



open**gov**intelligence

## OpenGovIntelligence

**Fostering Innovation and Creativity in Europe through Public  
Administration Modernization towards Supplying and Exploiting  
Linked Open Statistical Data**

---

### Deliverable 1.1

### OpenGovIntelligence challenges and needs

---

<b>Editor(s):</b>	Evangelos Kalampokis (CERTH)
<b>Responsible Organisation:</b>	CERTH
<b>Version-Status:</b>	V1.0 Final
<b>Submission date:</b>	31/07/2016
<b>Dissemination level:</b>	PU

## Deliverable factsheet

<b>Project Number:</b>	693849
<b>Project Acronym:</b>	OpenGovIntelligence
<b>Project Title:</b>	Fostering Innovation and Creativity in Europe through Public Administration Modernization towards Supplying and Exploiting Linked Open Statistical Data

<b>Title of Deliverable:</b>	D1.1 – OpenGovIntelligence challenges and needs
<b>Work package:</b>	WP1 – Challenges and needs identification
<b>Due date according to contract:</b>	31/07/2016

<b>Editor(s):</b>	Evangelos Kalampokis (CERTH)
<b>Contributor(s):</b>	Marijn Janssen (TUDelft) Tarmo Kalvet (TUT) Robert Krimmer (TUT) Ricardo Matheus (TUDelft) Bill Roberts (SWIRRL) Efthimios Tambouris (CERTH) Konstantinos Tarabanis (CERTH) Maarja Toots (TUT) Dimitris Zeginis (CERTH)
<b>Reviewer(s):</b>	Adegboyega Ojo (NUIG)
<b>Approved by:</b>	All partners

<b>Abstract:</b>	This document includes the results of the four tasks of WP1. In particular, it includes (a) the state of play on open data infrastructures, (b) challenges and needs related to data-driven public sector innovation and public service co-creation, (c) a thorough literature review in linked open statistical data (LOSD), (d) challenges and needs regarding the exploitation of LOSD, and
------------------	--

	(e) challenges and needs regarding the exploitation of open statistical data.
<b>Keyword List:</b>	Challenges, needs, linked data, statistical data, data integration, public services, co-production.

## Consortium

	<i>Role</i>	<i>Name</i>	<i>Short Name</i>	<i>Country</i>
1.	Coordinator	Centre for Research & Technology - Hellas	CERTH	Greece
2.	R&D partner	Delft University of Technology	TU Delft	Netherlands
3.	R&D partner	National University of Ireland, Galway	NUIG	Ireland
4.	R&D partner	Tallinn University of Technology	TUT	Estonia
5.	R&D partner	ProXML bvba	ProXML	Belgium
6.	R&D partner	Swirrl IT Limited	SWIRRL	United Kingdom
7.	Pilot Partner	Trafford council	TRAF	United Kingdom
8.	Pilot Partner	Flemish Government	VLO	Belgium
9.	Pilot Partner	Ministry of Interior and Administrative Reconstruction	MAREG	Greece
10.	Pilot Partner	Ministry of Economic Affairs and Communication	MKM	Estonia
11.	Pilot Partner	Marine Institute	MI	Ireland
12.	Pilot Partner	Public Institution Enterprise Lithuania	EL	Lithuania

## Revision History

<i>Version</i>	<i>Date</i>	<i>Revised by</i>	<i>Reason</i>
0.1	18/04/2016	CERTH	ToC
0.2	01/07/2016	CERTH	Challenges and needs regarding LOSD exploitation
0.3	11/07/2016	TUT	Challenges and needs regarding public service co-production
0.4	13/7/2016	TU Delft	Open Data Infrastructures
0.5	15/7/2016	CERTH	Challenges and needs regarding OSD exploitation, Literature review on LOSD
0.6	20/7/2016	CERTH	Final editing, version sent to NUIG for internal review
0.7	25/7/2016	CERTH, SWIRRL, NUIG	Comments from internal review
1.0	28/7/2016	CERTH	Final version

**Statement of originality:**

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

## Table of Contents

<b>DELIVERABLE FACTSHEET.....</b>	<b>2</b>
<b>CONSORTIUM.....</b>	<b>4</b>
<b>REVISION HISTORY .....</b>	<b>5</b>
<b>TABLE OF CONTENTS .....</b>	<b>6</b>
<b>LIST OF FIGURES .....</b>	<b>9</b>
<b>LIST OF TABLES .....</b>	<b>10</b>
<b>LIST OF ABBREVIATIONS.....</b>	<b>11</b>
<b>EXECUTIVE SUMMARY.....</b>	<b>12</b>
<b>1 INTRODUCTION .....</b>	<b>14</b>
1.1 SCOPE .....	14
1.2 AUDIENCE .....	14
1.3 STRUCTURE .....	14
<b>2 BACKGROUND .....</b>	<b>15</b>
2.1 OPEN GOVERNMENT DATA.....	15
2.2 OPEN STATISTICAL DATA.....	15
2.3 LINKED DATA .....	16
2.4 LINKED OPEN STATISTICAL DATA.....	16
<b>3 METHODOLOGY.....</b>	<b>18</b>
3.1 PA'S CHALLENGES & NEEDS REGARDING DATA-DRIVEN INNOVATION .....	18
3.2 TECHNICAL CHALLENGES & NEEDS OF LOSD.....	20
3.2.1 <i>Linked Open Statistical Data: State-of-the-Art</i> .....	20
3.2.2 <i>Challenges &amp; Needs regarding LOSD Interoperability</i> .....	21
3.3 USERS' CHALLENGES REGARDING THE EXPLOITATION OF OPEN STATISTICAL DATA .....	22
3.3.1 <i>Open Statistical Data Fragmentation</i> .....	22
3.3.2 <i>Challenges &amp; needs of software developers</i> .....	23
3.4 PILOTS' NEEDS ON DATA-DRIVEN INNOVATION .....	23
3.5 DATA INFRASTRUCTURES .....	24
<b>4 PUBLIC ADMINISTRATION'S CHALLENGES &amp; NEEDS REGARDING DATA-DRIVEN INNOVATION.....</b>	<b>25</b>
4.1 KEY CONCEPTS .....	25
4.2 LITERATURE REVIEW .....	26
4.3 SURVEY RESULTS.....	29
4.3.1 <i>Existing experience with open data and co-creation</i> .....	29
4.3.2 <i>Barriers and drivers of open data-driven public service co-creation</i> .....	32
4.3.3 <i>Needs and expectations for open data-driven innovation</i> .....	34
4.3.4 <i>Policy solutions: successful and unsuccessful examples</i> .....	35
4.4 DISCUSSION .....	39

<b>5</b>	<b>TECHNICAL CHALLENGES AND NEEDS OF LINKED OPEN STATISTICAL DATA .....</b>	<b>42</b>
5.1	LINKED OPEN STATISTICAL DATA: STATE-OF-THE-ART .....	42
5.1.1	<i>Conceptual framework</i> .....	42
5.1.2	<i>Quantitative Analysis</i> .....	42
5.1.3	<i>Qualitative Analysis</i> .....	44
5.2	CHALLENGES & NEEDS REGARDING LOSD INTEROPERABILITY .....	50
5.2.1	<i>Identification of conflicting practices</i> .....	50
5.2.2	<i>Understanding of conflicting practices</i> .....	61
5.2.3	<i>Consensus on common practices</i> .....	67
<b>6</b>	<b>USERS' CHALLENGES REGARDING THE EXPLOITATION OF OPEN STATISTICAL DATA .....</b>	<b>69</b>
6.1	OPEN STATISTICAL DATA FRAGMENTATION .....	69
6.2	CHALLENGES & NEEDS OF SOFTWARE DEVELOPERS .....	73
<b>7</b>	<b>PILOTS' NEEDS ON DATA-DRIVEN INNOVATION .....</b>	<b>81</b>
7.1	PILOT 1: THE GREEK MINISTRY OF INTERIOR .....	81
7.2	PILOT 2: ENTERPRISE LITHUANIA – LITHUANIAN MINISTRY OF ECONOMY .....	84
7.3	PILOT 3: TRAFFORD COUNCIL .....	85
7.4	PILOT 4: THE FLEMISH GOVERNMENT .....	86
7.5	PILOT 5: THE MARINE INSTITUTE .....	87
7.6	PILOT 6: THE ESTONIAN MINISTRY OF ECONOMICS .....	88
<b>8</b>	<b>DATA INFRASTRUCTURES: STATE-OF-THE-ART .....</b>	<b>92</b>
8.1	MATURITY LEVEL .....	92
8.1.1	<i>Linked Data Maturity Level</i> .....	94
8.2	DATA INFRASTRUCTURES .....	96
8.2.1	<i>Principles for opening data</i> .....	96
8.2.2	<i>The rise of Open Data Infrastructures</i> .....	97
8.2.3	<i>Infrastructure Operating in an Open Data Ecosystem</i> .....	97
8.2.4	<i>Open Data Infrastructures (ODI)</i> .....	98
8.2.5	<i>Open Data Lifecycle support by infrastructures</i> .....	103
8.3	INFRASTRUCTURE FUNCTIONALITIES IN THE LINKED OPEN DATA LIFECYCLE .....	106
8.3.1	<i>Comparing open data life cycles</i> .....	106
8.3.2	<i>An integrated open data life cycle</i> .....	106
8.3.3	<i>LOD Creation</i> .....	108
8.3.4	<i>LOD Publication</i> .....	109
8.3.5	<i>LOD Usage</i> .....	114
8.3.6	<i>Data Curation</i> .....	115
<b>9</b>	<b>CONCLUSION .....</b>	<b>116</b>
	<b>REFERENCES .....</b>	<b>118</b>
	<b>APPENDIX A: QUESTIONNAIRE FOR PA'S CHALLENGES .....</b>	<b>142</b>
	<b>APPENDIX B: QUESTIONNAIRE FOR LOSD TECHNICAL CHALLENGES .....</b>	<b>145</b>
	<b>APPENDIX C: QUESTIONNAIRE FOR DEVELOPERS .....</b>	<b>152</b>

**APPENDIX D: INTERVIEW FORM FOR PILOT PARTNERS ..... 156**



## List of Figures

FIGURE 1 MULTI-DIMENSIONAL DATA MODELLED AS A CUBE .....	15
FIGURE 2 THE RDF DATA CUBE VOCABULARY.....	17
FIGURE 3 CONNECTION OF WP1 TASKS TO THE SECTIONS OF THE DELIVERABLE.....	18
FIGURE 4 THE STEPS FOR ELICITING CHALLENGES AND NEEDS ON LOSD EXPLOITATION .....	22
FIGURE 5 THE LINKED OPEN STATISTICAL DATA FRAMEWORK .....	42
FIGURE 6 TYPES OF DATA INTEGRATION CONFLICTS .....	55
FIGURE 7 SEARCH RESULTS TO DATASETS & LINKS ABOUT UNEMPLOYMENT IN DATA.GOV.UK.....	70
FIGURE 8 PORTALS OF DATA.GOV.UK .....	71
FIGURE 9 PORTALS OF DATA.GOV.UK BY MODES OF PROVIDING RELEVANT & NON-PDF DATASETS .....	72
FIGURE 10 IMPORTANCE OF OPEN STATISTICAL DATA .....	73
FIGURE 11 DEVELOPERS' PERCEPTION OF THE VOLUME OF OSD .....	74
FIGURE 12 SOURCES OF OSD .....	74
FIGURE 13 NUMBER OF PORTALS PER APPLICATION .....	75
FIGURE 14 QUALITY OF OSD .....	75
FIGURE 15 CHALLENGES ON OSD DISCOVERY .....	76
FIGURE 16 CHALLENGES ON COMBINING OSD .....	77
FIGURE 17 STORAGE OF OSD .....	77
FIGURE 18 STATISTICAL ANALYSES ON OSD.....	78
FIGURE 19 VISUALISATIONS ON OSD.....	78
FIGURE 20 DEVELOPING TEAM OF OSD APPLICATIONS .....	79
FIGURE 21 PROFESSION OF QUESTIONNAIRE RESPONDERS.....	79
FIGURE 22 EMPLOYMENT STATUS OF QUESTIONNAIRE RESPONDERS.....	80
FIGURE 23 - OGD CLASSIFICATION SCHEME .....	92
FIGURE 24 - OPEN GOVERNMENT MATURITY MODEL (OGMM) .....	93
FIGURE 25 - INTEROPERABILITY MATURITY STAGES FOR BIG AND OPEN DATA.....	94
FIGURE 26 - LINKED DATA MATURITY LEVEL .....	94
FIGURE 27 - OPEN DATA ECOSYSTEM.....	98
FIGURE 28 - FACTORS THAT INFLUENCE OPEN DATA INFRASTRUCTURES.....	100
FIGURE 29 - OPEN GOVERNMENT DATA LIFECYCLE (TAKEN FROM ATTARD ET AL. (2015)).....	103
FIGURE 30 - GENERIC SCENARIO FOR DATA PROVISION .....	104
FIGURE 31 - OPEN DATA VALUE GENERATED BASED ON FOCUS AND STAKEHOLDER .....	105
FIGURE 32 - METADATA SYSTEM DESIGN.....	110

## List of Tables

TABLE 1 OFFICIAL LOSD PORTALS.....	52
TABLE 2 LITERATURE ON DATA INTEGRATION CONFLICTS.....	52
TABLE 3 - PROPOSED QUESTIONNAIRE FOR EVALUATING THE OPEN LINKED DATA MATURITY LEVEL OF PILOTS.....	95
TABLE 4- FACTORS THAT INFLUENCE OPEN DATA INFRASTRUCTURES.....	99
TABLE 5 - OBJECTIVES OF ODI.....	100
TABLE 6 - OVERVIEW OF REQUIREMENTS AND FUNCTIONALITIES FOR OPEN DATA INFRASTRUCTURE.....	101
TABLE 7 - STAGES OF OPEN GOVERNMENT DATA LIFECYCLE.....	105
TABLE 8 –COMPARING LINKED OPEN DATA LIFECYCLE APPROACHES.....	106
FIGURE 9 - STAGES OF LOD LIFECYCLE.....	107
TABLE 10 – LOD LIFECYCLES STAGES, STEPS AND DESCRIPTIONS.....	108
TABLE 11 - BENEFITS OF METADATA CREATION FOR LOD.....	111
TABLE 12 - DISADVANTAGES OF METADATA FOR LOD.....	112

## List of Abbreviations

The following table presents the acronyms used in the deliverable in alphabetical order.

<i>Abbreviation</i>	<i>Description</i>
DCLG	UK Department for Communities and Local Government
EC	European Commission
ICT	Information & Communication Technologies
ISTAT	Italian National Institute of Statistics
LOD	Linked Open Data
LOSD	Linked Open Statistical Data
NGO	Non-Governmental Organisation
ODI	Open Data Institute
OKFN	Open Knowledge Foundation
OSD	Open Statistical Data
OGD	Open Government Data
PA	Public Administration
QB	RDF Data Cube vocabulary
QUDT	Quantities, Units, Dimensions and Data Types
RDF	Resource Description Framework
SDMX	Statistical Data and Metadata eXchange
SKOS	Simple Knowledge Organization System
TR	Technical Report
WP	Work Package

## Executive Summary

This document is the first deliverable, entitled D1.1 “OpenGovIntelligence Challenges & Needs”, of the first work package of the OpenGovIntelligence project. The objective of the OpenGovIntelligence project is to provide a holistic approach for the modernisation of Public Administration (PA) by exploiting Linked Open Statistical Data (LOSD) technologies thus stimulate sustainable economic growth in Europe through fostering innovation in societies and enterprises.

Work package one (WP1) is responsible for eliciting the challenges and needs regarding political, legal, institutional, social and technical issues in opening-up and exploiting LOSD for the co-production of innovative data-driven public services. WP1 comprises four tasks with the following objectives: Task 1.1 aims to review existing government and other data infrastructures, Task 1.2 aims to elicit needs and expectations for service co-production, Task 1.3 aims to identify LOSD challenges and needs, and Task 1.4 aims to understand how PA can be transformed to innovation based on better use of data.

The methodology that was followed in this deliverable comprises the following actions:

- Identifying *PA’s challenges and needs* related to open data-driven innovation and public service co-creation. Towards this end, a thorough literature review was performed and an online survey aiming at both public and private sector was organised.
- Identifying *technical challenges and needs* of LOSD. Towards this end, the LOSD scientific literature was studied and 11 LOSD top experts from around the globe were involved.
- Identifying *users’ challenges and needs* regarding the exploitation of Open Statistical Data that are provided by Open Data Portals. Towards this end, we first played the role of a user and we studied the UK Open Data Portal in order to discover statistical data about a specific phenomenon. Moreover, we elicited the opinion of developers who have used Open Statistical Data to create data-driven applications.
- Identifying the *needs of the OpenGovIntelligence pilot partners* regarding data-driven innovation. Towards this end, a number of usage scenarios were specified. The scenarios describe the challenges of the pilots, the available datasets, and the final data-driven public service that the pilots want to co-produce.
- Identifying the *State-of-the-art of data infrastructures*.

These actions resulted in a set of outcomes that will feed future work in the project. More specifically, the results of this deliverable include:

- *PA’s challenges and needs* regarding data-driven innovation (Section 4). This includes challenges such as lack of political and administrative priority and leadership, incompatibility of existing administrative and organisational cultures with the idea of co-creation and resistance of public administrators to fundamental changes.
- *Technical challenges and needs of LOSD* that are presented in the following way (Section 5):
  - A mapping and analysis of the LOSD state-of-the-art. This includes a documentation of existing software tools, vocabularies, architectures, and use cases that can be re-used and re-purposed in the frame of the project.
  - A set of LOSD publishing practices that cause LOSD interoperability challenges. These technical challenges are accompanied by a list of advantages and disadvantages per

practice based on top experts' opinions. These results can feed the foreseen project standardisation activities in WP5, so that the community to agree on a list of best practices for LOSD publishing.

- *Users' challenges regarding the exploitation of statistical data* as they are provided through major Open Data Portals (Section 6). These challenges are presented in the form of:
  - An empirical description of the “Open Statistical Data Fragmentation” challenge based on the analysis of data.gov.uk. In summary, our research revealed that searching this portal for useful data on unemployment produces 122 results that provide access to 56 files and 610 links to 18 other portals (such as the Office for National Statistics and the National Archives) and by following the relevant links to more than 2,000 other files.
  - Analysis of the responses of 24 developers that have exploited Open Statistical Data in order to produce data-driven applications. Some of the challenges that they have faced include interoperability among datasets, the lack of metadata, and the quality of metadata.
- *Usage scenarios that reflect pilot's needs* regarding data-driven innovation in PA (Section 7). These scenarios will feed into WP4 towards the development of the evaluation scenarios.
- *The state-of-the-art regarding open data infrastructures* (Section 8). This comprises a maturity model for LOSD and a lifecycle for data infrastructures.

## 1 Introduction

The aim of this section is to introduce the background of the work pursued with WP1 “Challenges and needs identification” of the OpenGovIntelligence project. The scope and the objective that the current document has set out to achieve are presented in sub-section 1.1. The intended audience for this document is described in sub-section 1.2 while sub-section 1.3 outlines the structure of the rest of the document.

### 1.1 Scope

The present document is the deliverable “D1.1 - OpenGovIntelligence challenges and needs” (henceforth, referred to as D1.1) of the OpenGovIntelligence project. The main objective of D1.1 is to document the results of the four tasks of WP1:

- T1.1 State-of-the-art of data infrastructures
- T1.2 Elicitation of needs and expectations for service co-production
- T1.3 Identification of LOSD challenges and needs
- T1.4 Innovation in relation to decision making

These results will then feed into the creation of the OpenGovIntelligence Framework in WP2, the development of the OpenGovIntelligence ICT tools in WP3, and the specification of the pilots and evaluation plan in WP4.

### 1.2 Audience

The intended audience for this document is the OpenGovIntelligence consortium, the European Commission (EC) and those who are interested in challenges and needs for opening-up and exploiting LOSD for the co-production of innovative data-driven services on governments.

### 1.3 Structure

The structure of the document is as follows:

- Section 2 provides the basic background concepts that are needed for the understanding of the content of this deliverable.
- Section 3 presents in detail the methodology that we followed in order to achieve the objectives of the four tasks of WP1.
- Section 4 presents PA’s challenges and needs regarding data-driven innovation and public service co-creation.
- Section 5 documents the technical LOSD challenges.
- Section 6 presents users’ challenges regarding the exploitation of statistical data as they are provided through major Open Data Portals.
- Section 7 presents specific usage scenarios that reflect pilot’s needs regarding data-driven innovation in PA.
- Section 8 documents the state-of-the-art regarding open data infrastructures.
- Finally, section 9 draws conclusions.

## 2 Background

### 2.1 Open Government Data

The term “Open Data” springs from some of the same roots as “Open Source” or “Open Access”. Although “Open” in software normally means libre (i.e. free in the sense of having no restrictions), there is an increasing movement towards using “Open Access” to mean gratis (i.e. free in the sense of costing no money) and not libre. The GNU project suggests that Open Source (or Free) software is a matter of liberty, not price, and means that “the users have the freedom to run, copy, distribute, study, change and improve the software”.

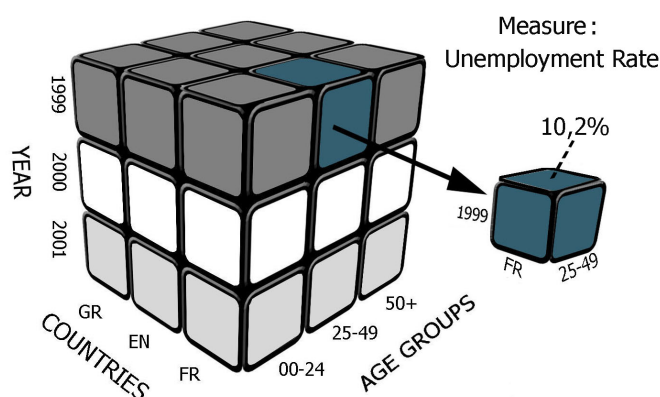
For the purposes of this deliverable, open data is as defined by the OKF: “data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike”.

We also use the term Open Government Data (OGD) to refer to government data in which the open data definition applies.

### 2.2 Open Statistical Data

A major part of open data concerns statistics such as demographics and economic indicators. For example, the vast majority of the datasets published on the open data portal of the European Commission are of statistical nature. Major providers of statistics at the international level include Eurostat, World Bank, OECD, and CIA’s World Factbook.

Statistical data is often organised in a multidimensional manner where a measured fact is described based on a number of dimensions, e.g. unemployment rate could be described based on geographic area, time and gender. In this case, statistical data is compared to a cube, where each cell contains a measure or a set of measures, and thus we onwards refer to statistical multidimensional data as data cubes or just cubes.



**Figure 1 Multi-dimensional data modelled as a cube**

A data cube is specified by a set of dimensions and a set of measures. The dimensions create a structure that comprises a number of cells, while each cell includes a numeric value for each measure of the cube. Let us consider as an example a cube from Eurostat with three dimensions, namely time

in years, geography in countries, and age group, that measures the employment rate. An example of a cell in this cube would define that the percentage of unemployed people between 25 and 49 years old in France in 1999 is 10.2%.

## 2.3 Linked Data

The term Linked Data refers to “data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external datasets, and can in turn be linked to from external datasets” (Bizer et al., 2009, p. 2).

More specifically, the four Linked Data principles as described by Berners-Lee (2010) are the following:

- All item should be identified using URIs;
- All URIs should be dereferenceable, that is, using HTTP URIs allows looking up the item identified through the URI;
- When looking up a URI it leads to more data, which is usually referred to as the follow your nose principle;
- Links to other URIs should be included in order to enable the discovery of more data.

## 2.4 Linked Open Statistical Data

The RDF Data Cube (QB) vocabulary (Cyganiak & Reynolds, 2014) is a W3C standard for modelling data cubes as graphs and thus adhering to the RDF model and Linked Data principles. Centric class in the vocabulary is `qb:DataSet` that defines a cube. A cube has a `qb:DataStructureDefinition` that defines the structure of the cube and multiple `qb:Observation` that describe each cell of the cube. The structure is specified by the abstract `qb:ComponentProperty` class, which has three sub-classes, namely `qb:DimensionProperty`, `qb:MeasureProperty`, and `qb:AttributeProperty`. The first one defines the dimensions of the cube, the second the measured variables, while the third structural metadata such as the unit of measurement.

