20150625.