Usually the values of the components are populated using predefined code lists that might formulate hierarchies such as a geographic or administrative division. These code lists can be specified by using either the Simple Knowledge Organization System (SKOS) (Miles & Bechhofer, 2009) vocabulary or the QB vocabulary. SKOS is a W3C standard used for expressing the basic structure and content of concept schemes such as thesauri, taxonomies, and classification schemes. The set of values is modelled as a `skos:ConceptScheme` and a value as a `skos:Concept`. In addition, `skos:broader` and `skos:narrower` are used to assert a direct hierarchical link between two `skos:Concepts`. In case of reusing RDF data that are not modelled using SKOS, the QB vocabulary introduced the `qb:HierarchicalCodeList` class that defines a set of root concepts in the hierarchy (`qb:hierarchyRoot`) and a parent-to-child relationship (`qb:parentChildProperty`).

At the moment, a number of statistical datasets are freely available on the Web as linked data cubes. For example, the European Commission’s Digital Agenda provides its Scoreboard as linked data cubes. An unofficial linked data transformation of Eurostat’s data, created in the course of a research project, includes more than 5,000 linked data cubes. Few statistical datasets from the European Central Bank, World Bank, UNESCO and other international organisations have been also transformed to linked data



in a third party activity. Census data of 2011 from Ireland and Greece and historical censuses from the Netherlands have been also published as linked data. Finally, the Department for Communities and Local Government (DCLG) in the UK also provides local statistics as linked data.

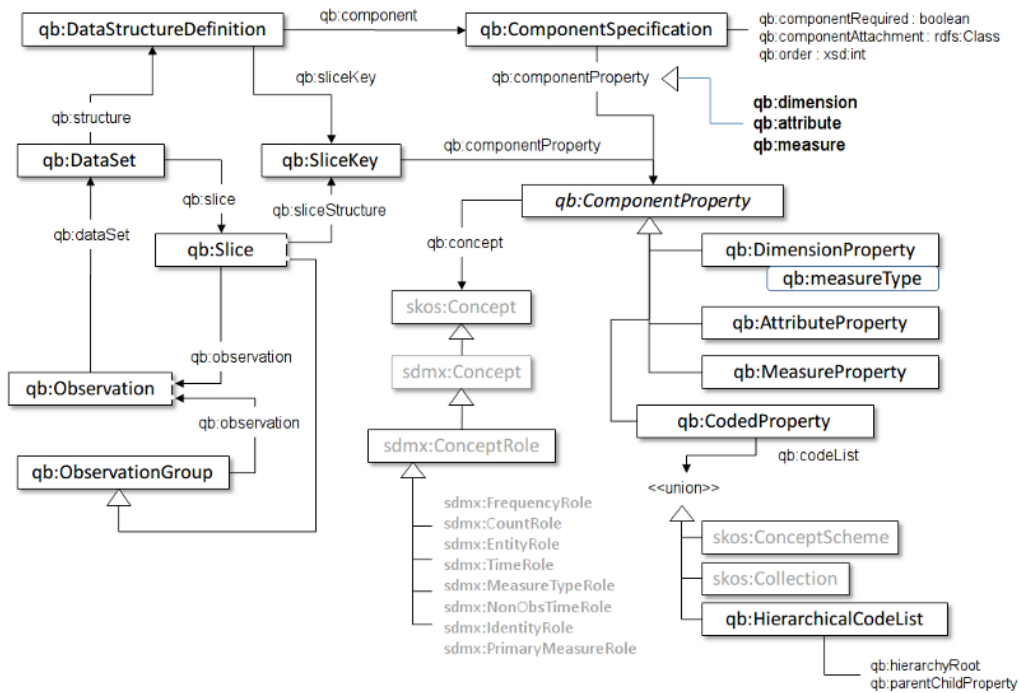


Figure 2 The RDF data cube vocabulary

### 3 Methodology

In this section we present the methodology that we follow in order to achieve the objectives of the four tasks of WP1. Figure 3 presents the connections of the four tasks of the WP to the steps of our methodology and the respective sections of this deliverable. The steps are described in detail in the rest of this section. The results of the steps are described in section 3-7 of this deliverable.

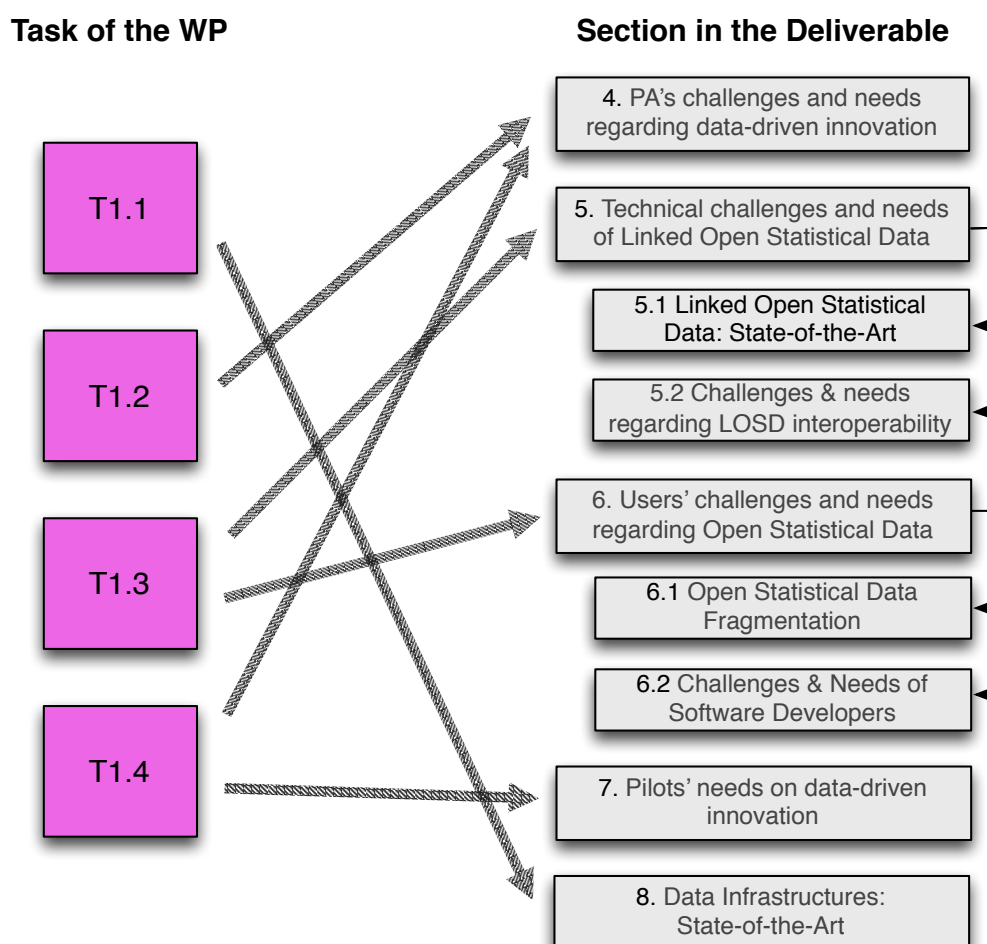


Figure 3 Connection of WP1 tasks to the sections of the deliverable

#### 3.1 PA's challenges & needs regarding data-driven innovation

The objective of Tasks 1.2 and 1.4 is to understand how service co-production can be performed in data-driven services, and understand innovation in relation to accountable decision-making based on better use of data. More specifically, Task 1.2 aims at eliciting core needs and expectations for service co-production and understanding how the co-production of public services can be applied to the production of qualitative data-driven public services. Task 1.4, on the other hand, aims to realize how public administration can innovate and transform its decision-making procedures based on better use of data.

Based on the above, a study was carried out to fulfil the objectives of Tasks 1.2 and 1.4, addressing the following research questions:

1. What are the core needs and expectations of different stakeholders in the data-driven co-production of public services?
2. How has co-creation/co-production been used in the design and provision of data-driven public services?
3. What drivers and barriers affect the co-creation/co-production of open data-driven public services?
4. What are stakeholders' expectations to public administrations with regard to innovating and transforming their decision-making procedures based on better use of data?
5. What policy solutions and initiatives could be taken to foster the co-creation of innovative services driven by the use of open data?

The methods used in data collection and analysis were qualitative in nature. The research process consisted of the following steps:

1. Literature review (March-May 2016)
2. Elaboration of a written interview/survey questionnaire (March 2016)
3. Ethics application and approval (April 2016)
4. Recruitment of participants (April-May 2016)
5. Dissemination of interview/survey questionnaire (May 2016)
6. Collection of responses (May-June 2016)
7. Analysis of responses, synthesis of literature review and survey results (June 2016)
8. Documentation of results (June 2016)

As the first step of the research, a review of relevant literature on the topics of open data, public service co-production and co-creation, public sector innovation, data-driven decision-making and other related topics was carried out. The aim of the literature review was to understand the way in which these concepts have been defined and addressed in literature, and develop working definitions of the concepts for the purposes of the OpenGovIntelligence project. The review included academic articles from databases such as Scopus/Elsevier, SpringerLINK as well as Google Scholar and to some extent also "grey" literature (relevant working papers and European Commission reports).

Most of the publications found were from the year 1995-2016. They were further narrowed down considering the focus of the OpenGovIntelligence project and direct relevance to the project. Altogether 91 academic and policy reports were found relevant and were synthesised. In addition, the conclusions of an extensive systematic literature review of public sector innovation by De Vries, Bekkers and Tummers (released in March 2016) were integrated into the review. They screened around 10,000 studies and carried out an in-depth analysis of 181 studies.

In parallel with the literature review, a written interview questionnaire was developed to investigate the core research questions in more depth, focusing more specifically on the context of the countries where the OpenGovIntelligence pilots will be carried out. The interview questions concerned the

respondents' understanding and views of issues related to open data, data-driven co-creation and innovation in the public sector; their awareness of open data and data-driven service innovation; the capacities and needs of their organisations concerning opening up data and engaging in service innovation; the potential and challenges of using open data in public sector innovation and co-creation; the main drivers and barriers of data-driven public sector innovation and co-creation; possible solutions (policy initiatives) for overcoming barriers and stimulating open data-driven innovation; relevant existing initiatives and good practices related to using open data for service co-creation, etc. The questionnaire comprised 11 questions in total, most of which were open-ended. No sensitive personal data was collected.

The respondents were hand-picked in consultation with the project partners in the pilot countries to involve the most relevant stakeholders in each country. In each pilot country, two main categories of stakeholders were invited to take part in the study:

- Representatives of public administration (including but not limited to the organisations directly involved in the pilots). This category included people holding relevant political and administrative positions, representing different levels of government (national, regional and local/municipal).
- Stakeholders external to public administrations. This category involved private companies, civil society organisations, universities and research organisations, and other non-governmental actors who had some expertise or experience in open data, service co-creation, public sector innovation or other domains relevant for the research.

The inclusion of the views of various stakeholder groups was aimed at in order to develop a better understanding of the different capacities, needs and expectations that affect open data-driven co-creation/co-production. Invitation to the survey was sent to *120 stakeholders from 6 countries*.

The respondents could either complete a Google Forms-based online survey or respond to the questions by e-mail. All respondents were also asked to fulfill an informed consent form to indicate their agreement to the process of data collection, analysis and storage.

The data was then analysed and processed by the Tallinn University of Technology in accordance with the Estonian Personal Data Protection Act and the university Rules on the Processing and Protection of Personal Data. For the purposes of the analysis the responses were anonymized, i.e. each participant was assigned a unique identification number not linked to any personally identifiable data. The results of the study are summarized in Section 4 of this deliverable.

## **3.2 Technical Challenges & Needs of LOSD**

### **3.2.1 Linked Open Statistical Data: State-of-the-Art**

During the last years a growing body of literature studies LOSD. The objective of this step of the methodology is to accumulate this body of knowledge and provide an analysis of the search results in the area so far. This will enable the identification of (a) the state-of-the-art in the area, and (b) open or unexamined research questions. Towards this end, we systematically reviewed the scientific literature to identify relevant studies. This activity is included in T1.3.

More specifically, in order to achieve our objective, we employ a method for conducting systematic literature review (Webster & Watson, 2003). Initially, we performed a systematic search in order to accumulate a relatively complete body of relevant scientific literature. Towards this end, we started with Google Scholar using the key words “linked data” AND “data cube” and we collected an initial pool of articles. Thereafter, we went backward by reviewing citations in the identified articles and forward by using Google Scholar’s functionality to identify articles citing the previously identified articles. We thereafter studied and filtered these initially identified articles in order to come up with the final set that was included in our research.

This approach resulted in a set of 138 papers that have been either published in scientific journals or presented in scientific conferences and workshops.

In order to synthesize the accumulated knowledge, we performed a concept-centric analysis. The main characteristics of the area were extracted and a conceptual framework for linked data cubes was created in order to structure the area. Finally, the framework was employed to categorise and further analyse the literature and to extract insights into the area of linked data cubes.

The result of this step of the methodology is documented in Section 5.1.

### 3.2.2 Challenges & Needs regarding LOSD Interoperability

The objective of this step of the methodology is to identify the *challenges of LOSD exploitation* and translate them to *needs of LOSD publishing*. The overall objective can be broken down to:

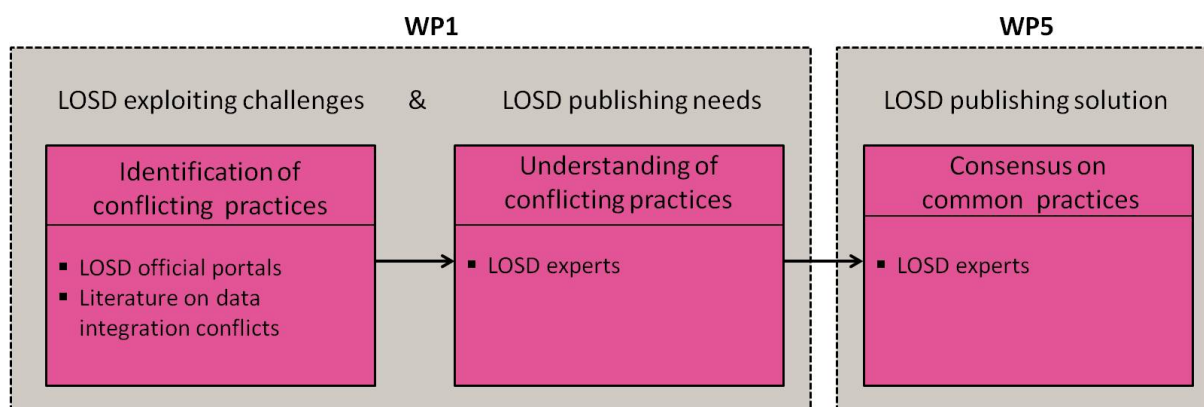
- Identify the challenges for exploiting LOSD from diverse portals
- Define the publishing needs based on the identified LOSD exploiting challenges
- Propose common practices to address the LOSD exploiting challenges and the corresponding LOSD publishing needs

During the last years, an increasing number of governments, public authorities and companies have started to publish linked statistical data using the RDF data cube (QB) vocabulary. However, in order to unleash the full potential of linked statistical data it is desirable to combine related data from different sources. This will enable, for example, the performing of combined analytics on top of multiple data cubes published by different authorities.

At the moment, many official portals launched by governmental organizations (owning the data) are using the QB vocabulary to publish their data as linked data cubes. For example, the Scottish Government, the UK Department for Communities and Local Government (DCLG), the Italian National Institute of Statistics (ISTAT), the Flemish Government, the Irish Central Statistics Office and the European Commission's Digital Agenda have published data using the QB vocabulary. Although all the above portals use the same vocabulary, they often adopt different practices, thus hampering their interoperability. The result is the creation of cubes that cannot be exploited together despite the use of linked data technologies.

In order to achieve the objectives of this study we adopt a three step methodology:

- **Identification of conflicting practices (WP1).** Identify the conflicting practices used by the existing portals. For this reason, we: i) detect official LOSD portals launched by governmental organizations, ii) conduct a state of the art analysis on data integration to identify the types of conflicts that might occur at LOSD, iii) based on the 2 previous steps, identify the conflicting practices that occur at the LOSD portals.
- **Understanding of conflicting practices (WP1).** Actively involve the LOSD experts (e.g. curators of linked data cube portals and publishers of data cubes) in order to understand the conflicting practices. The experts' involvement is achieved through a questionnaire where they describe the advantages, disadvantages and peculiarities of all the conflicting practices. The result of this step is the LOSD publishing needs.
- **Consensus on common practices (WP5).** Conclude at a set of common publishing practices in order to address the LOSD exploiting challenges. LOSD experts are involved in order to come up with a consensus on the practices. This step is still a work in progress and the final results will be presented at WP5 in the report of standardization efforts. However, some preliminary results are also presented at this deliverable.



**Figure 4 The steps for eliciting challenges and needs on LOSD exploitation**

The results of this step are documented in Section 5.2 of this deliverable.

### 3.3 Users' challenges regarding the exploitation of Open Statistical Data

In order to identify challenges related to the development of data-driven products by exploiting open statistical data we employed two methods:

- An in depth analysis of data.gov.uk in order to discover datasets about a real world problem.
- Questionnaires for developers that have used open statistical data to develop software products or applications.

#### 3.3.1 Open Statistical Data Fragmentation

For the first method, we conducted an in depth analysis of data.gov.uk in order to discover all the available datasets that are related to unemployment. In particular, the steps of the analysis are the following:

- First step is to find datasets about unemployment, so we searched for the keyword “unemployment” to data.gov.uk.
- Search results contain links directing to other portals. We documented these portals. Also all the links driving users from data.gov.uk to these portals are counted.
- Last step is to check the datasets, to which users have access through data.gov.uk, about relevance with the topic and their resource format.

### 3.3.2 Challenges & needs of software developers

For the second method, a questionnaire was designed in two topics (a) general questions about open data (b) specific questions about open statistical data. The questionnaire aimed at developers who created applications based on open statistical data such as participants in Open Data Hackathons and startups. Open questions with free text responses and closed questions with multiple choice responses were used. The questionnaire’s vocabulary was adapted to people who have technical skills in data-driven products and in data editing. After receiving the answers, visualizations were made through Excel in order to provide conclusions. The questionnaire can be found at Appendix C.

## 3.4 Pilots’ needs on data-driven innovation

The objective of this sub-step is to identify needs of users that aim to exploit open statistical data for the development of data-driven products. Pilots partners of the project from 6 different countries were picked in order to help identify these needs through interviews in which they participated:

- The Greek Ministry of Interior
- Enterprise Lithuania – Lithuanian Ministry of Economy
- Trafford Council
- The Flemish Government
- The Marine Institute
- The Estonian Ministry of Economics

An interview form was designed in the frame of this section that aim to identify needs of users that aim to exploit open statistical data. The form has three parts:

- **Problem Description.** A problem was asked to be described by pilot partners of the project which problem can be solved through the exploitation of statistical data. Questions covered many different areas such as problems of public administration, businesses and citizens.
- **Dataset Description.** Datasets that can solve the problems mentioned and their specific characteristics were asked to be described. Open questions like free texting were used while form’s vocabulary was adapted to people who have theoretical and technical knowledge of exploitation of statistical open data.
- **Final Product/Service Description.** The last component section of the interview form was about the final product/service. Descriptions of target users and characteristics of final product were asked in free text questions.

The interview form can be found at Appendix D.

### 3.5 Data Infrastructures

The Task 1 State-of-the-art of data infrastructure has the objective of reviewing the scientific literature, practitioner reports and projects regarding the data infrastructures of Open Data. The State-of-Art is the tool that will help to partially answer the following OpenGovIntelligence project questions presented in D1.1 and helped by other three tasks with different subjects and objectives also included in D1.1:

- How are existing data infrastructure currently deployed within the Public Sector?
- How are existing data infrastructures repurposed in the pursuit of innovation and creativity, following a public-private collaborative approach?

From these research-questions a research approach was established for the creation of the State-of-Art. The research/scientific papers used on this State-of-Art report were selected via a systematic literature review adapted from (Petticrew and Roberts 2008). Firstly, identification of keywords of the topics were found during the search in the database:

- 1) Open Data
- 2) Open Data Infrastructure
- 3) Open Government data
- 4) Open Government Data Infrastructures
- 5) Linked Data
- 6) Linked Open Data
- 7) Tool and Skill for Statistical Capacity Building Evaluation

The academic databases used for the literature searches of this document were the Electronic Government Reference Library (EGRL), the IEEE Xplore, the SCOPUS (Elsevier), the ACM Digital Library and Google Scholar that served to complement the results from the aforementioned library search. The EGRL gave us the wide range of literatures in the E-Government area. The IEEE and ACM libraries helped us to cover the technical perspectives. Scopus and Google Scholar provided us with an overview of related works and helped us to prioritize them according to the most cited literatures.

Moreover, there was a full reading of the abstracts and keywords in order to check their adherence and relevance to the area of study. Some papers were not related to the scope and objective of the OpenGovIntelligence project and therefore were excluded from the literature review.

Lastly, an organisation of papers was conducted, using the reference manager Mendeley, (<https://www.mendeley.com/>) and stored at the OpenGovIntelligence project folder on Google Drive cloud. These processes enabled the creation of a spreadsheet that controls the readings and the status of the articles found on the Google Drive (<https://goo.gl/tOjA84>), in order to give agility to the OpenGovIntelligence project and avoid any double reading between the consortium.



## 4 Public Administration's challenges & needs regarding data-driven innovation

As the first step of the study on service co-production, a literature review was conducted to delineate some of the central concepts of OpenGovIntelligence, such as public service and public value, data-driven public service innovation, co-production and co-creation. Based on literature, working definitions of these key concepts are proposed in the next section. This is followed by a discussion of the challenges and needs related to open data-driven public sector innovation and public service co-creation, based on literature and the results of the survey conducted in the framework of the project.

### 4.1 Key Concepts

**Public sector** – comprises the general government sector plus all public corporations including the central bank (OECD 1997).

**Public services** – services where government has a responsibility for the provision. Public services are those services which public bodies, such as central or local government, either provide themselves or commission others to provide. Public services may be produced and provided outside the actual public administration, by the private sector or by citizens. (OECD 2011)

**Public value** – various benefits for society that may vary according to the perspective or the actors, including the following: 1) goods or services that satisfy the desires of citizens and clients; 2) production choices that meet citizen expectations of justice, fairness, efficiency and effectiveness; 3) properly ordered and productive public institutions that reflect citizens' desires and preferences; 4) fairness and efficiency of distribution; 5) legitimate use of resource to accomplish public purposes; and 6) innovation and adaptability to changing preferences and demands (OECD 2014).

**Public sector innovation** – “the creation and implementation of new processes, products, services and methods of delivery which result in significant improvements in outcomes efficiency, effectiveness or quality” (Albury 2005).

**Private sector innovation** – “the implementation of a new or significantly improved product (good or service), or process, a new marketing method, or a new organisational method in business practices, workplace organisation or external relations” in firms (OECD 2005).

**Co-creation** – active involvement of end users in various stages of public policy and public service production, usually as co-initiators, co-designers or co-implementers of a policy or service. Co-creation can have different objectives, such as increasing the effectiveness or efficiency of policies and services, gaining customer satisfaction, increasing citizen involvement, etc. (Voorberg et al. 2014). In the context of public service provision, co-creation could be defined as the active involvement of end users (usually citizens or businesses) in the initiation, design or delivery of a public service.

**Co-production** – in literature, co-creation is often used interchangeably with the concept of co-production to indicate the active involvement of end users in various stages of public policy and public service production. According to Voorberg et al. (2014), the main difference between the two concepts seems to be that co-creation literature puts more emphasis on co-creation as a value. Hence,

the adoption of the concept of co-creation rather than co-production is suggested for the next steps of the project.

Co-creation and co-production can be distinguished from the concept of participation by their more specific focus on active involvement, while participation could also refer to passive involvement (Voorberg et al., 2014). Therefore, it should be emphasized that service co-creation presumes a two-way interaction between government and citizens or bottom-up interaction from citizens to government in which citizens (or other types of service users) have an active role of a participant or initiator, not just a passive receiver of information from the government.

**Data-driven** – at the core of this concept is the notion of being evidence-based and is often used in connection with journalism or learning. In the context of (e-) public service delivery it can be understood as a web service based on a database (Deutsch et.al., 2004) involving all levels of data processing, from information provision to transactions.

## 4.2 Literature review

In the context of the study, it is important to acknowledge that information and communication technologies (ICTs) hold an innovative potential both in terms of giving an impetus for novel public services and providing the means for transforming governance processes. Among other uses, ICTs enable the development of smarter and more collaborative methods of decision-making. This could lead to improved decision-making and service provision in the public sector, in particular through harnessing information, knowledge, skills and perspectives that have traditionally been external or unavailable to public sector organizations. Thus, ICTs enable the aggregation of knowledge through direct interaction with citizens, businesses, interest groups and public sector organizations, using methods such as e-consultations, wikis and crowdsourcing platforms, e-petitions or discussions in social media. Therefore, ICT-enabled co-creation has the potential to produce effective, high-quality public services and increase the perceived legitimacy of the decisions taken on behalf of the public. ICT applications also enable the collection, analysis and combination of vast amounts of public and private data such as open data, big data, linked data, or data crowdsourced directly from citizens and service users, which can provide governments hard evidence to back up policy decisions and indicate the aspects in which public services can be improved.

In the context of public sector ICT projects, the issue of success/failure and drivers/barriers has received considerable attention in academic literature, evolving into a research stream in its own right. Despite varying definitions in literature of what counts as a success or failure, there is a shared understanding that the failure rate of ICT projects continues to be globally high (Thomas and Marath 2013). This comes at a high price in terms of wasted resources, missed opportunities, unrealized benefits and damage to reputation (Heeks 2003, Sauer 1993). Therefore, much of information system (IS) management literature is devoted to researching the factors that affect IS success and failure. The issue has predominantly been approached from a rationalist angle, focusing on critical success/failure factors, which purport to predict the outcomes of ICT projects (Kautz and Cecez-Kecmanovic 2013). Although failure is a common problem in information systems projects both in the private and public sector, the issue has received much less attention in the context of public administration (Dwivedi et al. 2013). Several authors (Rochet et al. 2012, Dwivedi et al. 2013) emphasize the inherent complexity

of public sector ICT projects, owing to the environmental constraints specific to the public sector and the wide range of stakeholders involved. Therefore, the challenges seem to be especially complex and stakes particularly high in the public sector. This implies the need to consider the broader context of e-government in addition to information system success/failure factors.

This all has become very topical in public sector innovation research, concerned with “the creation and implementation of new processes, products, services and methods of delivery which result in significant improvements in outcomes efficiency, effectiveness or quality” (Albury 2005). From 2000s, the literature has rapidly grown on public sector innovation and on related changes in governance (see, e.g., Hartley 2005, Verhoest et al. 2006, Moore and Hartley 2008, Pollitt and Bouckaert, 2011).

One of the most recent systematic and detailed accounts of public sector innovation is a literature review (of 181 articles and books) by De Vries, Bekkers and Tummers (2016). It adopts the umbrella concept of “antecedents” to denote influential factors in public sector innovation at different levels. According to their definition, an antecedent can act either as a driver or barrier to innovation depending on the context and level of analysis. A similar idea has also been expressed by several other authors, e.g. Bekkers et al. (2013) and Nasi et al. (2015), who argue that it is often the specific context that determines whether a factor acts as a driver or barrier to innovation. De Vries et al. (2016) divide the antecedents into four main categories: 1) environmental level (the context external to public sector organizations); 2) organizational level (the structural and cultural features of an organization); 3) innovation level (intrinsic attributes of an innovation); and 4) individual level (characteristics of individuals who innovate).

In a similar vein, the main barriers highlighted in the European Commission’s report “Powering European Public Sector Innovation” (2013), which is perhaps the most influential policy document on the topic, can be categorized into environmental, organisational and individual-level factors. These include: (1) weak enabling factors or unfavourable framework conditions: scattered competences, ineffective governance mechanisms, diverse legal and administrative cultures, resource constraints to develop and deploy staff and to finance rollout, and inadequate coordination within and across organisations to share, spread and scale up successful initiatives; (2) lack of leadership at all levels: preference for caution and failure-avoidance to creativity (finding new paths to success), rigid rules and risk-averse managers who discourage staff and stifle the diffusion of innovative ideas; (3) limited knowledge and application of innovation processes and methods: an often absent access to capabilities (systems, skills, tools and methods), lack of collaboration (with other parts and levels of government, businesses, citizens and third sector organisations); (4) insufficiently precise and systematic use of measurement and data: inadequate information on sources of new and improved products, processes and services; lack of monitoring of the benefits for policy outcomes.

These approaches find support in numerous other studies, which conclude that innovation within the public sector is driven and constrained by multiple factors that can be related to the individuals involved in innovation as well as organisational and broader environmental context. Some of the frequently mentioned barriers include a lack of political and administrative triggers, legal culture of the public sector (factors such as standardization and formalization (e.g., Sørensen and Torfing 2011) and rule-driven ‘path dependencies’ (e.g., Bernier and Hafsi 2007), state, governance and civil service traditions and lack of necessary capacities (Pollitt and Bouckaert 2011), and lack of resources and

resource dependency within organizations and networks (Bekkers et al. 2013). At the same time, it is noteworthy that the same variable (e.g. limited resources) can act as a barrier as well as driver depending on the circumstances.

If we look more specifically at service co-production related literature, then most of the barriers are similar to any kind of ICT-led innovations in public sector. However, some barriers can be considered more specific to service co-production. These often include a lack of administrative and political championing, poor integration into organizational procedures and broader political processes, lack of easily demonstrable impact, unfavourable cultural context, hostile attitudes to citizen engagement, and the difficulty of matching different expectations and capabilities in designing systems intended to engage diverse user groups.

If we look more specifically at literature related to open data, further specific conclusions can be drawn. There are studies, for example, on the drawbacks of an arguably over simplistic view among many practitioners and scientists on the benefits and limitations of open data (Janssen et al. 2012). Various risks are identified and are related to the political, economic, institutional, legal and technological aspects, confirming the need to take an interdisciplinary look and addressing various levels and aspects.

One of the key issues identified is opposition from government agents themselves to publish data, the unpredictable nature of government support in the sphere, and lack of political communication between providers and re-users of open data (Martin et al. 2013, Barry and Bannister 2013). This is reinforced further by lack of knowledge on open data and open data based governance (for limited understanding of concept by policymakers see Janssen, 2011) and by the fact that in many cases shift to open data challenges existing organizational procedures and routines, while carrying risks (of failure). Misinterpretation of open data and open government concept among many public administration practitioners has been identified as clear barrier (Yu and Robinson 2012), including overly simplistic view on open data as a tool (Conradie and Choenni 2012) to promote transparency of government among many practitioners and academics (Janssen et al. 2012).

Low priority given to open data based public sector innovation by politicians can lead to legislative barriers. This is related both to the lack or ambiguity of regulatory basis of the open data-driven projects, challenging the flow of datasets from government agencies to other actors and in inconsistency of the policies and activities in the sphere (Ganapati and Reddick 2012). Even if legislative barriers do not exist, civil servants might expect further guidance in the form of strategies. Thus, one of the fundamental barriers identified relating to open data based public sector innovation is a lack of strategies on how to foster the re-use of open data by third parties, i.e. businesses and citizens as end-users (Veenstra and Broek 2013).

Also, information security and data protection issues are raised, especially in finding the balance between public and private information (Huijboom and Broek 2011), confidence in open data (O'Hara 2012) and possible abuse of open data.

Increasingly, barriers from user perspective are emphasized by various researchers. These include lack of user perspective vision in the government open data policies and barriers associated with the use

of data by the end-users, i.e. citizens and businesses such as the access to data and information justice (Johnson 2014), usability, misinformation, unfriendly interfaces, etc. (Zuiderwijk et al 2012).

### 4.3 Survey results

Our online survey on open data-driven co-creation and innovation yielded 63 responses from all six pilot countries, including 34 public administration representatives and 29 business, civil society and research actors. Participation activity was the largest in Greece (16 responses submitted), followed by Belgium, Ireland and the UK (10 participants from each country), Estonia (9 responses) and Lithuania (8 responses). 22 respondents represented central or federal government, 7 respondents represented regional government and 4 respondents the local government level. 15 respondents were from private companies, 7 represented non-governmental and civil society organisations and 8 represented universities and other research institutions. The following section gives a brief overview of the answers given to each survey question.

#### 4.3.1 Existing experience with open data and co-creation

Participants were first asked whether and how they had previously used open data in their work. While 5 respondents had not used nor produced open data in their organisations at all, others reported using open data mainly in internal organisational work processes such as data analytics, prediction, decision support and monitoring. Many have published open data for others to use, several also support public sector organisations in opening up data or provide tools and support to external users for re-using open data. Some organisations use open data in a variety of ways, for example:

*“We promote open knowledge and open data in Belgium, not only through events, but also by re-using open data. It goes from opening up data by scraping transport data (irail.be), to adding new data in OpenStreetMap (OSM.be) to using it in educational tools (datawijs.be) to re-opening consolidated legislation (<http://belaws.be/>), to using it during our Summer of code project.”*

*“We have used open data to access the areas that most need our focus, areas that have high levels of issues that we need to tackle or areas that suffer from elements of deprivation. We have used the information to make judgements on how to invest our time, money and efforts to their best use. We used the source to benefit the community and to offer a connectivity that might never have been visually able to see.”*

We also asked more specifically about the respondents’ experience with using open data for service co-creation or engaging in the co-creation of public services initiated by other organisations. As it turned out, 70% had taken part in some form of co-creation of public services based on the use of open data, while 30% had no such experience. Interestingly, all respondents from Belgium reported having participated or initiated the co-creation of data-driven services, which was not the case in other countries.

The reasons why different organisations had engaged in service co-creation were very different. Several mentioned direct benefits for the organisation, such as increased efficiency, data quality,

reaching a wider audience and increasing the impact of the organisation's work. Others referred to more general benefits for customers and society at large, such as enhanced transparency, building services rooted in user needs and increasing service quality for customers, promoting public participation, driving digital innovation and stimulating economic growth:

*"In order to serve the community needs (Citizens, Customers, Business, Public Sector) in a most efficient and less bureaucratic way."*

*"We are convinced that open data is a major driver for innovation and enabler of digital transformation in public administrations."*

*"To promote the idea and mind-set of eParticipation and thus, promote also Open and Participatory Governance. To actually empower citizens and motivate them to be actively involved in local issues."*

A number of respondents also mentioned legal obligation to publish data as a factor driving their efforts to make data open and build services on top of open data:

*"That's the law. And it provides value to the citizens. And we as a regulator must set an example."*

*"The ministry has the main responsibility on open access and re-use of public sector documents, information and data."*

*"Primarily by being required to by EU directives but also due to the beliefs and opinions of dedicated individuals."*

For some organisations, open data is an inherent part of their daily work and organisational strategy:

*"Because it's our mission, and the reason we were set up: to realise the open web of data. We try to demonstrate the value of openness, and open data, in everything we do."*

For organisations that had not participated in open data-enabled service co-creation, the reasons were mostly related to open data and service provision being out of the scope of their organisation's work:

*"As an NGO by definition we do not involve ourselves in the creation of public services."*

*"Because our organisation doesn't provide public services in which open data could be used."*

Some also mentioned lack of priority, lack of perceived benefits or lack of funding as reasons why they have refrained from such initiatives:

*"Not yet found an advantage of such an initiation that would contribute to the fulfilment of the Authority's mission."*

*“Being private organisation the cost is major issue in partaking the co-creation of the public service using open data.”*

Finally, participants were asked to express their views on the current level of involvement of businesses and citizens in the co-creation of open data-driven public services in their country. Interestingly, although the question was posed as an open question, the majority of answers across countries and sectors could be summarized in one word: “low”. However, there are slight variations in different countries. For example, most respondents in Belgium expressed the opinion that while stakeholder involvement is still low, it is also clearly on the rise:

*“More can always be done, but both the local and regional governments take several initiatives already to engage stakeholders at events, workshops and so on. Generally, I think Belgium has made big strides in the last two years.”*

However, some also suggested that participation is mostly limited to a small community of activists, while the general public remains passive or uninvolved:

*“... to my knowledge, there are only a handful citizens / businesses / organisations into data-driven co-creation. But those who are, tend to be very active and very knowledgeable.”*

At the same time, the Estonian respondents pointed to the unavailability of open data and government support as a possible factor impeding more active participation:

*“There are not enough data made available as open data yet in Estonia. That withholds also the innovative thinking and participation of citizens.”*

While the vast majority of respondents from Greece described stakeholder involvement as “low” or “poor”, some also referred to a developing culture in the field of open data and data-driven co-creation, for example:

*“A culture of open data is being created in Greece during the last years, but that culture remains retained inside a few businesses and organizations.”*

Quite similarly, Irish participants mostly characterized the current level of involvement as low. However, a few different positions stood out. For example, one participant seemed to be satisfied with the way citizens and businesses are currently involved in producing open data-driven services:

*“Citizens and Businesses are represented in the Open Data Governance Board and all our public consultation processes are open to all.”*

Another respondent suggested that although involvement is currently low, it should not be considered a problem in itself, adding:

*“I believe we need to see open data being considered as part of national data infrastructures, where the primary target is to get thematic/expert users making use of open data for evidence based decision making. In other words, open data*

*should be a key part of public sector data infrastructure efforts, with any business or citizen engagement as an offshoot of that process"*

Even in the case of the UK, which according to our overall observations from the survey seems to be comparatively advanced in terms of open data and data-driven co-creation, most participants seemed to agree there is room for improvement, in particular as regards engaging a wider community of citizens and businesses:

*"Poor but improving. Great central government support to open data. Little support to act upon it consistently."*

*"From my experience there are isolated pockets of interested parties that wish to engage with open data."*

*"Business involvement tends to be large international corporates not smaller local organisations."*

#### **4.3.2 Barriers and drivers of open data-driven public service co-creation**

Another set of questions concerned the barriers that the respondents view as adversely affecting the use of open data for the co-creation of public services, as well as the drivers that enable and boost open data-driven service innovation. The survey revealed that respondents see a number of barriers that hinder the use of open data for the co-creation of services, which concern very different levels – from the technology itself to stakeholder perceptions about open data and co-creation. Very broadly, the barriers could be divided into two main categories: 1) technology/ICT-related factors; and 2) contextual factors. The latter, in turn, can be further categorized into groups of barriers related to a) administrative and organisational context, including existing administrative and organisational processes and procedures, resources and culture; b) legal context, and c) impediments related to the stakeholders involved in the provision or use of open data or in the co-creation of innovative data-driven services.

Some of the frequently mentioned technology-related factors include the unavailability of open data, lack of data quality, lack of data coherence (different standards and methods for collecting data in different jurisdictions), technological immaturity, under-developed open data infrastructures and lack of uniform standards. In one case, the lack of availability and documentation of APIs was mentioned as a barrier, another respondent made an interesting case for the difficulties related to re-using data:

*"Often it is difficult to use open data for an innovative purpose that wasn't anticipated by the collector/ publisher – e.g. because the schema only has a limited scope for re-interpretation."*

A number of barriers emerge from the legal, cultural and organisational context. In many cases a broader use of open data seems to be hindered by existing proprietary business models that are based on selling key data. This is highly related to the issue of resources – by making data open, many organisations would lose an important source of revenue. Moreover, even if no business models are involved, opening and publishing data is associated with high costs:



*“Lack of resources to produce and publish linked open data and create public services based on these data”*

At the same time, these costs are often seen to exceed the potential benefits:

*“Resources need to be invested yet return on investment not always clear.”*

Several respondents mentioned legal issues as a barrier, in particular existing legislation related to sharing and licences and the related privacy and security concerns. In addition to legislation, many of the barriers mentioned have to do with the perceived incompatibility of existing administrative procedures and organisational processes with collaboration, co-creation and involvement of stakeholders:

*“Re-users have a hard time letting governments know about what data is lacking and governments have no or little insights in what the re-user is searching. A decent feedback loop is needed.”*

*“It changes the relationship of the public service (government) with the public, making the creation of new processes that take into account feedback from citizens necessary. Public organizations are not used to reacting to public input.”*

A particular bundle of barriers can be related to the different stakeholders around open data, such as data providers and users, policy-makers, public services providers and (potential) co-creators. According to the survey, there is still very little awareness of the existence of open data and its potential among all groups of stakeholders. Many respondents also referred to the lack of necessary skills to open up data and make use of open data in innovative ways. A crucial barrier seems to be the lack of perceived usefulness and business value of open data, lack of political priority and support, administrators' lack of openness to open data and co-creation and fears of data misuse:

*“Lack of awareness of the benefits by both users and the public service providers”*

*“There is still a lot of concern about the usefulness of open data.”*

*“Our administrations are driven by politicians that don't see the long term value of open data”*

*“corporate culture is not compatible with sharing and openness”*

*“failure to recognise the value, lack of relevant responsibilities, lack of funding possibilities or specific know-how”*

Other barriers that were mentioned less frequently but still seem to pose problems for some countries include the perceived lack of demand for open data (this was mentioned in particular by Estonian respondents), lack of initiative and ideas, time constraints to participation, lack of supporting policies and strategies, lack of legal responsibility to publish open data, lack of governance of open data programmes and lack of assurance that data will remain open in the long term.

In many ways, the factors that are seen as important drivers of open data-driven (public service) innovation are the barriers in reverse. For example, an important set of drivers is associated with data

and technology, including the availability of open data to start with, the existence of easy-to-use, well-maintained, well-documented, high-quality datasets, the provision of easy-to-adopt standards for data formats and data exchange, prototyping and the existence of concrete applications to showcase open data solutions and interactive data visualisation.

Other key drivers can be associated with various contextual factors, from the legal, policy and administrative environment to stakeholders and innovators. The existence of enabling legislation, policies, strategies and policy measures was considered important by the majority of respondents. This includes open data legislation, statutory obligations to publish open data, open standards policy, open data action plans, European open data policy (in particular the Directive on the re-use of public sector information, also known as the “PSI directive”). Some also mentioned the importance of a broader openness and transparency agenda in addition to open data policies and benchmarks with other countries as a measure to foster open data policies. Several saw the dissemination of best practices and real use cases as a way to drive further innovation. Not surprisingly, in addition to the existence of favourable policies, the availability of funding was seen as a key driver by many.

Another important set of enabling factors was, again, associated with the stakeholders involved in data and public service provision, co-creation and the use of data and services. A major precondition for any open data innovation seems to be that different kinds of stakeholders perceive open data as valuable and beneficial in the first place. Some of the key benefits that were mentioned as driving the use of open data include the perceived ability of open data to support administrative efficiency and automation of organisational processes, better information and evidence-based policymaking, better products and services, possibility to answer real needs, public value, cultural and societal potential, transparency, more participatory and open governance. Open data is also seen to create economic opportunities, enable the creation of cheaper and simpler web applications and private sector-driven commercialized solutions that stimulate digital economy.

In fact, stakeholders’ beliefs, priorities, preferences and actions were among the most frequently mentioned drivers for open data-driven innovation. Visionary policy-makers and administrators and their personal enthusiasm and ambition seem to be considered the key force driving the exploitation of open data, regardless of the country context. In relation to that, political priority, senior buy-in and in-house innovation champions were mentioned. A clear demand from the private sector and broader community were also considered important, as well as the existence of a wider open data movement and the presence of knowledge and skills for actually working with open data among all stakeholders.

#### **4.3.3 Needs and expectations for open data-driven innovation**

Participants were also asked about the main needs of their organisations and the capacities they feel would need to be developed with regard to engaging in open data-driven innovation. Although a few respondents claimed to have all the capacities they need, most mentioned several unmet needs. The ones mentioned the most frequently include the availability of open, high-quality, up-to-date and interoperable data, more coherent datasets and more providers of the same type of data, which are seen as important enablers for the re-use of data to trigger innovation. This should be supported by a full understanding of the benefits of open data, knowledge and data literacy within the organisations, as well as skilled manpower, in particular data analysts (however, it was also stated that the availability

of easily usable data reduces the need for qualified specialists). Time and money for innovative activities were also seen as lacking in many organisations. In addition to resources, respondents also mentioned the need for a supportive organisational culture and innovative ideas, backed up by senior management and staff commitment and capable change management. Some also saw a need for a more holistic approach to the use of open data, such as making open data part of the general digital policy and core tasks of public organisations and adopting clear strategies for the exploitation of open data. With regard to the external context, collaboration between public authorities, demand and support from stakeholders and capacity-building were perceived as missing factors. The latter, as emphasized by many, should involve concrete examples and instructions for extracting, opening and exploiting data.

Regarding the participants' expectations to the ways in which public sector organisations could use open data in decision-making, a number of possible areas were mentioned in which governments could make better use of open data. These areas could broadly be divided into two areas: 1) data-driven and evidence-based decision-making, and 2) stakeholder participation and government transparency. The general view seemed to be that open data can enhance the efficiency and effectiveness of public decisions, help root public decisions in real needs, and increase transparency and trust in public institutions. According to the respondents, data-driven decision-making can be supported by data analytics, visualization tools, KPIs and dashboards. Open data was also considered useful for strategic development, risk assessment, impact assessment and estimation of public acceptance of government decisions and policies. Governments were also suggested to involve third parties (including citizens, businesses, other stakeholders) in data aggregation to generate new insights. For citizen and stakeholder engagement, methods such as crowdsourcing could be used to improve public services. Citizens' participation could also be used to identify problem areas by stakeholders. On the other hand, the transparency of decision-making processes could be advanced by publishing data that decisions are based on in the form of open data or linked open data. Respondents also suggested re-using data from other sources such as research organisations, public sector organisations and social media as input to decision-making. Finally, respondents made suggestions for facilitating the use of data in decision-making – among other ways, this could be done by sharing open data and good practices with other public administrations, as well as gather inspiration from grassroots open data initiatives.

One participant also expressed the opinion that the crucial question was not in whether governments use open or closed data to make decisions but in how they use it:

*"It's the processes that are taken to make the decision. Hence, sharing and opening the decision making process is a step that should be taken - "open data" is a red herring in this context."*

#### **4.3.4 Policy solutions: successful and unsuccessful examples**

In identifying successful policies or initiatives that have been used to promote open data-driven innovation in public services and decision-making, the following conclusions can be drawn on the basis of responses. The respondents were also asked for the reasons these measures were successful.

Many respondents emphasized the importance of setting proper legal framework on strategies.

*“The “open by default” policy of our new government has been doing great work. All of the municipalities are now realising they'll have to open up, and see the benefit of it (or sometimes need examples from front-runners).”*

And, the importance to have strategic, comprehensive, systematic and well integrated with current state of the art as well as future trends in information technology open data policy.

*“In my opinion, UK has a number of successful policies used to promote open data. UK government issued several important documents and requirements for opening up data and for providing open data driven services. For example, see Open Data White Paper – Unleashing the Potential, HM UK Government (June 2012), Information principles for the UK public sector (30 April 2012), etc. I do not know any other country in Europe who has been so systematic in open data and also linked data strategy. By the way, it is important to notice that UK universities are well engaged into this initiative.”*

*“The most known initiatives in Greece regarding open data is the Transparency Project. It combined both regulatory actions and technical solutions to give birth to a highly estimated completely open system. The transparency project give birth to surrounding ecosystem of small applications that further advances the open data initiative and the related effects.”*

Although, in order to have it implemented, competent people are needed.

*“In almost all the cases, the main reason was the existence of people within a public organisation believing in the power and potential of open data. Believers are what make it succesful.”*

*“Vilnius Municipality hired a developer-open data evangelist to open municipality's data and Open Knowledge (OKFN) organisation.”*

And, concrete handbooks should be provided (e.g., <http://opendatahandbook.org/guide/en/what-is-open-data/#what-data-are-you-talking-about>) that explain open data to administrative local civil servants and to politicians.

The respondents argued if public services offering organisations are pushed to open data, this will be picked up by other stakeholders. The threshold for creating new services is low (data is free, structured, easily accessible) and that enables to create new services by everyone.

*“In most cases, successful open data products or services come about as a result of a specific (and often personal) frustration. Apps are being developed by a number of people who e.g. had it with untransparent train delays, or don't want to take a certain route when there's actually a better way to get from a to z.”*

*“Finland discontinued charging for digital maps in 2012. This resulted in increased use of spatial data by 50 times in three months; creation of first application within a month, recruitment of application creators by small enterprises, according to the estimation by Finnish ETLA, the sales of companies dealing with*

*map data will increase 15% faster than the sales of companies operating in other fields.”*

Hackathons were mentioned by many respondents (with references, e.g., to <http://appsforghent.be/>, OKFN, Open data hackathons organized by Startup Lithuania). These were considered important for ideation and awareness. Also, other public co-creation events were mentioned (e.g., <http://2016.summerofcode.be/>, IT4GOV, ODI)

*“IT4GOV is an initiative/competition organized by the Ministry of Interior in Greece in order to find innovative services that are not available on the market and that aim at structuring new alternative models for procedure simplification in public administration and effective eGovernment services which will provide innovative solutions for useable services to citizens and businesses. This kind of initiatives definitely helps to promote open data-driven innovation.”*

Finally, awareness-raising events, such as workshops, conferences, etc. were considered as well. For example, Open Belgium conference is about bringing together the data owners and re-users/open communities and open data day Flanders is more focused on the data owners and best practices. FIWARE /OASC/Creative Ring were also mentioned as initiatives stimulating open data/open services platforms and urban innovation, emphasising cross-hub/cross-country collaboration and interoperability.

Several respondents were convinced that such hackathons, public co-creation events and workshops, mostly driven by organisations outside the public administrations are important. They are driven by innovators, seeking collaboration, interaction and building bridges, and should be continued to fund, to enable them to mature, to scale, to generate tangible outcomes, to disrupt the existing and overcome resistance to change. Such initiatives provide good motivation for citizens’ to be engaged: crowdsourcing competition prizes, potential of forming a start-up or expand the research conducted in a particular field, consume useful information about everyday issues.

At the same time, several respondents were also very critical of hackathons, mainly as they

*“...tend to [be] one-off and tokenistic, [while being] useful for exploration or demonstration but not sustainable or commercially viable”*

in particular, when connected to

*“[...] big prizes (e.g. +10.000 euro). Those competitions are very distant, impersonal and will only attract ideas and applications from organisations that do not necessarily care about open data, [but] only about the possible prizes.”*

As the last question in the survey, we asked participants whether any promising policy instruments or initiatives were currently missing from the scene either at the national or EU level, which might be effective in advancing the use of open data for the co-creation of innovative services. We received a number of suggestions concerning different levels of government and different types of instruments, from broad strategies to specific tools.

At the EU level, respondents were generally satisfied with the existing policies, in particular the PSI directive. However, suggestions were made to give more attention to the day-to-day implementation of the directive and even consider an update to the directive to force member states to make all state information public free of charge. Respondents also frequently highlighted the need for data standardisation policies, which should be tackled at a cross-border rather than national level.

As to national policies and strategies, a number of respondents emphasized the importance of the open-by-default policy and the introduction of legislation to oblige governments to publish government data in open data formats, which is not yet the case in many member states. This obligation, according to several suggestions, should go hand in hand with providing a central free open data portal where local and national governments could publish their data. More specific suggestions for these kinds of central data portals included ensuring their ability to host data, sign-post to remote data, cache datasets to build resilience and offer on-the-fly data transformation across various formats or via various web services requests.

Other specific data-related policies that were seen to be missing at the national level concerned data infrastructure legislation for the maintenance and accessibility of high value data assets (e.g. geospatial data), the creation of a central data repository with API capabilities, implementation of the “API first” policy (i.e. the principle according to which governments should prioritize providing good APIs along with open data that everyone can use and develop new services upon). Others also found it potentially helpful to introduce a centralized semantic classificatory for easier linking between different databases and analytical decision-making applications that would utilize data across government agencies and administrative areas.

However, the realm of national government-level policies that could support open data-driven innovation is broader than that. For example, several suggestions were made to qualify public grant submissions and public tenders against open data or even oblige open data publication as part of public procurement and funding schemes. Several respondents also emphasized the need to review licencing and copyright policies – first of all in the sense of the modernization of copyright legislation in relation to public interest and new business models, and secondly encouraging a widespread adoption of free software licences with minimal restrictions and maximum compatibility, such as the MIT Licence.

In addition to broader policies, respondents frequently mentioned the need to offer data providers and users clear guidelines and optimal methods for providing and using open data (such as are currently being prepared by the Information Society Development Committee in Lithuania). These should be accompanied with efforts to publish and disseminate case studies and best practices as well as funding schemes and incentives to support the publication of open data in organisations. The adoption and funding of different forms of collaboration (cross-border, cross-sectoral, inter-organisational) were seen as an important measure to enable learning from others’ experience, adopt common methodologies, enhance cooperation between data producers and data users and even “co-creating a cross-border open data policy”.

Another broad set of measures which according to many respondents should be further strengthened concern building human capital and open data-related competences. For example, several

respondents found that public administration organisations need more professional data management and archiving capacities. Some suggested the establishment of a dedicated team of developers to focus on helping public institutions open their data, others proposed the creation of innovation teams around internal change-agents, who should be given sufficient freedom to experiment with open data in innovative ways. Systematic training programs and courses on open data, linked open data and general digital skills were suggested as an effective measure by many, which should not only be targeted to developers but also public servants and civil society.

#### 4.4 Discussion

Comparing the survey responses with literature on public sector innovation, co-creation and open data, it appears the survey results largely confirm the importance of the key barriers, drivers and needs that are frequently mentioned in literature. For example, both our literature review and the survey strongly suggest that the *lack of political and administrative priority and leadership* is a key barrier for any innovative undertaking in the public sector, including open data innovation. This seems to be closely related to the perceived *incompatibility of existing administrative and organisational cultures* with the idea of co-creation. Many public sector organisations still exhibit a hostile attitude to more open policy-making and service production and seem to lack understanding of the value that co-creation and user engagement could bring to existing processes.

What our survey confirmed is that a similar opposition is also present with regard to publishing open data. According to the survey as well as previous findings from literature, this tends to be associated with the *public administrators' lack of knowledge of the potential value and benefits* of open data. At the same time, the *costs of opening up data* and keeping it open in the long term are perceived to be high. Coupled with a *limited understanding of the processes and tools* which could be used for opening and linking data, the environment for innovative uses of data in public sector organisations does not yet seem to be ripe.

The opposition from public sector organisations is also possibly closely related to the complexity of organisational innovation and the *resistance of public administrators to fundamental changes* in their daily administrative procedures and business models. By opening up data, many organisations would lose an established source of revenue and for the moment there still seems to be very *limited understanding of the new business models* that could be built on open data. Literature suggests that the incompatibility of existing administrative procedures and resistance to organisational change is not only a problem for open data-driven innovation but a more general barrier to public sector innovation in general, particularly collaborative kinds of innovation that presume the involvement of non-governmental stakeholders in governance processes. However, as long as open data-driven service co-creation is not integrated into the daily work of organisations, there is high risk that such innovation will not become sustainable practice.

According to the survey, *the limited and unpredictable support of governments to opening up public data and assuring the availability* of open data in the long run also decreases the incentives for private companies and non-governmental organisations to build new services based on open data. At the same time, responses suggest that a lack of knowledge and skills related to open data is also

widespread among businesses and citizens. Therefore, some of the crucial barriers to open data-driven innovation emerge from a general *lack of awareness* across different stakeholder groups.

Somewhat differently from broader public sector innovation literature where technology seems to be less of a barrier than the context in which it is implemented, many of the barriers to open data innovation still seem to be related to the qualities of data and data infrastructures. Access to open data and its poor usability are often mentioned as barriers in open data literature. Our survey further reinforces the understanding that a large part of why open data-driven innovation has not taken off is simply due to the *lack of open data to begin with*. The unavailability of important data in open formats, the poor quality of data that has been opened, and lack of coherent and comparable datasets across different jurisdictions and organisations were the factors that were the most frequently cited as barriers by the survey participants.

At the same time, the key drivers for open data innovation (including LOSD innovation) are also very much associated with technological factors. According to the survey, this includes the availability of open data in high-quality and easily usable forms, which should be supported by the availability of concrete tools and applications to showcase and facilitate the use of open data.

The survey also suggests that technology needs to be supported by a favourable context. This includes legislation (in particular the legislative obligation to publish government data in (linked) open data formats for free) as well as specific policies and strategies to strengthen the implementation of legislation, such as an open standards policy to support the publication of open data, harmonization of standards across borders, and a broader open government agenda to support co-creation with non-governmental stakeholders. Naturally, funding mechanisms for making data open and engaging in open data innovation were also seen as an important driver.

One of the main source of innovation is seen in the stakeholders involved in different stages of the provision, design and use of data and services. Here, the perceived value of open data for individuals and organisations seems to be a key driver. Stakeholders' beliefs, priorities, preferences and actions were among the most frequently mentioned drivers for open data-driven innovation in our survey. In this context, it was considered particularly important that political and administrative managers and senior staff believe in the usefulness of open data and take an active role in championing open data innovation. However, there also seems to be a vicious circle involved – as long as public managers do not perceive a sufficient societal demand for open data, they remain reluctant to embrace open data innovation; however, as long as public sector organisations do not provide and demonstrate the various possible uses of open data, public demand will remain low due to the lack of knowledge and understanding of the benefits of open data.

According to the survey, the most urgent needs of organisations and stakeholders for engaging in open data-driven innovation are related to the availability and quality of data; the need for a better understanding of the value and potential uses of open data; open data-related skills; time and financial resources (in innovation literature often referred to as organisational slack); supportive organisational culture; senior management support; competent change management; integration of open data in the government's digital policy and organisational strategies; active cross-sectoral collaboration, and concrete guidelines for working with open data.



Therefore, we see that the key challenges and needs for open data innovation, including LOSD innovation, can be related to three main categories:

1. Technology (open data and data infrastructures);
2. The context of technology implementation (including the legal, administrative, cultural and organisational environment, stakeholders, etc.);
3. Processes for innovation implementation (including processes for opening up and exploiting data, organisational innovation, best practice-sharing, user involvement and co-creation).

This suggests the need for a holistic framework for supporting LOSD innovation, which is the key task to be addressed in WP2. Such a framework should take into account 1) the nature of data and data infrastructures that are the essential components of LOSD innovation; 2) the barriers and drivers emerging from the existing attitudes, cultures, knowledge, skills, legislation, policies, strategies, organisational procedures and routines, etc.; and 3) the precise processes that can be employed to enable the creation of high-quality, meaningful, coherent, useful and usable datasets, facilitate the use of such datasets for public sector innovation, enable the involvement of users in the co-creation of services based on linked statistical open data, and support organisational change and innovation management. As the survey demonstrated, the majority of the needs and challenges are similar across countries. However, there are also slight differences in the existing level of open data awareness, “culture”, legislation and policies. Therefore, the framework also requires validation in local contexts.

## 5 Technical challenges and needs of Linked Open Statistical Data

### 5.1 Linked Open Statistical Data: State-of-the-Art

#### 5.1.1 Conceptual framework

In order to synthesize the accumulated knowledge, we performed a concept-centric analysis. The main characteristics of the area were extracted and a conceptual framework for linked data cubes was created in order to structure the area. Finally, the framework was employed to categorise and further analyse the literature and to extract insights into the area of “linked open statistical data” or “linked data cubes”.

The proposed framework (Figure 5) comprises three dimensions:

- The type of contributions in the literature. This includes software, architecture, formal theory, vocabularies, and cases. Some of these categories are further divided into sub-category. For example, the vocabulary type can be related either to a definition of a vocabulary or the application of a vocabulary.
- The linked data cube analysis steps. These include data cube publishing, exploiting, combining, quality assurance, and access control. Two of the steps are further divided. The exploit step can refer to (a) browsing, (b) OLAP analysis, (c) visualisation, and (d) statistical analysis. Moreover, quality assurance can refer to either instance-level quality or schema-level quality.
- Application areas such as health, finance, and environment.

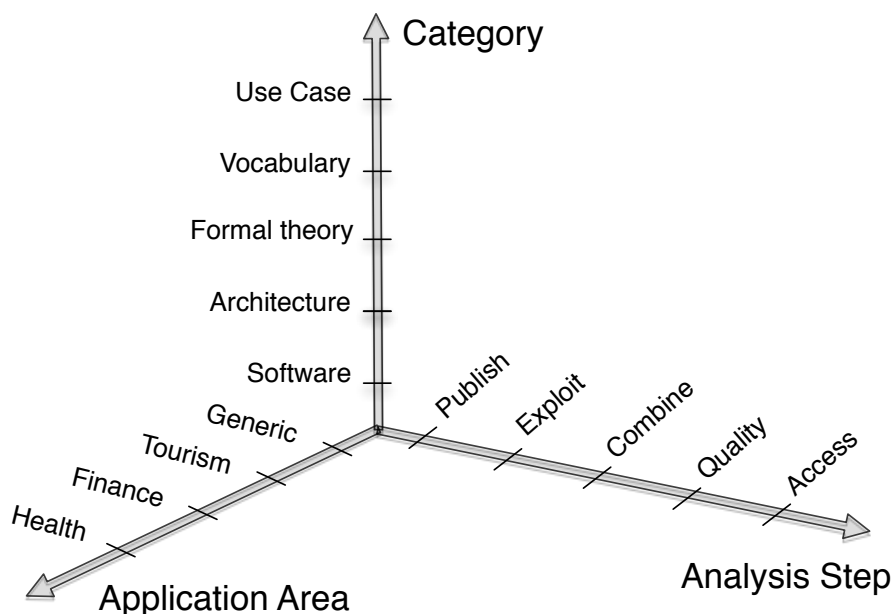


Figure 5 The linked open statistical data framework

#### 5.1.2 Quantitative Analysis

We now employ the framework in order to gain insight into the research on linked data cubes. We initially categorise the article based on the three dimensions.

Using only the “Analysis Step” dimension we can categorise the literature into the five steps. The vast majority of the contributions are related to publishing (42%) and exploitation (37%). The rest of the contributions are related to combining (13%), “Quality” (5%), and “Access” (3%). If we drill-down in the “Exploit” step, then we can see that 38% of the contribution refer to visualisation, 29% to OLAP analysis, 26% to statistical analysis, 7% to browsing.

Using only the “Category” dimension of the framework we can categorise the literature based on the different types of contributions. The vast majority of the contributions describe use cases (38%) and present software tools (38%). Other types of contributions include vocabularies (15%), formal theory (7%), and architectures (2%).

Interestingly, taking into account only the “Application Area” dimension we can see that 40% of the contribution are domain-specific. These contributions span a wide range of domains including health, policy-making, government, environment, economics, biology, and tourism. The most notable domain is health that characterise 25% of all domain-specific contributions.

By synthesizing the results, we can further analyse the literature and make some interesting observations regarding the research contributions so far. First, it is very important to combine the “Category” and “Analysis Step” dimensions. By doing so we can get two different views of the literature, (a) the types of contributions in relation to the analysis step and (b) the analysis steps in relation to the types of contributions.

The vast majority (64%) of the use cases in the literature are about publishing and only 20% and 13% respectively describe cases of exploiting and combining. Software contributions mainly focus on requirements related to exploitation (56%). Out of these exploitation software contributions 55% enables creating visualisations, 24% enables performing statistical analysis, 13% enables browsing, and 8% enables performing OLAP operations on top of linked data cubes. The rest of the software contributions facilitate publishing (25%), support combining (19%), ensure cubes’ quality (6%), and enable controlling access (3%) to linked data cubes. Interestingly, half of the contributions that introduce vocabularies support the publishing of linked data cubes (such as the RDF data cube vocabulary). The rest of the vocabulary related contributions focus on (a) facilitating the exploitation (36%) of linked data cubes, (b) supporting quality assurance (11%), and (c) enabling access control (3%). Finally, formal theory related contributions focus only on issues related to linked data cube integration and exploitation.

If we study the analysis steps according to the types of contribution, we can come up with some interesting results. First, the majority of the publishing contributions are cases (58%), while software (22%) and vocabularies (18%) follow. The majority of the exploitation contributions are software (57%) with use cases (21%), vocabularies (15%), and formal theories (7%) coming next. Finally, regarding the contributions about combining linked data cubes, 38% are use cases, 29% are software tools, and 25% are theoretical contributions.

We can also combine the “Application Area” dimension with the “Category” dimension. If we do so, we will see that 77% of all software tools are domain independent. The same metric is 82% and 79% in the case of software tools for publishing and exploitation respectively. Interestingly, however, this percentage goes down to 57% in the specific case of software tools that support combining of linked

data cubes. On the other hand, 78% of all vocabulary contributions are domain independent. More specifically, domain-specific vocabularies have been developed only for publishing and access control.

### 5.1.3 Qualitative Analysis

We now focus on specific types of contributions of major importance, namely vocabularies, software, architecture, formal theory, and use cases.

#### 5.1.3.1 Vocabularies

**Publish.** Several RDF vocabularies have been proposed to model statistical data as RDF. Two of the first vocabularies introduced were the Statistical Data and Metadata eXchange (SDMX)<sup>1</sup> information model that was proposed to represent statistical data and make them available using web services (Cyganiak et al., 2010) and the Statistical Core Vocabulary (SCOVO)<sup>2</sup> for modeling and publishing statistical data (Hausenblas et al., 2009). Other vocabularies proposed in literature include DDI-RDF Discovery Vocabulary (Bosch & Cyganiak, 2013) for the discovery of statistical data and metadata, Open Cubes (Etcheverry and Vaisman, 2012a), the vocabulary that models Data Warehouses' cubes (Diamantini and Potena, 2009), the SCOVOLink ontology that extends SCOVO in order to allow the creation of links between the data and the described entities (Vrandečić et al., 2010) and the LOIUS ontology that also extends SCOVO to describe statistics about University student activities (Pirrotta, 2010). However the mostly used vocabulary is the RDF Data Cube vocabulary (QB)<sup>3</sup> (Cyganiak and Reynolds, 2014) that models statistical data as RDF. The first version of QB was released on April 2012 but the vocabulary has been a W3C standard since January 2014.

As QB is the most widely used vocabulary in literature, a number of studies extend it in order to overcome some of the QB's limitations or satisfy specific domains' needs and requirements. For example, the Linked Clinical Data Cube (LCDC) is a vocabulary that combines QB and DDI-RDF in order to allow the publication of clinical data as linked data (Lefort & Leroux, 2014). Moreover, Bandholtz et al.'s vocabulary is also a SCOVO-based vocabulary that models German Environmental Specimen Bank (ESB), data that describe the accumulation of pollutants/substances in test subjects at specific places with respect to time (Bandholtz et al., 2009).

**Exploit OLAP.** The REA (Resources, events, agents) - based model for OLAP cubes (Schütz et al., 2013). Prat et al.'s model that represents OLAP cubes as an OWL-DL ontology (Prat et al., 2012a; Prat et al., 2012b). Another example is the QB4OLAP vocabulary, that extends QB in order to allow implementing OLAP operations on cubes such as rollup, slice, dice and drill-across using SPARQL queries (Etcheverry and Vaisman, 2012b; Etcheverry et al., 2014; Etcheverry et al., 2015).

**Exploit Statistical Analysis.** Follenfant et al.'s model (2011) is an extension to the QB vocabulary that enables the description of analytical processes (e.g. data analysis) that can be performed within reporting tools.

---

<sup>1</sup> <http://purl.org/linked-data/sdmx>

<sup>2</sup> <http://vocab.deri.ie/scovo>

<sup>3</sup> <https://www.w3.org/TR/vocab-data-cube/>

**Quality.** A number of vocabularies or extensions to existing vocabularies in literature aim to increase the quality of data cubes. For example, Meroño-Peñuela et al. (2015b) propose an extension to the QB vocabulary so as to facilitate the identification of inconsistencies in data cubes. Moreover, Gayo et al. (2013) propose an extension to the QB vocabulary to represent computational index structures. The vocabulary can be used to compute and validate any type of statistical index. Additional studies propose RDF constraints that can be used to evaluate the quality of statistical data modeled as cubes (Hartmann et al., 2015a and Hartmann et al., 2015b).

**Access Control.** LiMDAC metadata model for describing medical data cubes and LiMDAC access policy model for medical data proposed by Kamateri et al. (2014) can be used to define the data to be protected and to whom access is granted or denied.

Recently, however, the focus has been moved from the definition to the application of vocabularies. For example, Becker et al. (2015a) performed a quantitative survey on how the QB vocabulary is applied to model multidimensional data. Kalampokis et al. (2015) also investigated the challenges related to the different practices that can be followed in applying the QB vocabulary.

#### 5.1.3.2 Software

**Publish.** A number of software solutions have been developed that aim to facilitate the publishing of multidimensional data. For example, the OpenCube toolkit<sup>4</sup> (Kalampokis et al., 2014; Kalampokis et al., 2016) includes a number of open source software components that have been developed to enable data cubes publishing. Specifically, the OpenCube toolkit includes the TARQL extension<sup>5</sup> for the conversion of legacy tabular data to RDF, the D2RQ data cubes extension<sup>6</sup> for the conversion of relational databases to RDF, the JSON-stat2qb data cubes extension<sup>7</sup> for the conversion of JSON-stat files into RDF and the R2RML data cubes extension for data transformation of cubes structured in tabular sources to linked data cubes. Another example is the LOD2 Statistical Workbench<sup>8</sup> (Janev et al., 2014) that includes a set of tools for accessing, manipulating, exploring and publishing statistical data. Specifically, LOD2 includes tools for importing and editing cubes, and managing their dimensions and code lists. Moreover, LODStats<sup>9</sup> (Ermilov et al., 2013) is a web-based tool that collects and publishes statistics about the LOD cloud. At the same time, OLAP2DataCube (Salas et al., 2012a; Salas et al., 2012b) and the CSV2DataCube (Salas et al., 2012a) are both plug-ins for the Ontowiki<sup>10</sup> tool (Frischmuth et al., 2015) for extracting and publishing statistical data in RDF. Specifically, the two plug-ins enable the publishing of OLAP databases to RDF and CSV data to RDF respectively. TabLinker<sup>11</sup> (Meroño-Peñuela et al., 2013) is also a tool for publishing Excel data as data cubes. Another tool for

---

<sup>4</sup> <http://opencube-toolkit.eu/>

<sup>5</sup> <http://opencube-toolkit.eu/tarql-extension-for-data-cubes/>

<sup>6</sup> <http://opencube-toolkit.eu/d2rq-extension-for-data-cubes/>

<sup>7</sup> <http://opencube-toolkit.eu/json-stat2qb-extension-for-data-cubes/>

<sup>8</sup> <http://lod2.stat.gov.rs/lod2statworkbench>

<sup>9</sup> <http://stats.lod2.eu/>

<sup>10</sup> <http://aksw.org/Projects/OntoWiki.html>

<sup>11</sup> <https://github.com/Data2Semantics/TabLinker>

publishing is also developed by the Italian National Institute of Statistics (Istat). The proposed solution aims to facilitate the mapping from the SDMX data model to the QB vocabulary.

**Exploit Browse.** A number of software solutions also enable the browsing of data cubes. For example, LSD Dimensions (Meroño-Peñuela et al., 2014) is a web-based tool for monitoring dimensions and codes (i.e. variables and values) of data cubes. It provides users with a list of the dimensions of cubes stores in Datahub.io and allows users to browse them. Linked Data Cubes Explorer (LDCX)<sup>12</sup> (Kämpgen & Harth, 2014) is another tool that enables the browsing of data cubes. It allows users to select a number of datasets and show and explore their dimensions and measures. The OpenCube Browser (Kalampokis et al., 2016) is also a tool that provides functionalities for exploring linked open statistical data cubes. Finally, the Data catalogue management solution<sup>13</sup>, also part of the OpenCube toolkit, provides wiki pages and templates for viewing data catalogue concepts, such as catalogues, datasets, data distributions, etc.

**Exploit OLAP.** The OpenCube OLAP Browser<sup>14</sup> is a tool that enables using linked data cubes to perform OLAP operations (such as drill-down, roll-up and pivot). Saad et al. (2013) also implement a prototype that enables performing OLAP operations on top of data cubes. Specifically, the prototype allows users to select a data cube and then select the OLAP operation he wants to apply on it.

**Exploit visualizations.** A number of tools have been developed aiming at producing meaningful charts, maps and other visualizations out of statistical data cubes. Visualizations vary from charts to maps. For example, the OpenCube MapView<sup>15</sup> (Tambouris et al., 2015), part of the OpenCube toolkit, enables visualizations of linked data cubes with a geo-spatial dimension on maps. The Interactive chart visualization widgets<sup>16</sup> which is also part of the OpenCube toolkit, allows the visualization of RDF data cubes, in particular, time series data using charts. Another example is the LOD2 Statistical Workbench<sup>17</sup> (Janev et al., 2014) that provides CubeViz (Mader et al., 2014; Martin et al., 2015), a tool for visualizing a data cube's observations with suitable charts. Moreover, Map4rdf (Leon et al., 2012) is a browsing tool that allows the exploring and visualization of RDF data cubes that include geospatial information using maps. Map4rdf supports Google Maps<sup>18</sup> and OpenStreetMap<sup>19</sup>. Map4rdf has also a mobile compatible counterpart, Map4RDF-iOS app which can be installed in mobile devices with iOS and offers similar functionalities with Map4rdf. Another visualization tool is the Linked Data Visualisation Model (Helmich et al., 2014) that also includes a plug-in to facilitate the publication non-data cubes as cubes. The CODE Visualisation Wizard (Mutlu et al., 2013; Mutlu et al., 2014; Tschinkel et al., 2014) is also a platform that imports statistical data, transforms them to data cubes and suggests and creates visualizations (bar charts, lines, pies etc.) on top of them. ETIHQ visual dashboard (Sabou et al., 2015) is also a software that allows the visualization of tourism indicators modeled as data cubes

---

<sup>12</sup> <http://km.aifb.kit.edu/projects/ldcx/>

<sup>13</sup> <http://opencube-toolkit.eu/data-catalogue-management-solution/>

<sup>14</sup> <http://opencube-toolkit.eu/opencube-olap-browser/>

<sup>15</sup> <http://opencube-toolkit.eu/opencube-map-view/>

<sup>16</sup> <http://opencube-toolkit.eu/interactive-chart-visualization-widgets/>

<sup>17</sup> <http://lod2.stat.gov.rs/lod2statworkbench>

<sup>18</sup> <https://developers.google.com/maps/>

<sup>19</sup> <http://www.openstreetmap.org/#map=5/51.500/-0.100>

and was developed to enhance the decision making of Destination Marketing Organizations. qb.js (McCusker et al., 2013) is also another tool that enables the creation of visualizations from data cubes without requiring knowledge of linked data or semantic tools.

**Exploit Statistical Analysis.** Software solutions described in literature also allow performing various types of analytics on top of data cubes. For example, the Linked Open Data Extension of Rapidminer adds to Rapidminer the Data Cube Importer operator (Ristoski et al., 2015). The operator enables the importing of data model using the QB vocabulary so as to perform a plethora of statistical analyses and predictive analytics functions. Moreover, the R statistical analysis module<sup>20</sup> of the OpenCube toolkit allows applying statistical analysis methods to data represented as RDF data cubes.

**Combine.** A lot of tools enable the integration of data cubes. For example, the OpenCube Compatibility Explorer<sup>21</sup> of the OpenCube toolkit allows users to identify compatible cubes for potential merge and then establish typed links to facilitate discovery. Moreover, the OpenCube Aggregator<sup>22</sup> of the OpenCube toolkit enables the aggregation of cubes across a dimension or across a hierarchy. The OpenCube expander<sup>23</sup> is another tool included in the OpenCube toolkit that allows searching for compatible cubes and creating new expanded cubes by merging two compatible cubes. Another example is the LOD2 Statistical Workbench<sup>24</sup> (Janev et al., 2014) includes tools for interlinking the dimensions of two data cubes and for enriching data cubes with external data (e.g. data from dbpedia). Bacon (Bayerl et al., 2015) is also an open source software that enables the fusion of semantically associated cubes and the integration of related cubes into a single cube. The discovery of relative cubes is based on the structure as well as the content of the cubes and the efficiency of the integration is increased by allowing the modification of the structure of and also by detecting duplicate information in the integrated cube.

**Quality - Schema.** Relevant studies propose validation tools such as the RDF Data Cube Validation tool which is part of the LOD2 Statistical Workbench and can be used to validate the integrity constraints defined in the QB vocabulary specification and also to automatic repair identified errors (Janev et al., 2013) and Vital<sup>25</sup> that supports data publishers in the detections of bugs in data cubes such as : insufficient documentation, wrong data types, syntax errors in URIs and inconsistencies between the data structure specifications and the observations (Daga et al., 2014).

**Quality – Data.** Relevant tools for the assessment of the quality of data cubes include Computex<sup>26</sup> (Gayo et al., 2013), a service that allows the validation of statistical index data represented as data cubes.

---

<sup>20</sup> <http://opencube-toolkit.eu/r-statistical-analysis-module/>

<sup>21</sup> <http://opencube-toolkit.eu/opencube-compatibility-explorer/>

<sup>22</sup> <http://opencube-toolkit.eu/opencube-aggregator/>

<sup>23</sup> <http://opencube-toolkit.eu/opencube-expander/>

<sup>24</sup> <http://lod2.stat.gov.rs/lod2statworkbench>

<sup>25</sup> <http://data.open.ac.uk/demo/vital/>

<sup>26</sup> <http://computex.herokuapp.com/>

**Access Control.** Kamateri et al. (2014) propose LiMDAC, a proof-of-concept platform developed to evaluate their framework that enables controlling the access to medical data cubes.

#### 5.1.3.3 Architecture

**Publish.** Ruback et al. (2013) propose an architecture that acts as a mediator for publishing statistical data stored in relational databases as data cubes.

**Exploit OLAP.** Kämpgen & Harth (2014) propose OLAP4LOD architecture to assist developers create applications over statistical data modeled as data cubes. The architecture can be used, for example, to create applications that exploit and analyse statistical data modeled using the QB vocabulary.

#### 5.1.3.4 Formal Theory

**Exploit OLAP.** Kämpgen et al. (2015) are using OLAP algebra to formally define OLAP operations on data cubes. They also present a way to transform OLAP queries to SPARQL queries. In addition, Saad et al. (2013) also use OLAP algebra to translate OLAP operations to SPARQL queries that can be applied to data cubes.

**Combine.** A number of studies present work related to the formal theory of cubes' integration. For example, Meimaris et al. (2014, 2016) propose three methods of identifying complementary and containment relationships between QB observations. A containment relationship captures whether an observation contains aggregated information with respect to other observations while a complementarity relationship captures whether the measures of two observations can be combined together, providing comparable data for QB observations. Moreover, Do et al. (2015b) develop mechanisms for the interconnection of data sets based on their metadata (i.e. components such as dimensions, measure and attributes or values of dimensions or values of attributes). Their approach is based on spatial and temporal dimensions.

#### 5.1.3.5 Use Cases

**Publish.** A large number of studies in literature present the results of different cases of statistical data publishing as cubes. Most of them make use of the QB vocabulary to publish statistical data as RDF especially after the QB vocabulary became a W3C recommendation. These studies are usually applied in different domains and, sometimes, in specific countries. For example, in the health domain, Zaveri et al. (2011) integrate and publish data 1) regarding diseases and 2) the respective research investments as cubes in order to reduce the research-disease burden and Zaveri et al. (2013) publish clinical data from the Global Health Observatory (GHO) of the United Nations' World Health Organization (WHO) as RDF data cubes.

Census data is also another category of data that are widely used in relative cases of data cubes' publishing. For example, Meroño-Peñuela et al. (2015a) publish CEDAR, the Dutch historical censuses as data cubes in order to detect extensional concept drift, Petrou et al. (2013) publish Greek residential statistical data as linked data, Janev et al. (2012) publish Serbian open statistical data as data cubes, Aracri et al. (2014) on behalf of Istat (aka the Italian National Institute of Statistics) published Italian population and housing census data as linked data and Zancanaro et al. (2013) also present a case for



publishing multidimensional statistical linked data. In their case, they identify Brazilian open government data and published them as data cubes using the QB vocabulary.

In the domain of environment and climate Lefort et al. (2012) publish ACORN-SAT, a dataset that contains Australia's daily temperature over the last 100 years and Lefort et al. (2013) publish data cubes using datasets of the Australian Bureau of Meteorology about daily temperature over the last 100 years. They create cubes using the Semantic Sensor Network ontology and the QB vocabulary.

In addition, Ceolin et al. (2010) used the RDF data cube vocabulary to model data regarding annotations of experts for museum artifacts, Hallo et al. (2015) in their study use the QB vocabulary to publish statistical scientific data from Open Journal systems, Höffner et al. (2015) publish finance data from the OpenSpending.org platform as data cubes creating the LinkedSpending dataset with more than five million planned and carried out financial transactions in 627 data sets from all over the world from 2005 to 2035. In the domain of tourism, Sabou et al. (2015) publish tourism-related data as data cubes also using the QB vocabulary. Vilches-Blázquez et. (2014) also present a case of publishing geospatial data from Spanish National data sets.

**Exploit Visualization.** A number of cases in literature also try to prove the importance of re-using data cubes using visualizations. For example, Koho et al., (2012) visualized bird observations and related weather data to assist ornithologists, bird watchers and researchers to explore these data.

**Exploit Statistical Analysis.** A number of studies in literature showcase the statistical analysis of data cubes. For example, Zapilko & Mathiak's approach (2011) aims at assisting researchers perform statistical analysis on linked data. Specifically, they propose using SPARQL queries in order to combine data from distributed sources and then apply simple statistical calculations, such as the computation of the variance of data or the creation of linear regression models. To this end, they propose a three step process: 1) create a SPARQL query using UNION operator to select numerical values from data, 2) store results in arrays and 3) perform the statistical calculations on results. Moreover, Celino & Calegari (2014) present a case of exploiting Milano's geo-statistical data. Specifically, they used k-means clustering to prove the correlation between population demographics from the Italian Institute for Statistics and mobile phone activity data provided by the Telecom Italia mobile operator regarding the city of Milano. They represented the results of the correlation as cubes. Kalampokis et al. (2013) also describe a case study related the general elections of the United Kingdom. They used data from data.gov.uk, the official UK's portal that were modeled as data cubes in order to prove the value of exploiting data cubes. McCusker et al. (2013) also created a case that explores the hypothesis that youth tobacco access laws have consistent, measurable impacts on the rate of change in cigarette smoking among high school students over time. To this end, they modeled relevant statistical data as data cubes and explored the correlation between them using linear regression.

**Combine.** A number of cases in literature also showcase the importance of the integration of data cubes. For example, Do et al. (2015a) integrate data cubes coming from multiple data sources, published in varying formats, use heterogeneous scales, and are accessible by different means. Sato et al. (2013) also carried out a matching between data cubes from different data sources using country codes of area dimensions of the two statistics and year codes of time dimensions of them. They found that it would be easy to identify possible matching between data cubes if the appropriate upper-level

resources were referred to and if the usage of external codes were identified in the schema-level. In addition, Becker et al. (2015b) present a case of data cubes integration. Their approach allows vertical (i.e. union of observations from different data cubes that use the same data structure definition) as well as horizontal integration (i.e. integration datasets with different structure definitions) of data cubes. Vilches-Blázquez et al. (2014) also present a case of integrating geospatial Spanish National data sets coming from different sources (heterogeneous, multidisciplinary, multitemporal, multiresolution, and multilingual).

## 5.2 Challenges & needs regarding LOSD interoperability

The literature review as well as our experience in the development of LOSD software tools reveals that LOSD interoperability is one of the major challenges that the community needs to address in order to achieve the wide adoption of LOSD. The achievement of LOSD interoperability will enable (a) the combination of data cubes across multiple portals and (b) the development of generic software tools that can be widely re-used. At this section we present the results of the study that we performed towards this direction. The first two steps (“Identification of conflicting practices” and “Understanding of conflicting practices”) are presented in detail, while the last step “Consensus on common practices” is still a work in progress, so we present the results so far. The complete results will be presented at WP5.

### 5.2.1 Identification of conflicting practices

The LOSD exploiting challenges occur due to different publishing practices adopted by the existing portals that create non-interoperable data. In this section we aim at identifying all these conflicting practices. As a first step we identify all the existing LOSD official portals, and then we identify the types of conflicts that might occur at LOSD through a state of the art analysis. Finally, we identify the conflicting practices that occur at the LOSD portals using the outcome of the two previous steps.

#### 5.2.1.1 LOSD official portals

At the moment, a number of LOSD are available on the Web through dedicated portals. Some of them are official efforts launched by the organizations that own the data. At the following paragraphs we briefly describe each of the existing official LOSD portals.

The Scottish Government<sup>27</sup> provides the data behind their official statistics on “Neighborhood Statistics” as linked data. They offer 131 linked data cubes categorized to 15 themes (e.g. housing, transport). The cubes comprise 17 distinct measures and 84 dimensions. The offered cubes contain in average 657717 observations, while the geospatial dimension has 8475 distinct values and the time dimension 139 distinct values. The geography (e.g. Parliamentary Constituencies, Council Areas) and time (e.g. 2002, 2002-Q1) has values at different levels.

The UK Department for Communities and Local Government (DCLG)<sup>28</sup> provides official linked open data of a selection of statistics including local government finance, housing and homelessness, well-

---

<sup>27</sup> <http://statistics.gov.scot/>

<sup>28</sup> <http://opendatacommunities.org/>

being and deprivation. They offer 233 linked data cubes categorized to 14 themes (e.g. homelessness, societal well-being). The offered cubes comprise 106 distinct measures, 94 dimensions and contain in average 14437 observations per cube. The geography dimension has 89869 distinct values and the time dimension 110 distinct values, while both have values at different levels e.g. County, Region and 2002, 2002-Q1.

The Italian National Institute of Statistics (ISTAT)<sup>29</sup> makes available the Italian Population and Housing Census 2011 as linked data. They offer 8 linked data cubes that comprise 8 distinct measures and 20 dimensions. The offered cubes contain in average 7060284 observations, while the geography dimension has 426725 distinct values at different granularity e.g. region, province. The time dimension has a fixed value since data are only for 2011.

The Irish Census 2011<sup>30</sup> is published as linked data cubes providing a comprehensive picture of the social and living conditions of the people. They offer 682 linked data cubes that comprise 19 distinct measures and 50 dimensions. The offered cubes contain in average 5292 observations, while the geography dimension has 4806 distinct values at different levels e.g. County, Electoral Division etc. The time dimension does not exist since data are only for 2011. A peculiarity of the Irish Census data cubes is that they offer different data cubes for each of the geographical levels. For example, they offer 12 cubes that measure the unemployment one for each of the 12 geographical levels.

The European Commission's Digital Agenda<sup>31</sup> provides its Scoreboard as linked data cubes. They offer 4 linked data cubes that comprise 7 distinct measures and 16 dimensions. The offered cubes contain in average 145155 observations, while the geography dimension has 61 distinct values and the time dimension 77 distinct values. The geography (e.g. Greece, European Union 28) and time (e.g. 2002, 2002-Q1) has values at different levels. A peculiarity of the Digital Agenda data cubes is that they use a "super-dimension" to embrace the values of dimensions other than time and geography. This means that many cubes are conceptually integrated together through the use of the super-dimension e.g. "Individuals who are born in non-EU country", "Individuals with high formal education", "Unemployed".

The Flemish Government<sup>32</sup> makes available some of their data as link data cubes. The cubes have been published through the OpenCube project with the official permission and contribution of the Flemish Government. Specifically, they offer 11 linked data cubes that comprise 27 distinct measures and 27 dimensions. The offered cubes contain in average 69687 observations, while the geography dimension has 589 distinct values and the time dimension 25 distinct values. The geography dimension has values at different levels e.g. region, province, district.

At the same time, a number of datasets have been transformed to linked data cubes in third parties activities (unofficial). For example, a linked data transformation of Eurostat's data, which was created in the course of a research project, includes more than 5,000 linked data cubes. Moreover, few statistical datasets from the European Central Bank, World Bank, UNESCO and other international organizations have been also transformed to linked data in a third party activity. In addition, census

---

<sup>29</sup> <http://datiopen.istat.it/>

<sup>30</sup> <http://data.cso.ie/>

<sup>31</sup> <http://digital-agenda-data.eu/>

<sup>32</sup> <http://data.opendataforum.info>

data of 2011 from Greece and historical censuses from the Netherlands are available as linked data cubes.

**Table 1 Official LOSD portals**

	Scottish	DCLG	ISTAT	Irish CSO	Flemish	Digital Agenda
<b>Data</b>	Neighbour- hood Statistics	Finance, wellbeing etc.	Italian Census 2011	Irish Census 2011	Flemish Gov. Datasets	Digital Agenda Score- board
<b>Curator</b>	Scottish Govern- ment	DCLG	ISTAT	Irish Central Statistics Office	Flemish Govern- ment	European Commis- sion
<b>Cubes</b>	131	233	8	682	11	4
<b>Measures</b>	17	106	8	19	27	7
<b>Dimensions</b>	84	94	20	50	18	16
<b>Observations</b>	84845456	3075142	56482270	3609306	766552	580620
<b>Triples</b>	901538411	126242629	800369986	20202132	7652149	4767031
<b>GeoValues</b>	8475	89869	426725	4806	589	61
<b>TimeValues</b>	139	110	-	-	25	77

#### 5.2.1.2 Data integration conflicts

At this step we aim at identifying data interoperability and integration conflicts at the literature. Towards this direction, we conduct a literature review on database and data-warehouses focusing on conflicts which may appear when integrating various data sources. The conflicts should also be applicable to the data cube context (i.e. conflicts that are not applicable to data cubes are limited out). We adopt the state-of-the-art analysis method proposed by (Webster & Watson 2002). Initially we performed a systematic search in order to accumulate a complete body of relevant scientific papers. We thereafter studied and filtered these initially identified papers in order to come up with the final set of 18 papers that was included in our research.

Table 2 presents the identified conflicts classified to categories, as well as the papers they appear to.

**Table 2 Literature on Data Integration Conflicts**

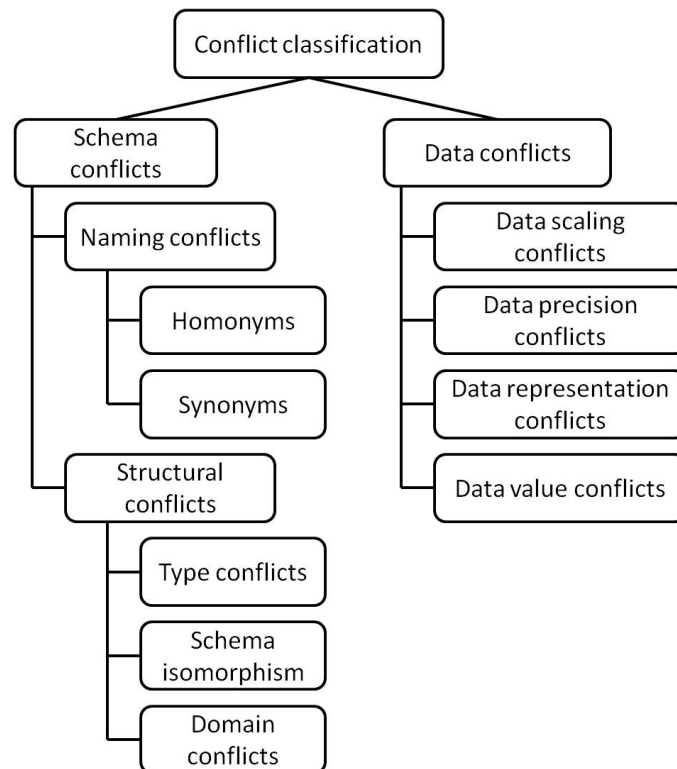
	Schema conflicts				Data conflicts			
	Naming	Structural			Scaling	Precision	Representation	Data value
		Type	Isomorphism	Domain				
(Tseng 2005)	✓	-	-	✓	✓	-	-	✓
(Kim 1991)	✓	✓	✓	✓	✓	✓	✓	✓
(Berger 2012)	✓	✓	✓	✓	✓	✓	-	✓
(Ram 2004)	✓	-	✓	-	✓	✓	✓	✓
(Reddy 1994)	✓	✓	-	-	✓	✓	-	-
(Batini 1986)	✓	✓	-	-	-	-	-	-
(Sheth 1992)	✓	✓	✓	✓	✓	✓	✓	✓
(Channah 1995)	✓	-	-	-	-	✓	-	-
(Doan 2005)	✓	✓	-	-	-	-	-	✓
(Bruckner 2001)	✓	-	-	✓		-	-	-
(Spaccapietra 1992)	✓	✓	-	✓	✓	-	-	-
(Lee 1999)	-	✓	✓	-	✓	✓	✓	-
(Lee 2002)	✓	✓	✓	✓	✓	✓	✓	-
(Sboui 2007)	✓	-	-	-	✓	-	-	-
(Mangisengi 2001)	✓	-	-	-	-	-	-	-
(Diamantini 2014)	✓	-	✓	✓	✓	-	-	✓
(Neumayr 2010)	-	-	-	-	✓	-	-	-
(Torlone 2009)	✓	-	-	✓	✓	-	-	✓

Based on the state of the art analysis the following conflicts have been identified. At high level they are classified to schema and data level conflicts.

- **Schema-level conflicts** result from the use of different schema definitions.
  - **Naming conflicts:** One of the principles of linked data is to use URIs as names for things. As a result, naming conflicts occur at the URI level. At the case of linked data cubes, naming conflicts may appear at all the cube's elements that use URIs i.e. dimensions, measures, measure units, dimension levels and dimension values. Two types of naming conflicts exist:
    - **Homonyms:** two semantically unrelated cube elements use the same URI. For example sdmx-measure:obsValue can be used to encode semantically unrelated measured variables (unemployment, poverty etc).

- **Synonyms:** two semantically similar linked data cube elements use different URIs. For example `sdmx-dimension:refArea` and `eg:geo` can be used to encode the geographical dimension.
- **Structural conflicts:** Structural conflicts occur when different modeling approaches are followed to construct a linked data cube. The following structural conflicts have been identified:
  - **Type conflicts:** the same concept is represented using different classes of the QB vocabulary. For example, the cube dimensions could also be represented as dimension values.
  - **Schema isomorphism:** two linked data cubes have different number of components (i.e. dimension, measure, attribute) at their observations although they have the same components at their data structure definition. For example, two cubes may have the same number of measures, however one can use one measure per observation while the other multiple measures.
  - **Domain conflicts:** two semantically similar dimensions have different set of allowed values. For example, different code lists can be associated with the same dimension at two separate cubes.
- **Data-level conflicts** are due to incompatible or inconsistent data.
  - **Data scaling conflicts:** Data stored using different units of measure. For instance, sales can be measured in euro or in dollars.
  - **Data precision conflicts:** Data stored using different precisions. For example, the "weight" may have values like "heavy", "medium", "light" or values in kilograms e.g. 40 kg.
  - **Data representation conflicts:** Data stored using different formats. For example, the VAT can be represented as a percentage value (23%) or as a decimal (0.23). Another example occurs for dates where different formats can be used e.g. "dd/mm/yy" vs "mm/dd/yy".
  - **Data value conflicts:** Data that have measurements with conflicting values. Such conflicts appear due to wrong, obsolete data or due to different statistical methods employed. For example, the unemployment in a country may be computed using data collected quarterly or annually, these two statistical methods may lead to different measurements.

The following figure shows the classification of the types of conflicts identified at the literature.



**Figure 6 Types of data integration conflicts**

#### 5.2.1.3 Conflicting publishing practices

We use the type of conflicts identified at the literature in order to detect relevant conflicts at the LOSD portals. The outcome is a set of publishing practices that cause conflicts at data integration and interoperability. These practices are presented below.

**P1:** A cube contains a set of measures that represent the phenomena being observed. The following practices exist to define a measure:

**P1.1** Re-use sdmx-measure:obsValue e.g.:

```
eg:obs1 a qb:Observation;
    sdmx-measure:obsValue "0.17"^^xsd:double.
```

**P1.2** Define a new measure based on types of units (e.g. count, ratio). This measure is also defined as subproperty of sdmx-measure:obsValue. Use the sdmx-attribute:unitMeasure to specify the type of measure e.g.:

```
eg:ratio a qb:MeasureProperty;
    rdfs:subPropertyOf sdmx-measure:obsValue.
eg:obs1 a qb:Observation;
    eg:ratio "0.17"^^xsd:double;
    sdmx-attribute:unitMeasure eg:unemployment.
```

**P1.3** Define a new measure. This measure is also defined as subproperty of sdmx-measure:obsValue e.g.:

```
eg:unemployment a qb:MeasureProperty;  
    rdfs:subPropertyOf sdmx-measure:obsValue.  
eg:obs1 a qb:Observation;  
    eg:unemployment "0.17"^^xsd:double.
```

**P2:** Data may contain more than one measure. The following practices exist to represent them:

**P2.1** Create several data cubes with one measure each

**P2.2** Create one data cube with multiple measure

**P3:** The following practices exist to represent multiple measures per cube:

**P3.1** Multi-measure observations (proposed by QB vocabulary). Define multiple qb:MeasureProperty at the data structure definition and use all measures to every observation e.g.:

```
eg:unemployment a qb:MeasureProperty.  
eg:poverty a qb:MeasureProperty.  
eg:obs1 a qb:Observation;  
    eg:unemployment "0.17"^^xsd:double;  
    eg:poverty "0.25"^^xsd:double.
```

**P3.2** Measure dimension (proposed by QB vocabulary): Define multiple qb:MeasureProperty at the data structure definition and use an extra dimension, the qb:measureType, to denote which particular qb:MeasureProperty is being conveyed by the observation e.g.:

```
eg:unemployment a qb:MeasureProperty.  
eg:poverty a qb:MeasureProperty.  
eg:obs1 a qb:Observation;  
    eg:unemployment "0.17"^^xsd:double;  
    qb:measureType eg:unemployment.  
eg:obs2 a qb:Observation;  
    eg:poverty "0.25"^^xsd:double;  
    qb:measureType eg:poverty.
```

**P3.3** Indicator dimension. Use sdmx-measure:obsValue along with a qb:DimensionProperty that indicates the measure being conveyed by the observation e.g.:

```
eg:indicator a qb:DimensionProperty.  
eg:obs1 a qb:Observation;  
    sdmx-measure:obsValue "0.17"^^xsd:double;  
    eg:indicator eg:unemployment.
```



```
eg:obs2 a qb:Observation;  
  sdmx-measure:obsValue "0.25"^^xsd:double;  
  eg:indicator eg:poverty.
```

**P4** An official portal defines a dimension that contains non-associated values (i.e. eg:female, eg:age18-25 eg:large-enterprises). In this case multiple dimensions are embraced together.

**P5** Time, geography, sex and age dimensions as well as the unit of measure attribute are very common among the published cubes. The following practices exist to represent them:

**P5.1** Re-use SDMX e.g.:

```
eg:obs1 a qb:Observation;  
  sdmx-dimension:refArea eg:Greece;  
  eg:unemployment "0.17"^^xsd:double.
```

**P5.2** Define a new dimension property. This property is also a subproperty of SDMX e.g.:

```
eg:geo a qb:DimensionProperty;  
  rdfs:subPropertyOf sdmx-dimension:refArea;  
  qb:codeList eg:Geography.  
eg:obs1 a qb:Observation;  
  eg:geo eg:Greece;  
  eg:unemployment "0.17"^^xsd:double.
```

**P6:** The QB vocabulary enables the declaration of the unit of measure at different levels. The following practices exist to represent them:

**P6.1** Declare the unit at qb:DataSet level e.g.:

```
eg:dataset1 a qb:DataSet;  
  sdmx-attribute:unitMeasure eg:Percent.  
eg:unemployment a qb:MeasureProperty.  
eg:obs1 a qb:Observation;  
  qb:dataSet eg:dataset1;  
  eg:unemployment "0.17"^^xsd:double.
```

**P6.2** Declare the unit at qb:MeasureProperty level e.g.:

```
eg:dataset1 a qb:DataSet.  
eg:unemployment a qb:MeasureProperty;
```

```
sdmx-attribute:unitMeasure eg:Percent.  
eg:obs1 a qb:Observation;  
qb:dataSet eg:dataset1;  
eg:unemployment "0.17"^^xsd:double.
```

**P6.3** Declare the unit at qb:Observation level e.g.:

```
eg:dataset1 a qb:DataSet.  
eg:unemployment a qb:MeasureProperty.  
eg:obs1 a qb:Observation;  
qb:dataSet eg:dataset1;  
eg:unemployment "0.17"^^xsd:double;  
sdmx-attribute:unitMeasure eg:Percent.
```

**P7:** Time dimension values can be represented either as URIs e.g. <http://example.com/2016> or as xsd:date e.g. "2016-01-01"^^xsd:date.

**P8:** The values of the time dimension can be drawn from a code list (in the case that URIs are used). The following code lists are used:

**P8.1** Use reference.data.gov.uk e.g.:

```
http://reference.data.gov.uk/id/year/2015
```

**P8.2** Use DBpedia e.g.:

```
http://dbpedia.org/resource/2015
```

**P8.3** Define a new code list

**P9:** Some datasets contain data with a single value for a dimension (e.g. census data has a single value for the time dimension). The following practices exist:

**P9.1** Express the single value (e.g. use a single time dimension value for census data)

**P9.2** Do not express the single value

**P10:** The values of the sex dimension can be drawn from a code list. The following code lists are used:

**P10.1** Use SDMX code list e.g.:

```
sdmx-code:sex-F, sdmx-code:sex-M
```

**P10.2** Define a new code list

**P11:** The values of the measure units can be drawn from a code list (in the case that URIs are used). The following code lists are used:

**P11.1** Use QUDT (<http://qudt.org/>) e.g.:

```
http://qudt.org/vocab/unit#Euro
```

**P11.2** Use DBpedia e.g.:

```
http://dbpedia.org/resource/Euro
```

**P11.3** Define a new code list

**P12:** The QB vocabulary allows three different practices for defining the values of a dimension:

**P12.1** Use the property `qb:codeList` to associate a `skos:ConceptScheme` to the `qb:DimensionProperty`

```
sdmx-code:sex a skos:ConceptScheme.  
eg:sex a qb:DimensionProperty, qb:CodedProperty;  
    qb:codeList sdmx-code:sex.
```

**P12.2** Use the property `rdfs:range` to associate a `skos:Concept` to the `qb:DimensionProperty`

```
sdmx-code:Sex a rdfs:Class.  
eg:sex a qb:DimensionProperty, qb:CodedProperty;  
    rdfs:range sdmx-code:Sex.
```

**P12.3** Use both

```
sdmx-code:sex a skos:ConceptScheme.  
sdmx-code:Sex a rdfs:Class.  
eg:sex a qb:DimensionProperty, qb:CodedProperty;  
    qb:codeList sdmx-code:sex;  
    rdfs:range sdmx-code:Sex.
```

**P13:** In some cases aggregated values (e.g.:total) are used at the dimensions. For example the sex dimension may have values `sdmx-code:sex-F`, `sdmx-code:sex-M` and `eg:total`.

**P14:** The dimension values often have hierarchical relations i.e. generalization and specialization e.g. Greece is part of Europe. They are also organized into hierarchical levels e.g. region, district. The following practices exist to express them:

**P14.1** Use SKOS to express hierarchical relations and `rdf:type` to express hierarchical levels e.g.:

```
eg:Country rdf:type rdfs:Class.  
eg:Continent rdf:type rdfs:Class.  
eg:Europe rdf:type eg:Continent.  
eg:Greece rdf:type eg:Country;  
           skos:broader eg:Europe.
```

**P14.2** Use XKOS to define hierarchical relations and levels e.g.:

```
eg:Country rdf:type xkos:ClassificationLevel.  
eg:Continent rdf:type xkos:ClassificationLevel.  
eg:Europe skos:member eg:Continent.  
eg:Greece skos:member eg:Country;  
           xkos:isPartOf eg:Europe.
```

**P14.3** Define new properties to express hierarchical relations and `rdf:type` to express hierarchical levels e.g.:

```
eg:Country rdf:type rdfs:Class.  
eg:Continent rdf:type rdfs:Class.  
eg:Europe rdf:type eg:Continent.  
eg:Greece rdf:type eg:Country;  
           eg:within eg:Europe.
```

**P14.4** Use QB vocabulary to define hierarchical relations and `rdf:type` to express hierarchical levels e.g.:

```
eg:geoHierarchy a qb:HierarchicalCodeList;  
               qb:hierarchyRoot eg:Europe;  
               qb:parentChildProperty eg:within.  
eg:Country rdf:type rdfs:Class.  
eg:Continent rdf:type rdfs:Class.  
eg:Europe rdf:type eg:Continent.  
eg:Greece rdf:type eg:Country;  
           eg:within eg:Europe.
```

**P15:** An official portal defines separate cubes for each hierarchical level e.g. defines a cube that contains the regions and another that contains the districts.

### 5.2.2 Understanding of conflicting practices

In order to understand the conflicting practices (i.e. when and why they are used which are their advantages/disadvantages) we actively involve the LOSD experts. The experts' involvement is achieved through a questionnaire.

A set of 11 experts have participated at the study. They are all highly related to the study including curators of linked data cube portals, and data cube publishers. Specifically, the experts that participate are:

- 1 Stefano Abruzzini (Digital Agenda)
- 2 Sarven Capadisli (University of Bonn)
- 3 Oscar Corcho (Universidad Politécnica de Madrid)
- 4 Frank Cotton (INSEE)
- 5 Richard Cyganiak (TopQuadrant)
- 6 Adrian Gschwend (Zazuko)
- 7 Paul Hermans (ProXML)
- 8 Andrei Melis (Eau de Web)
- 9 Dave Reynolds (Epimorphics)
- 10 Bill Roberts (Swirrl)
- 11 Luca Valentino (ISTAT)

A questionnaire was created and distributed to the experts in order to collect their feedback i.e. advantages, disadvantages and peculiarities of all the conflicting practices. The results of the questionnaire are presented below:

**P1: A cube contains a set of measures that represent the phenomena being observed. The following practices exist to define a measure:**

**P1.1** Re-use sdmx-measure:obsValue e.g.:

- Advantages:
  - Easy to implement
  - Easy convert SDMX data to QB. Volume of data available as SDMX are much higher than RDF data cubes
- Disadvantages:
  - Need to attach additional metadata to the dataset to see what is measured
  - Dataset cannot contain more than one measure
  - Prevent advanced operations on data cubes e.g. identifying cubes related by measure in order to merge or compare them

**P1.2** Define a new measure based on types of units (e.g. count, ratio). This measure is also defined as subproperty of sdmx-measure:obsValue. Use the sdmx-attribute:unitMeasure to specify the type of measure:

- Advantages:
  - Treats the unit as an identifiable and reusable concept.
  - Make it easy to access information in a generic structure for driving a user interface
  - Having datasets with e.g. measure = count and unit = people gives you the option to compare/ combine those datasets

- Determine programmatically if observations with a particular measure are suitable for aggregation - so to differentiate 'counts' and 'ratios'.
- Disadvantages:
  - It has limitations given that the RDF Data Cube vocabulary is based off SDMX 2.0
  - Requires a vocabulary for unit types.
  - sdmx-attribute:unitMeasure should point to the units of measurement not the thing measured.
  - With ratio type data there can get a lot of different units

**P1.3** Define a new measure. This measure is also defined as subproperty of sdmx-measure:obsValue e.g.:

- Advantages:
  - reusable and extensible i.e. add additional properties to the measure
  - It aids readability to define a specific qb:MeasureProperty
  - Ideal when publishing new data directly as RDF QB
- Disadvantages:
  - Create a very long list of such measures
  - Define measures as subProperty of sdmx-measure:obsValue is not a good practice especially for multi-measure cubes

**P2:** Data may contain more than one measure. The following practices exist to represent them:

**P2.1** Create several data cubes with one measure each

- Comment:
  - Use separate cubes when the measures are essentially independent questions.

**P2.2** Create one data cube with multiple measure

- Comment:
  - Use one data cube when data has truly more than one measures, in the sense that values are closely related to a single observational event e.g. sensor network measurements and forecasting

**P3:** The following practices exist to represent multiple measures per cube:

**P3.1** Multi-measure observations

- Advantages:
  - Reduced space. It produces low number of observations.
- Disadvantages:
  - Observations cannot be easily re-used in other context since they contain many measures
  - Applicable if measures are unit-less or share units and other attributes, and they are measured together

**P3.2** Measure dimension

- Advantages:
  - Observations can easily be re-used in other context
  - Enables more control over attributes - the ability to annotate individual measurements
- Disadvantages:
  - Increased space. It produces high number of observations.

**P3.3** Indicator dimension.

- Advantages:
  - Observations can easily be re-used in other context
  - Enables more control over attributes - the ability to annotate individual measurements
- Disadvantages:
  - Not a practice proposed by the QB vocabulary.
  - Increased space. It produces high number of observations.

**P4** An official portal defines a dimension that contains non-associated values (i.e. eg:female, eg:age18-25 eg:large-enterprises). In this case multiple dimensions are embraced together.

- Advantages
  - When converting to RDF existing datasets published as SDMX or relational databases this is the only feasible option not requiring manual intervention
- Disadvantages
  - The possible values of a coherent dimension should be comparable in some sense.

**P5** Time, geography, sex and age dimensions as well as the unit of measure attribute are very common among the published cubes. The following practices exist to represent them:

**P5.1** Re-use SDMX

- Advantages
  - Easier to compare data from different places
- Disadvantages
  - Cannot add extra properties to dimensions e.g. qb:codeList, rdfs:comment, rdfs:label

**P5.2** Define a new dimension property. This property is also a subproperty of SDMX

- Advantages
  - Can add extra properties to dimensions e.g. qb:codeList, rdfs:comment, rdfs:label
- Disadvantages
  - Leads to similar dimensions meaning the same thing

**P6:** The QB vocabulary enables the declaration of the unit of measure at different levels. The following practices exist to represent them:

**P6.1** Declare the unit at qb:DataSet level

- Advantages
  - Identify available units from the Data Structure Definition (no need to search the observations)
- Disadvantages
  - Cannot be used if there are more than one measure
  - Cannot be used if the measure has more than one unit

**P6.2** Declare the unit at qb:MeasureProperty level

- Advantages
  - Identify available units from the Data Structure Definition (no need to search the observations)
  - Can be used if there are more than one measure
- Disadvantages
  - Cannot be used if a measure has more than one unit

**P6.3** Declare the unit at qb:Observation level

- Advantages
  - Can be used if there are more than one measure

- Can be used if a measure has more than one unit
- Disadvantages
  - Cannot identify available units from the Data Structure Definition (need to search the observations)

**P7:** Time dimension values can be represented either as URIs e.g. <http://example.com/2016> or as `xsd:date` e.g. `"2016-01-01"^^xsd:date`.

**P7.1** XSD (e.g. `xsd:date`, `xsd:gYearMonth` and `xsd:gYear`)

- Advantages
  - Can refer to a specific point in time
  - Easier to query with SPARQL
- Disadvantages
  - Can't specify time periods
  - Cannot define further properties e.g. label
  - Cannot represent hierarchy of time intervals

**P7.2** URI

- Advantages
  - Can define a time period (e.g. year, semester, quarter etc.). You can define precisely the start and end instants of the time interval
  - Can define further properties e.g. label
  - Enable the representation of a hierarchy of time intervals
- Disadvantages
  - Not easy to SPARQL

**P8:** The values of the time dimension can be drawn from a code list (in the case that URIs are used).  
The following code lists are used:

**P8.1** Use [reference.data.gov.uk](http://reference.data.gov.uk) e.g.:

- Advantages
  - URIs are dynamically created and the service is really stable
  - Defines a wide range of time intervals
  - The URIs are dereferenceable
- Disadvantages
  - Cannot be queried via a SPARQL endpoint
  - Some intervals are specific to the UK public sector such as 'government-year' ( 1 April to 31 March)
  - There is no HTML representation of the URIs

**P8.2** Use DBpedia

**P8.3** Define a new code list

- Advantages
  - Enables the definition of new properties e.g. add `prefLabel` in a different language or `altLabel`'s and links to legislation
- Disadvantages
  - Hamper interoperability between different datasets



**P9:** Some datasets contain data with a single value for a dimension (e.g. census data has a single value for the time dimension). The following practices exist:

**P9.1** Express the single value (e.g. use a single time dimension value for census data)

- Advantages
  - Enables the addition of more observations with different values for the dimension
  - Enables the “merge” of data with different dimension values
- Disadvantages
  - Increased space

**P9.2** Do not express the single value

- Advantages
  - Decreased space
- Disadvantages
  - Does not enable the addition of more observations with different values for the dimension
  - Does not enable the “merge” of data with different dimension values

**P10:** The values of the sex dimension can be drawn from a code list. The following code lists are used:

**P10.1** Use SDMX code list e.g.:

- Advantages
  - Re-usability and interoperability between different datasets
- Disadvantages
  - Does not enable the definition of nuanced notions of sex i.e. biological gender (e.g. hermaphroditism, transgender), self-identification (e.g. asexual)
  - Does not enable the definition of new properties e.g. add prefLabel in a different language or altLabel's and links to legislation

**P10.2** Define a new code list

- Advantages
  - Enables the definition of nuanced notions of sex i.e. biological gender (e.g. hermaphroditism, transgender), self-identification (e.g. asexual)
  - Enables the definition of new properties e.g. add prefLabel in a different language or altLabel's and links to legislation
- Disadvantages
  - Hamper interoperability between different datasets

**P11:** The values of the measure units can be drawn from a code list (in the case that URIs are used). The following code lists are used:

**P11.1** Use QUDT (<http://qudt.org/>) e.g.:

- Advantages
  - It covers a wide range of units
  - Units are well organized
- Disadvantages
  - It is not complete

**P11.2** Use DBpedia e.g.:

**P11.3** Define a new code list

- Advantages
  - Enables the definition of new properties e.g. add prefLabel in a different language or altLabel's and links to legislation
- Disadvantages
  - Hamper interoperability between different datasets

**P12:** The QB vocabulary allows three different practices for defining the values of a dimension:

**P12.1** Use the property qb:codeList to associate a skos:ConceptScheme to the qb:DimensionProperty

- Advantage: Enable the definition of hierarchies

**P12.2** Use the property rdfs:range to associate a skos:Concept to the qb:DimensionProperty

Disadvantage: Does not enable the definition of hierarchies

**P12.3** Use both

- Advantage: Enable the definition of hierarchies
- Disadvantage: Redundancy

**P13:** In some cases aggregated values (e.g:total) are used at the dimensions. For example the sex dimension may have values sdmx-code:sex-F, sdmx-code:sex-M and eg:total.

- Advantages
  - It's easier to have that information explicitly in the dataset rather than have to calculate it
- Disadvantages
  - Need to differentiate the "total" code from "normal" codes in the code list.
  - Aggregated values from different datasets cannot be compared or merged since e.g. not in all cases the sum, for instance, will be equivalent to the total.

**P14:** The dimension values often have hierarchical relations i.e. generalization and specialization e.g. Greece is part of Europe. They are also organized into hierarchical levels e.g. region, district. The following practices exist to express them:

**P14.1** Use SKOS to express hierarchical relations and rdf:type to express hierarchical levels

- Advantages
  - Re-use a standard vocabulary
- Disadvantages
  - Do not enable the representation of more nuances e.g. contained v.s. administeredBy
  - Can represent only tree structures not graphs
  - The use of rdf:type to express hierarchical levels seems too generic and may be used for other purposes (a dimension value can have in theory multiple rdf:type e.g. skos:Concept, sdmx:Concept, sdmx-code:Sex),

**P14.2** Use XKOS to define hierarchical relations and levels

- Advantages
  - Re-use a common vocabulary
  - Use the same vocabulary to define both hierarchical relations and levels
- Disadvantages
  - Do not enable the representation of more nuances e.g. contained v.s. administeredBy
  - Can represent only tree structures not graphs
  - Verbose and a bit difficult to work with

**P14.3** Define new properties to express hierarchical relations and `rdf:type` to express hierarchical levels

- Advantages
  - It enables the representation of more nuances e.g. contained v.s. administeredBy
  - Enables the representation of graph structure rather than a tree
- Disadvantages
  - Needs extra logic in case one wants to let the machine do some work

**P14.4** Use QB vocabulary to define hierarchical relations and `rdf:type` to express hierarchical levels

- Advantages
  - It enables the representation of more nuances e.g. contained v.s. administeredBy
  - Enables the representation of graph structure rather than a tree
  - Re-use a standard vocabulary
- Disadvantages
  - The use of `rdf:type` to express hierarchical levels seems too generic and may be used for other purposes (a dimension value can have in theory multiple `rdf:type` e.g. `skos:Concept`, `sdmx:Concept`, `sdmx-code:Sex`),

**P15:** An official portal defines separate cubes for each hierarchical level e.g. defines a cube that contains the regions and another that contains the districts.

- Advantages
  - Better performance since it generates cubes with low number of observations.
  - Common approach in official statistical portals
- Disadvantages
  - Complex for anyone who wants the details because they have to reassemble the hierarchical cube from the different splits.

### 5.2.3 Consensus on common practices

After identifying and understanding the conflicting practices there is a need to conclude at a set of common publishing practices to address the LOSD exploiting challenges. The common practices do not aim to solve conflicts of already published data, but they intend to be widely followed when publishing in order to create interoperable data. Towards this direction, we employ the Delphi method (Hsu, 2007) in order to reach a consensus on the common practices. Delphi requires the involvement of experts through a questionnaire with multiple iterations to collect feedback until a consensus is reached. The feedback process allows experts to reassess an initial judgment based on the anonymous comments and feedback provided by other experts. The first round begins with an open-ended questionnaire while the next rounds narrow down the options.

The Delphi study is still a work in progress and the final results will be presented at WP5. However, some preliminary results (i.e. common practices already with a consensus) are also presented at this deliverable as a set of Best Practices (BP):

**BP1: Defining a measure.** Re-use existing measures where possible. If not possible, then define new specific measures corresponding to the phenomenon being observed. The measure should NOT be a subproperty of `sdmx-measure:obsValue`. Defining a measure as subproperty of `sdmx-measure:obsValue` does not add any more semantics than defining it as a `qb:MeasureProperty`.

```
eg:unemployment a qb:MeasureProperty;  
    rdfs:label "unemployment"@en.  
eg:obs1 a qb:Observation;  
    eg:unemployment "0.17"^^xsd:double.
```

**BP 2: Multiple measures.** A data cube should contain multiple measures only when they are closely related to a single observational event e.g. sensor measurements.

**BP 3: Coherent dimension.** All values of a coherent dimension should be comparable in some way. Dimensions with non-associated values should be avoided.

**BP 4: Time dimension values.** XSD (e.g. `xsd:date`) should be used when referring to a specific point in time while URIs should be used to express time periods and/or hierarchies e.g. year, semester, quarter etc.

**BP 5: Code lists.** Re-use existing code lists. If not sufficient define your own code lists and map to the existing where possible. The following code list should be preferred:

1. Time dimension: `reference.data.gov.uk`
2. Sex dimension: SDMX
3. Unit of measure: QUDT

**BP 6: Sex dimension.** The `sdmx-dimension:sex` is associated with the SDMX sex code list (i.e. sex-F, sex-M). If the code list is not sufficient e.g. more nuanced notions of sex are needed like: i) biological gender (hermaphroditism, transgender), ii) self-identification (e.g. asexual), then a new dimension should be defined and associated with the new code list.

## 6 Users' challenges regarding the exploitation of Open Statistical Data

In this section we define the challenges regarding the exploitation of Open Statistical Data by taking into account the perspective of the actual users. We divide this section into two sub-sections:

- In the first case we play the role of a user and we go directly to an advanced Open Data Portal in order to discover statistical data regarding to a specific phenomenon.
- In the second case we elicit the opinion of developers who have used Open Statistical Data to create a product or service.

### 6.1 Open Statistical Data Fragmentation

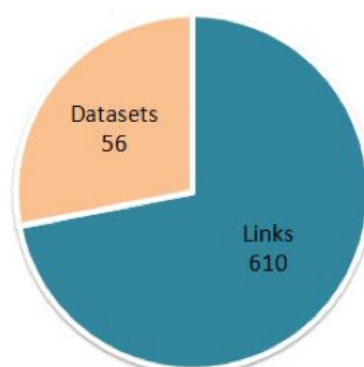
In this step of our analysis we searched on the UK Open Data Portal (i.e. [data.gov.uk](http://data.gov.uk)) for datasets about unemployment in order to understand users' challenges related to searching for statistical data on Open Data Portals. In summary, our research revealed that searching this portal for useful data on unemployment results in large numbers of datasets and links. In particular, if we search on <http://data.gov.uk> for datasets using the keyword "unemployment," we will come up with 122 results that provide access to 56 files and 610 links to 18 other portals (such as the Office for National Statistics and the National Archives) and by following the relevant links to more than 2,000 other files.

We call *open data fragmentation* the situation where collections of relevant open data are broken down into many pieces that are not close together. This definition is actually an adaptation of the definition of data fragmentation in computing.

In particular, the outcome of this activity includes 122 search results that contain either datasets or links. Search results containing one or more datasets may also include an additional link directing users to another portal where they can download the same datasets. Datasets come in various technical formats such as XLS, CSV, PDF, etc. These datasets are directly downloaded from [data.gov.uk](http://data.gov.uk).

On the other hand, search results can contain one or more links, which direct the users to other portals where they can find links or datasets, in order to download them.

The 122 search results lead to 56 datasets, which can be downloaded directly from [data.gov.uk](http://data.gov.uk) and 610 links, where users can follow in order to find and download datasets. It is noticed that some search results encapsulate exactly the same data but in different format. In that case, datasets and links having the exact same data are counted as one.

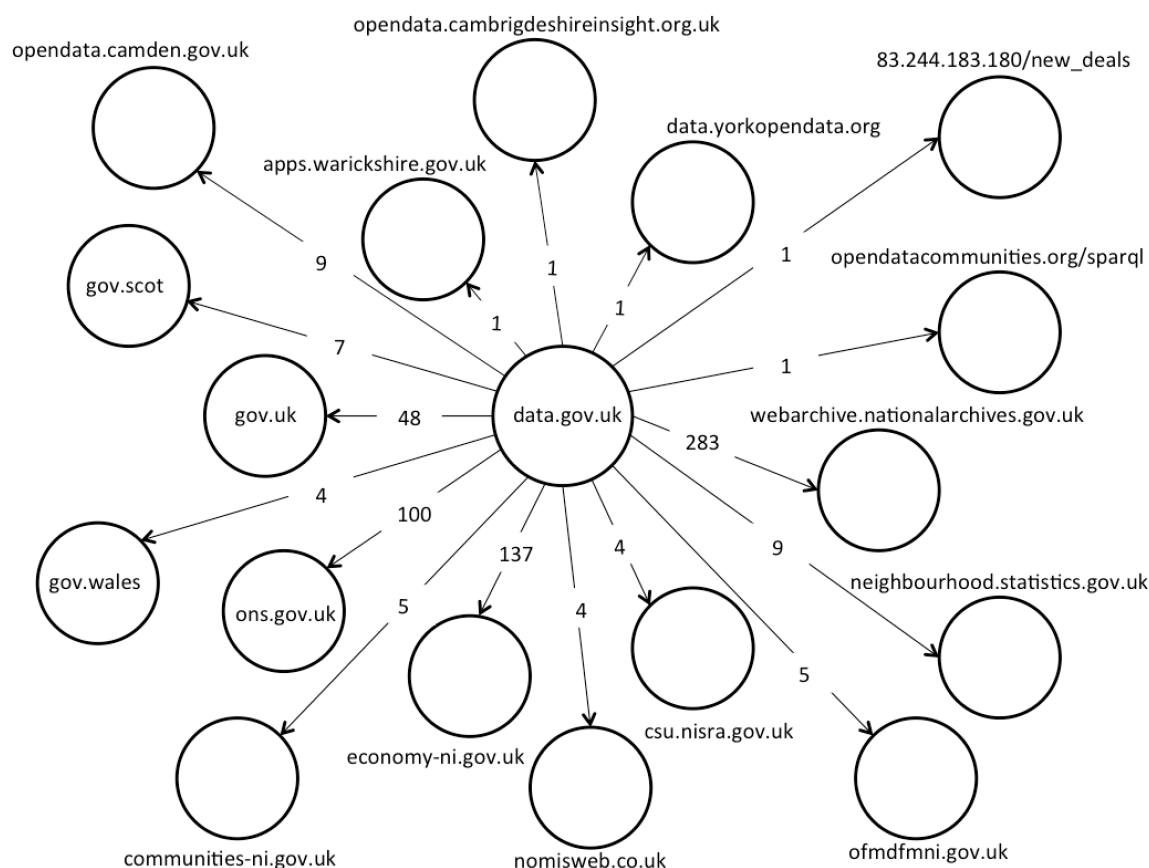


**Figure 7 Search results to datasets & links about unemployment in data.gov.uk**

Search results containing links direct users from data.gov.uk to other portals. For search results that contain datasets, there are often additional links (or another link in the details of search result). We thereafter counted every link that leads to the portals that supply data.gov.uk with data. The identified links in the results lead users to the following portals.

- |  |   |
|--|---|
| 1. <a href="https://opendata.camden.gov.uk">opendata.camden.gov.uk</a>   | 10. <a href="https://csu.nisra.gov.uk">csu.nisra.gov.uk</a>   |
| 2. <a href="https://gov.scot">gov.scot</a>                               | 11. <a href="https://ofmdfmi.gov.uk">ofmdfmi.gov.uk</a>   |
| 3. <a href="https://gov.uk">gov.uk</a>                                   | 12. <a href="https://neighbourhood.statistics.gov.uk">neighbourhood.statistics.gov.uk</a>             |
| 4. <a href="https://gov.wales">gov.wales</a>                             | 13. <a href="https://webarchive.nationalarchives.gov.uk">webarchive.nationalarchives.gov.uk</a>       |
| 5. <a href="https://ons.gov.uk">ons.gov.uk</a>                           | 14. <a href="https://opendatacommunities.org/sparql">opendatacommunities.org/sparql</a>               |
| 6. <a href="https://communities-ni.gov.uk">communities-ni.gov.uk</a>     | 15. <a href="https://83.244.183.180/new_deals">83.244.183.180/new_deals</a>                           |
| 7. <a href="https://economy-ni.gov.uk">economy-ni.gov.uk</a>             | 16. <a href="https://data.yorkopendata.org">data.yorkopendata.org</a>                                 |
| 8. <a href="https://nomisweb.co.uk">nomisweb.co.uk</a>                   | 17. <a href="https://opendata.cambridgeshireinsight.org.uk">opendata.cambridgeshireinsight.org.uk</a> |
| 9. <a href="https://apps.warickshire.gov.uk">apps.warickshire.gov.uk</a> |   |

Moreover, if we count the number of links to different pages of a portal we come up with the following graph (Figure 8).



**Figure 8 Portals of data.gov.uk**

Portals are presented on the graph as nodes. At the centre of the graph is data.gov.uk, which we consider as the entry point of users. Each edge presents the connection between data.gov.uk and portals. The number on each edge is the amount of links directing to a portal. The links directing to the exact same page of a portal are counted as one. For example, there is a relationship between data.gov.uk and gov.uk as is shown in Figure 8.

The users can either download datasets directly from data.gov.uk or follow a link to other portals. That means each portal provides datasets to data.gov.uk in two ways (a) directly from data.gov.uk (search results containing datasets) and (b) exclusively from a third portal (search results containing links).

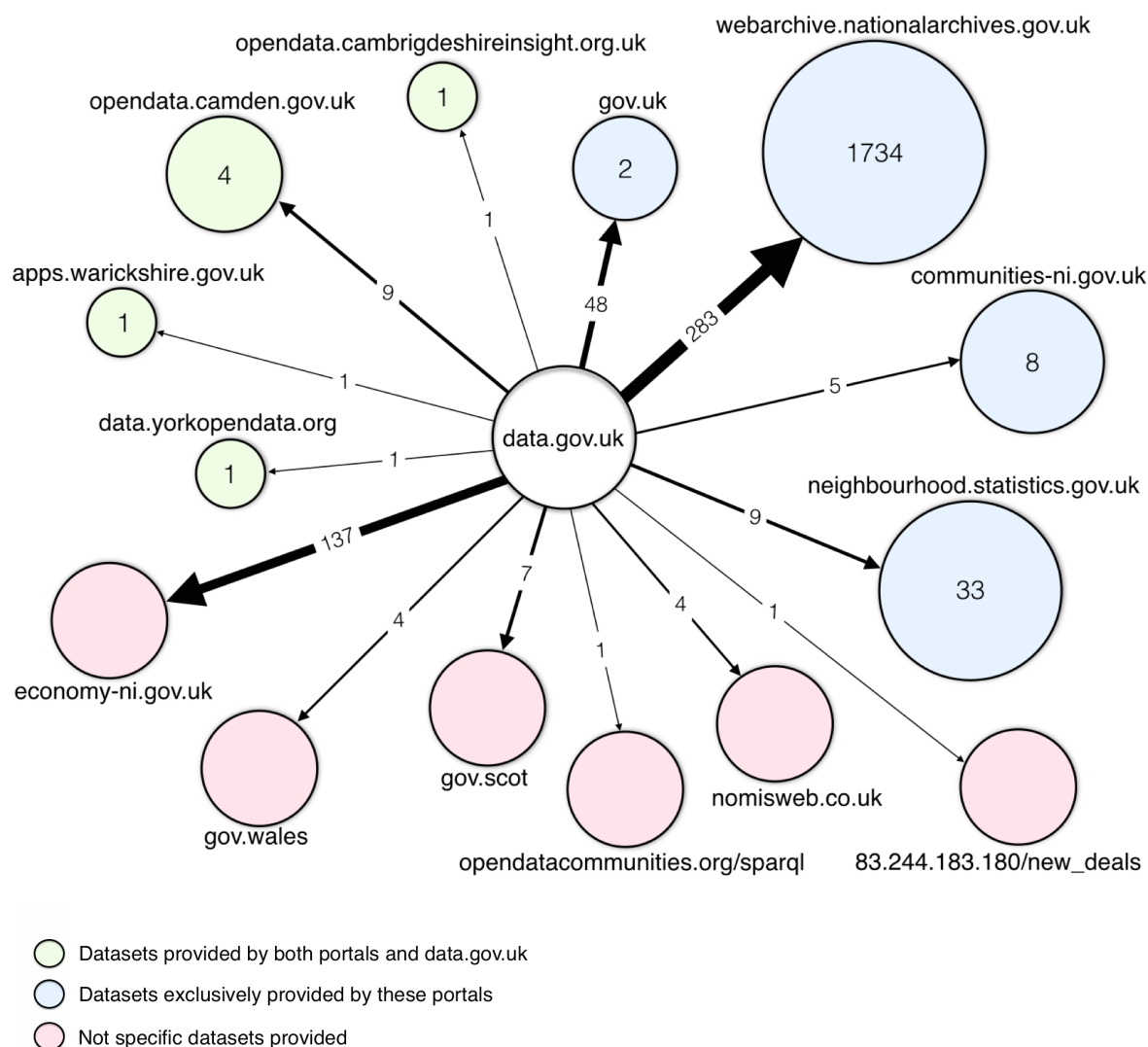
It was noticed, however, that some links lead to an error message. In addition, many links can drive the users to the homepage or the search page of a portal.

Our analysis focuses on the 56 datasets that the users can download directly from data.gov.uk and the datasets that users can download instantly after clicking on a link.

After analysing these datasets, we separated them according to their relevance to unemployment and their resource format.

Datasets that are irrelevant to unemployment or are in PDF format (even if they are relevant) are not taken into account. Datasets that are relevant to unemployment and are in a non-PDF format like XLS, CSV and RDF are counted.

After clarifying the way datasets are provided by portals and focusing on the relevant datasets, which are not in pdf format, the following graph is created (Figure 9):



**Figure 9 Portals of data.gov.uk by modes of providing relevant & non-PDF datasets**

The exact number of all portals is seventeen (as shown in Figure 8). The graph in Figure 9, however, presents just the fourteen sources that supply data.gov.uk with only relevant and non-PDF datasets. The three portals that supply data.gov.uk with only irrelevant datasets or datasets in PDF format, or their links that may result to an error message, are not included in Figure 9.

The number inside a node presents the amount of relevant and non-PDF datasets in each portal that the users can reach through data.gov.uk. That means these portals may include much more datasets than the ones accessed through data.gov.uk. It should be mentioned that the number of datasets in a node could be the result of just one link between data.gov.uk and a portal.

Portals of relevant and non-PDF datasets are divided into three categories:

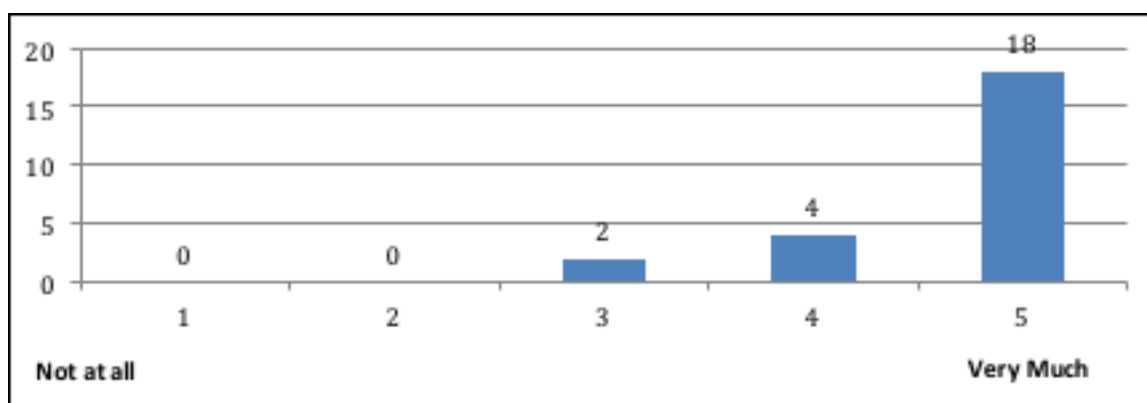


1. Green nodes present the portals in which datasets can be also downloaded from data.gov.uk.
2. Blue nodes present the portals that provide datasets exclusively by their portal. The users follow a link from data.gov.uk to “blue portals” where they can download the available datasets instantly. Datasets are not available to data.gov.uk; the users must follow the link.
3. Pink nodes present the portals that do not supply users instantly with datasets but instead they have to search for them. The users follow a link that directs them to a portal where there are no datasets to download directly. Usually it is a home page or a search page of these portals or even a SPARQL endpoint. The user must navigate to detect the datasets. In other words, these portals may feature relevant and non-PDF datasets but the users have to search for them in order to download them.

## 6.2 Challenges & Needs of Software Developers

In this section we present the result of the online survey that we performed aiming at eliciting the opinion of developers that have used Open Statistical Data (OSD). In total, 24 developers from Belgium, Greece, Brazil, England and Wales answered the questionnaire. These answers were collected and analysed:

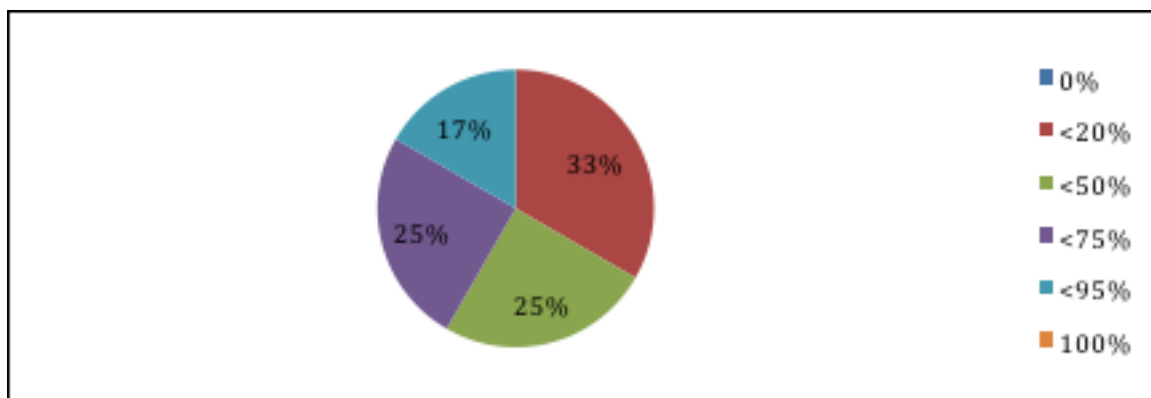
*How important are the open statistical data that are produced by governments and organizations, according to your opinion?*



**Figure 10 Importance of Open Statistical Data**

The potential of open statistical data which are produced by Governments and organizations is recognized by most of developers as they believe (more than 70%) that OSD are very important (Figure 10).

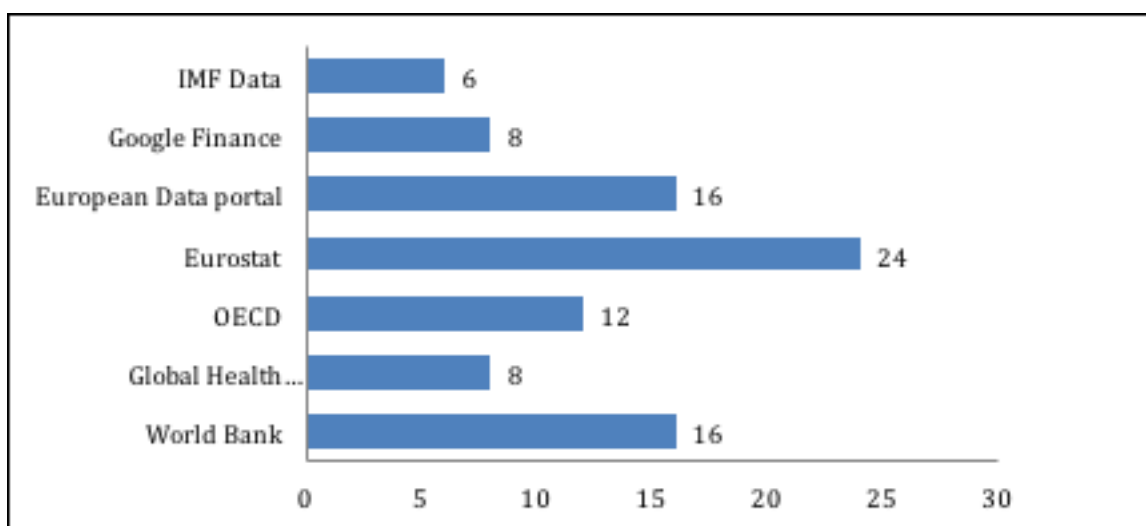
*What percentage of the data provided by national open data portals is statistical according to your perception?*



**Figure 11 Developers' perception of the volume of OSD**

More than the half of developers believes that the percentage of provided data which are statistical is lower than 50%. So, the problem of national portals of open data which do not provide statistical data is recognized.

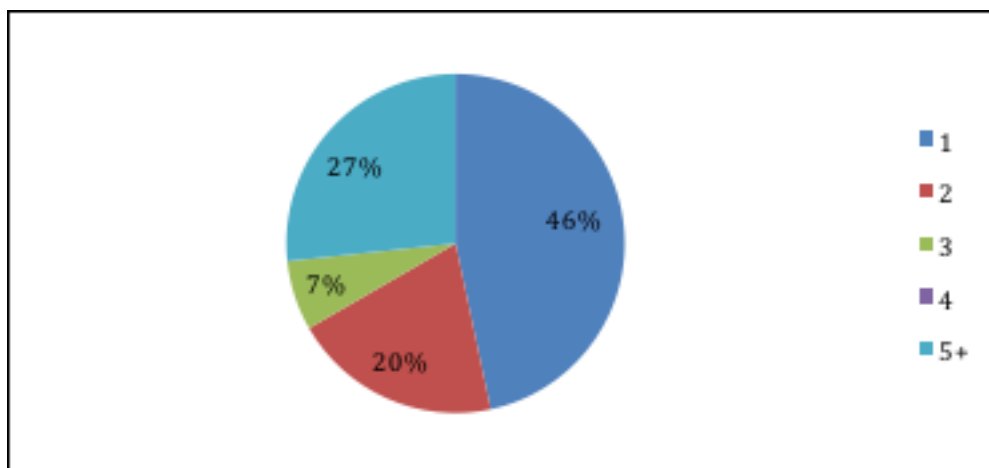
*Which of the following open statistical data sources are you aware of?*



**Figure 12 Sources of OSD**

Eurostat, World Bank and European Data portal are at the most popular open statistical data sources for which developers are aware of. OECD, GHO, Google Finance and IMF data come to a lower recognition by developers as more than 20% and less than 50% of them are aware of these statistical data sources.

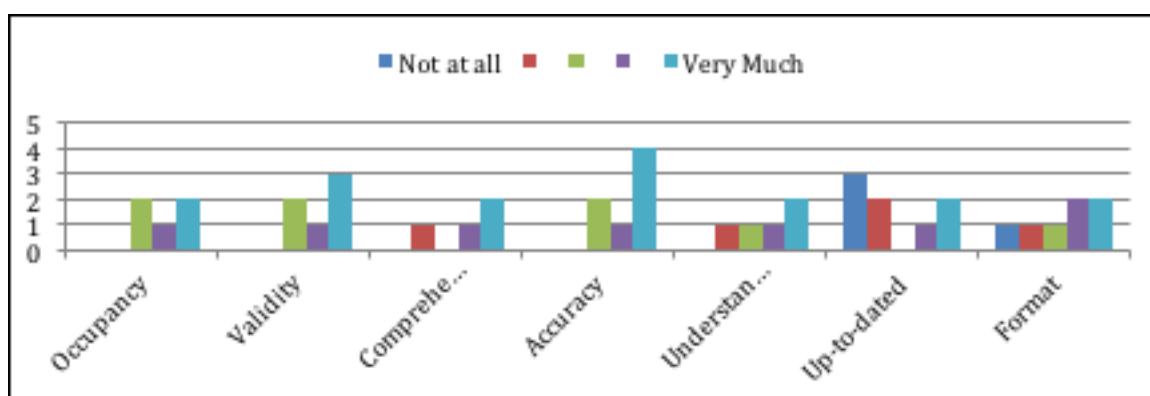
*During the development of your application, how many portals did you use in total to collect your data and how many datasets? How many of them were open statistical data?*



**Figure 13 Number of portals per application**

Just one portal was used by almost 50% of developers to collect their data for their applications. Developers used 189 datasets in total, 161 from which were open statistical data. Thus, they used datasets which had at 85% open statistical data during the development of their applications.

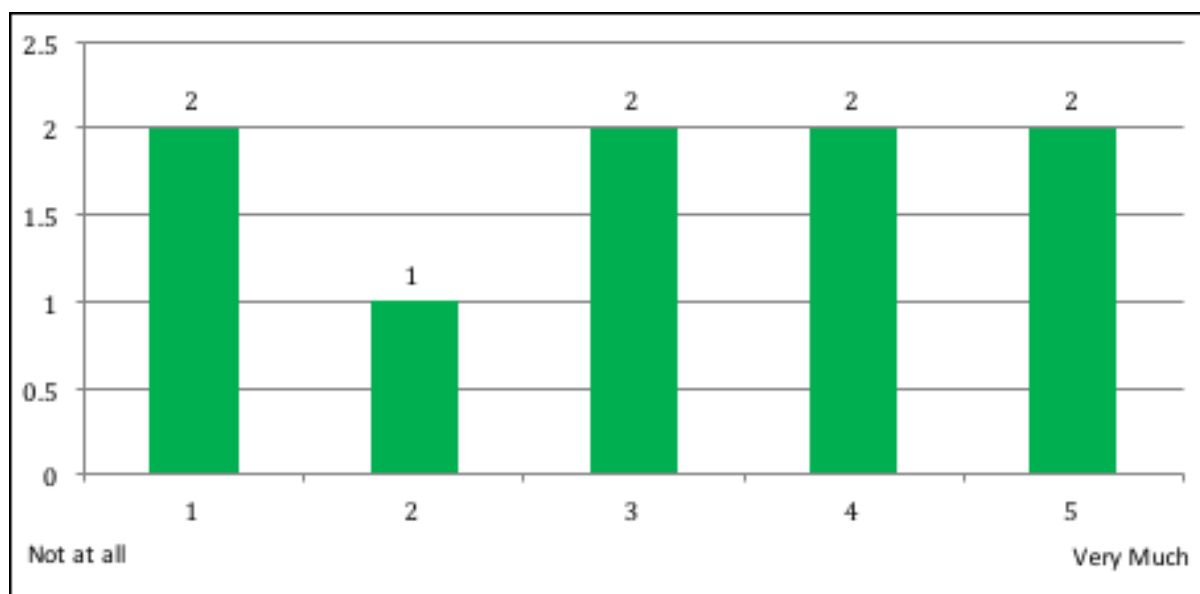
*How satisfied are you with the quality of the open statistical data that you used in your application?*



**Figure 14 Quality of OSD**

Quality of data encompasses some measures in which developers are fully satisfied according to the bar chart. The only exception is the attribute of “up-to-dated” data in which developers are at 50% dissatisfied enough.

*How difficult was it for you to find the open statistical data that you used for your application and which was the greatest difficulty you encountered during this process?*



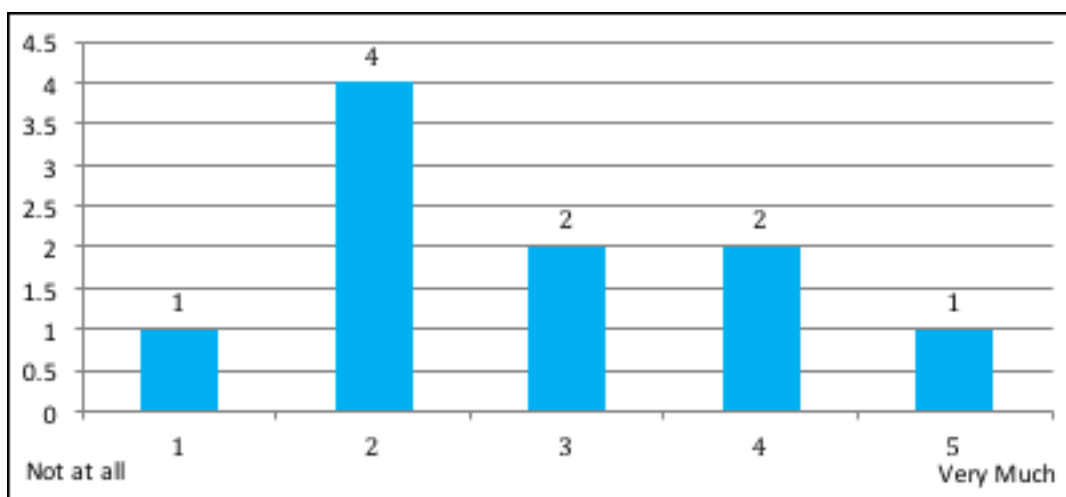
**Figure 15 Challenges on OSD discovery**

The difficulty of finding OSD are almost equally divided at the rating scale by the developers without let us make some important conclusions.

However, some of the challenges that the developers mentioned include:

- Finding the right organisation that owns the data
- Finding the right portal that provides the data
- They do not have access to the data
- The data is not well advertised
- The search functionality of data portals is not very effective
- The technical formats of the available data

*How difficult was it for you to combine the open statistical data that you used for your application and which was the greatest difficulty that you encountered during this process?*

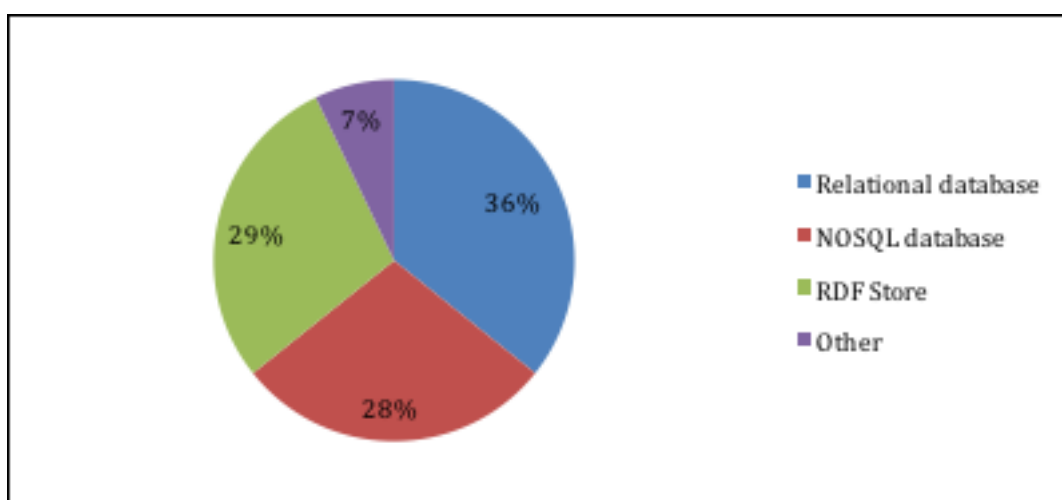


**Figure 16 Challenges on Combining OSD**

Developers did not face many difficulties trying to combine the OSD that they needed for their application. Some problems that were mentioned include:

- Interoperability among datasets
- Lack and quality of metadata
- Technical formats that do not facilitate integration
- No streaming data

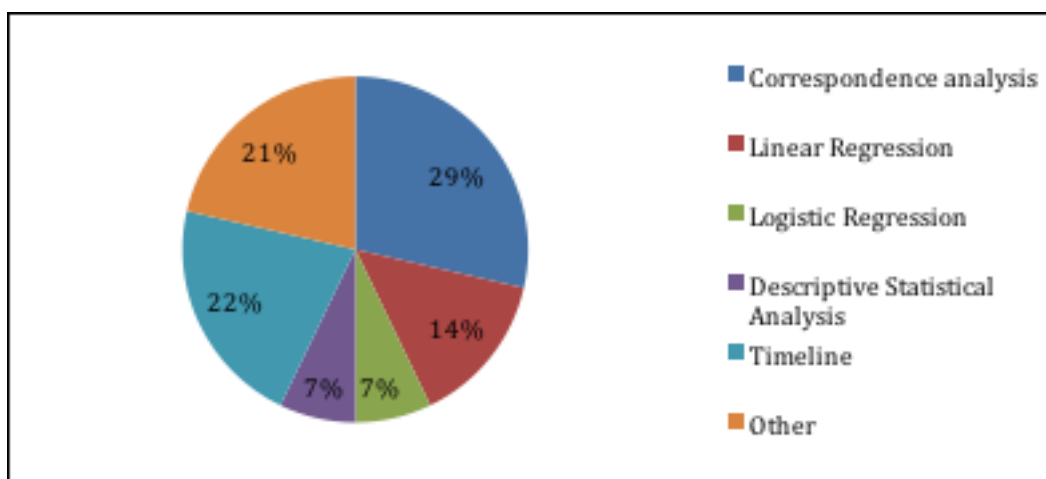
*Where did you store the open data for usage by the application?*



**Figure 17 Storage of OSD**

Both relational and NOSQL databases and RDF store are used by developers to store the open data for their application.

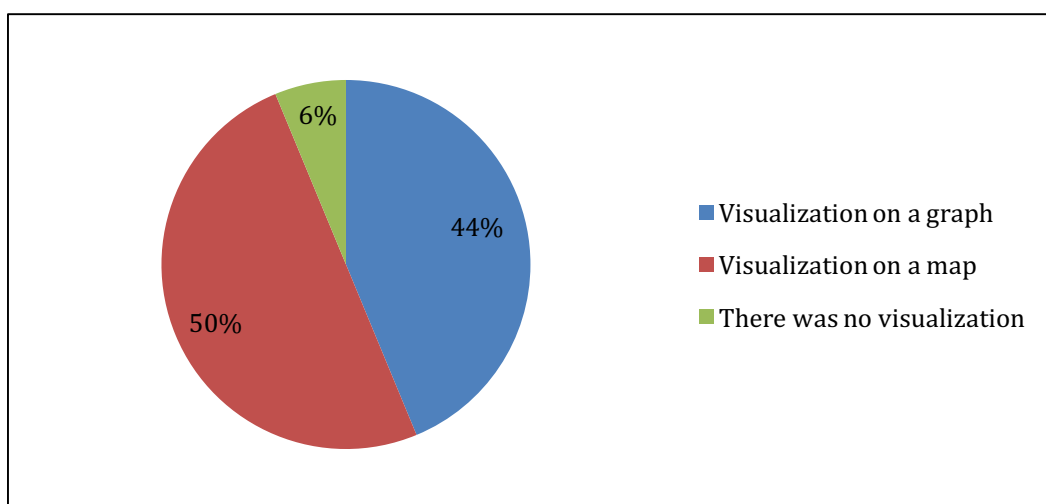
*Did you make any sort of statistical analysis in these data?*



**Figure 18 Statistical Analyses on OSD**

Most of developers made statistical analysis in the data they used. They preferred correspondence analysis at a rate of 29% and timelines at a rate of 22%. They used logistic regression and descriptive statistical analysis less at a rate of 7%.

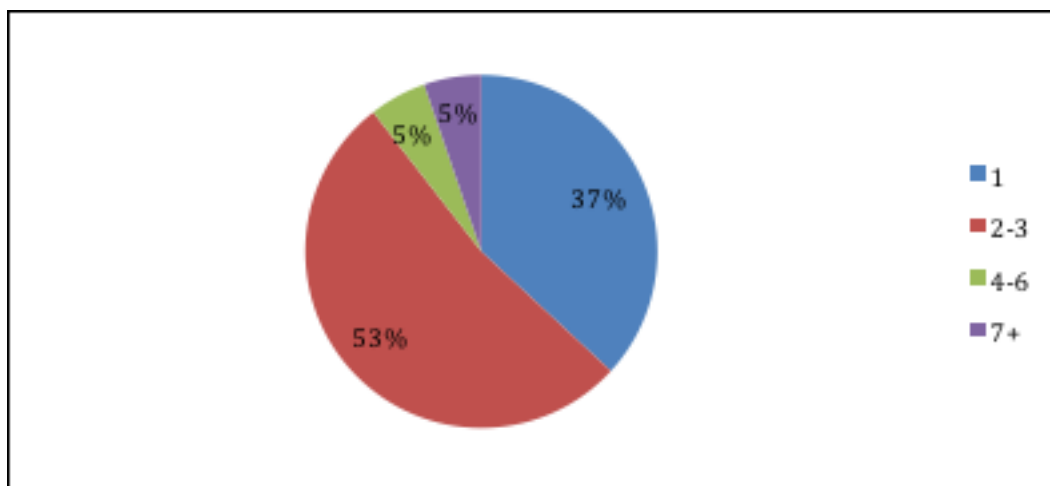
*Was there any visualization of the data? If yes, what kind of visualization did you make?*



**Figure 19 Visualisations on OSD**

The majority of developers (more than 90%) created visualizations of the data they used. Most of them made visualizations on a map at a rate of 50% while enough (44%) preferred to make visualizations on a graph.

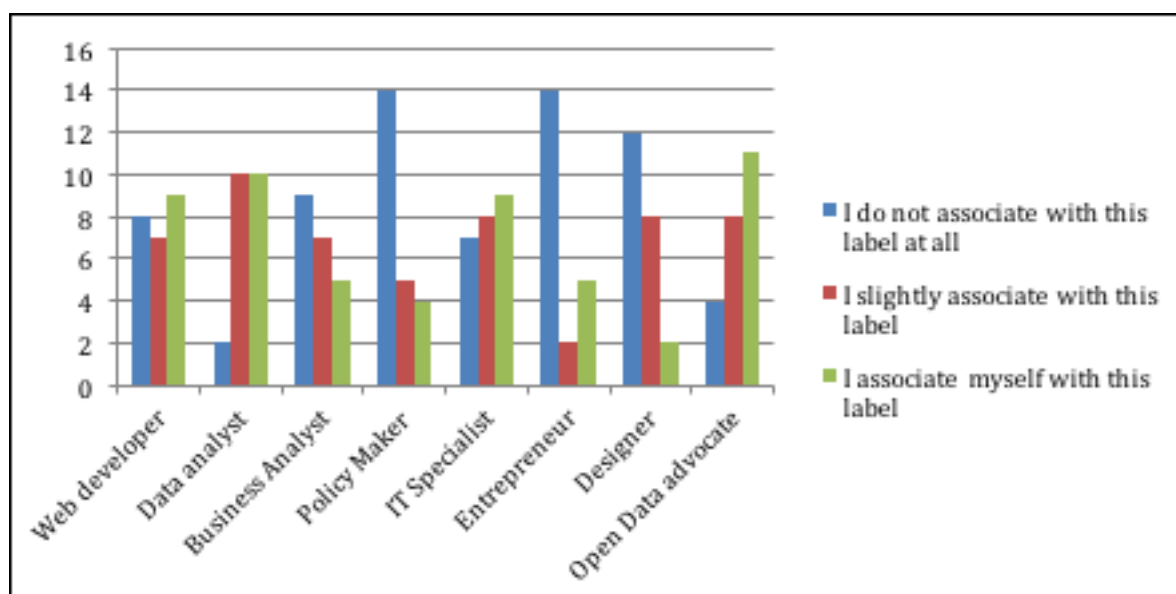
*How many individuals did you collaborate as a team in order to conclude your application?*



**Figure 20 Developing team of OSD applications**

A team of 2-3 individuals have a rate of 53% while just the developer alone has a rate of 37%. Developers prefer to get the job done by themselves or with more 2 individuals and not develop their application with a big team of 4 or 6 individuals.

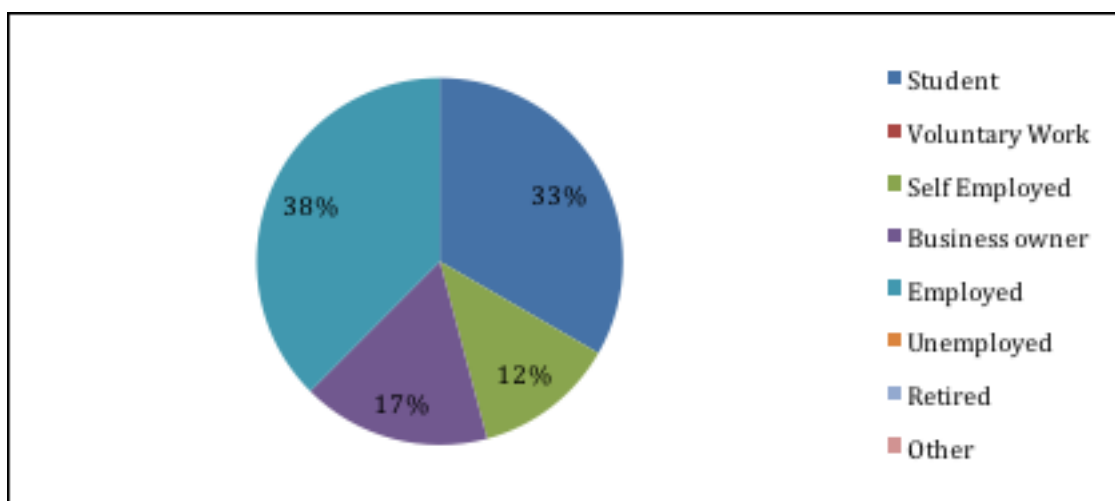
*Thinking about yourself, on a scale of 1 to 3, to what extent do you associate with the following labels?*



**Figure 21 Profession of questionnaire responders**

Web developer, data analyst, IT specialist and open data advocate are among the most popular labels which developers think they associate with. On the contrary, business analyst, policy maker, entrepreneur and designer are labels that developers think they do not associate at all.

*What is your employment status?*



**Figure 22 Employment status of questionnaire responders**

Most of developers (50%) who try to use OSD in their applications are either employed or self-employed people. Admirable is the fact that students hold such a big part among the developers list (33%). Business owners are also among them with a rate at almost 20%.



## 7 Pilots' needs on data-driven innovation

In this section we describe the results of the activities performed to identify challenges and needs of the pilot partners of the consortium. Pilot partners of the project described problems that can be solved through the exploitation of statistical data covering many different areas such as problems of public administration, businesses and citizens.

### 7.1 Pilot 1: The Greek Ministry of Interior

#### Description

The Ministry of Interior and Administrative Reconstruction is in charge of monitoring and managing an approximate number of 11.500 government vehicles which are used by all Greek Public Agencies. The data sets it possesses originate from different sources and have not yet been properly defined, structured and combined in order to be converted to meaningful information, which will facilitate internal decision making and increase transparency towards the public. Furthermore, the non-existence of structured and well-defined data obliges the Ministry to follow document-centric processes, which lack in efficiency and effectiveness and also fail to facilitate the monitoring and management of Government Vehicles. This problem can be further analyzed in the following bullet points:

- The inefficiency of accurate data regarding government vehicles
- The inability to provide accurate responses regarding certain measures of government vehicles (e.g. count of government vehicles, number of vehicles per region, etc)
- Limited ability to match the demand and offer of government vehicles
- Reduced control of government vehicles operational costs
- Limitations in policy making in the area of government vehicles
- Limited access of the public to data concerning government vehicles

The first problem will be attempted to be tackled as a short-term goal, the second bullet will be the subject of a medium-term goal and the following three problems will constitute a long-term goal of the project. The final problem will be attempted to be solved during the whole life cycle of the project, provided that Ministerial approval will be given.

#### Datasets

1. Data describing Government Vehicles
2. Data on the lifecycle of Government Vehicles
3. Data on the operation and maintenance of Government Vehicles
4. Statistical data on Greek Public Agencies and their personnel
5. Statistical data on Greek Municipalities, Prefectures and Regions (describing their population, their topography, their climate etc.)

The use of the aforementioned datasets to solve the problems of the previous section is displayed in the following table:

	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5
The inefficiency of accurate data regarding government vehicles	X	X	X		
The inability to provide accurate responses regarding certain measures of government vehicles (e.g. count of government vehicles, number of vehicles per region, etc)	X			X	X
Limited ability to match the demand and offer of government vehicles	X	X		X	
Reduced control of government vehicles operational costs	X		X	X	
Limitations in policy making in the area of government vehicles	X	X	X	X	X
Limited access of the public to data concerning government vehicles	X	X	X	X	X

1. **Data describing Government Vehicles:** Closed Data. Readily available. Data Owner: Ministry of Interior and Administrative Reconstruction

Measured Variable: Number of Government Vehicles

Dimensions:

- Count of Government Vehicles
- Count of Government Vehicles per Public Agency
- Count of Government Vehicles per Region/Prefecture/Municipality
- Proportion of the number Government Vehicles to the number of Civil servants per Public Agency
- Proportion of the number Government Vehicles to the number of Inhabitants per Region/Prefecture/Municipality

Measured Variable: Engine Displacement (in cubic centimeters)

Dimensions:

- Average Engine Displacement per Region/Prefecture/Municipality
- Average Engine Displacement according to altitude

2. **Data on the lifecycle of Government Vehicles:** Will be available when Government Vehicles' Information System will be operational. Owner: Ministry of Interior and Administrative Reconstruction
3. **Data on the Operation and Management of Government Vehicles:** Will be available when Government Vehicles' Information System will be operational. Owner: Ministry of Interior and Administrative Reconstruction

Measured Variable: Fuels Consumption

Dimensions:

- Yearly consumption of all Government Vehicles
- Average yearly consumption per vehicle type
- Average yearly consumption per region/prefecture/municipality
- Average yearly consumption according to altitude

Measured Variable: Maintenance cost

Dimensions:

- Yearly maintenance cost of all Government Vehicles
- Average yearly maintenance cost per vehicle type
- Average yearly maintenance cost per region/prefecture/municipality
- Average yearly maintenance cost according to altitude

Measured Variable: Vehicle Insurance cost

Dimensions:

- Yearly insurance cost of all Government Vehicles
- Average yearly insurance cost per vehicle type
- Average yearly insurance cost per region/prefecture/municipality

Measured Variable: Toll fee costs

Dimensions:

- Yearly Toll fee cost of all Government Vehicles
- Average yearly Toll fee cost per vehicle type
- Average yearly Toll fee cost per region/prefecture/municipality

Measured Variable: Distance travelled

Dimensions:

- Yearly/monthly/weekly/daily total distance travelled
- Average yearly/monthly/weekly/daily distance travelled per municipality/prefecture/region

Average yearly/monthly/weekly/daily distance travelled per Public Agency

### **Final Product**

The main category of target users of the final product will belong to all Greek Public Agencies who use government vehicles. These users will use the product that will be developed in order to obtain measures and reports on their use of government vehicles.

The Ministry of Interior and Administrative Reconstruction also belongs to the target users, as it will use the final product to produce managerial reports that will facilitate them in their decision making process.

The final product/service will also be targeted to citizens as users who might be interested to obtain open data regarding Government Vehicles. Additionally, companies engaging in activities relating vehicles, such as toll-management companies, technical maintenance companies and insurance companies may be interested to obtain and use the open data provided by the final product.

The final product/service will be able to combine the aforementioned data sets to produce statistics that will provide added value both for the Greek public sector, as well as for citizens and companies.

In the short term the product will be able to produce statistics on descriptive data of government vehicles, while in the medium and long-term it will produce statistics and measures on operational data and costs of government vehicles. On the long-term the final product will also be able to produce managerial reports, make forecasts and examine trends on certain measures of the entire life cycle of government vehicles, as well as examine and study causalities between the dependent and independent variables. Part of the data and their statistical processing results will also be open to the public and will be at the disposal of citizens/businesses to process them or use them in other systems or applications.

Statistical Analysis Methods that need to be applied for the final product/service to be produced:

- OLAP Analysis for the development of the first scenario (short-term goal)
- Correlation Matrix for the development of the first scenario (short-term goal)
- Regression Analysis for the development of the second scenario (medium-term goal)
- Decision Tree for the development of the third scenario (long-term goal)

Tools for the graphical representation of data include statistic software packages, such as R-project etc.

## 7.2 Pilot 2: Enterprise Lithuania – Lithuanian Ministry of Economy

### Description

Market research and decision making process problem was described as a national business problem. Entrepreneurs in Vilnius city have no information about the opportunities and competition in the areas they want to start their businesses. They need to invest a lot of resources in order to find out if their idea has any potential. Linked open statistical data can be used to simplify market research and decision-making process during the business planning stage.

### Datasets

<i>Dataset</i>	<i>Description</i>	<i>Measure and dimensions</i>	<i>Available at</i>
Advertisement permits Data RDF	Open Data	List of advertisement permits issued to businesses in Vilnius city	
Permits - Hygiene Passports RDF	Closed Data	List of permits - hygiene passports issued to	

		businesses in Vilnius City	
--	--	-------------------------------	--

### Final Product

The final product/service will let the users to navigate the Vilnius city map and see all active businesses from up to 5 most popular business areas in the city. The target users of this product/service will be entrepreneurs, who are planning to start or expand their businesses in Vilnius city. Visualization tools are needed to develop the final product/service.

## 7.3 Pilot 3: Trafford Council

### Description

A distribution problem of Job Centers in England was described as a national problem of public administration and citizens' problem. The Department for Work and Pensions (DWP) is a central government department, who maintains around 800 Job Centre Plus in England where people can claim out of work benefits, receive advice on CV-writing and interviews, and apply for jobs. The DWP is reviewing the distribution of these Job Centers and wants open statistical data to help this job done.

### Datasets

<i>Dataset</i>	<i>Description</i>	<i>Measure and dimensions</i>	<i>Available at</i>
Demographics	Counts of people by ethnicity, age, gender, by Lower Super Output Area (LSOA)	Open Data	NOMIS web portal
Benefits	Counts of people claiming different benefits by broad age band, how long they've been claiming for, by LSOA	Open Data	NOMIS web portal
Qualifications	Counts of people with highest qualification level, by LSOA	Open Data	NOMIS web portal
Businessess	Counts of businesses by type, size, number of	Open Data	<ul style="list-style-type: none"> <li>Business demography dataset</li> </ul>

	births and deaths (of businesses), by Local Authority area		<ul style="list-style-type: none"> <li>Companies House open data</li> </ul>
16-18 year olds not in education, employment or training	Counts of people by Middle Super Output Area (MSOA)	Open Data	data.gov.uk
Physical assets	Details of public sector assets by address	Open Data (may be difficult to get used)	Transparency code
Voluntary groups	Details of voluntary/charity groups by address	Open Data	Trafford Innovation and Intelligence Lab portal

### Final Product

The final product/service will allow exploration of data, especially from a spatial point of view - allowing people to see where needs are greater, and where there are available assets or groups who could support an alternative model of delivering Job Centre Plus services. The tool will also provide a dashboard for decision-makers to get the most up-to-date information about worklessness that they can explore, drill-through, etc. The target users of this product/service will be local DWP teams responsible for reconfiguring the Job Centers, in conjunction with Local Authority leads for worklessness. Ultimately, the public are the users of the Job Centre Plus. Visualization tools to include JavaScript, HTML and CSS, leaflet mapping library, d3 visuals, and custom JavaScript tools are needed to develop the final product/service. Publish My Data linked data platform, using APIs from NOMIS and other Central Government portals and upload pipelines for data without APIs will be needed too

## 7.4 Pilot 4: The Flemish Government

### Description

A problem about reported and permitted emission was described as a national problem of public administration, business and citizens' problem. Someone wants to compare reported and permitted emissions or emission ratios between geographical regions. He maybe wants to link emission data with population density or integrate the emission data with health data and look for correlations.

### Datasets

<i>Dataset</i>	<i>Description</i>	<i>Measure and dimensions</i>	<i>Available at</i>
----------------	--------------------	-------------------------------	---------------------

Environmental permissions	Open Data*		
Reported emission data	Open Data*	Dimensions are: reference year, emitted substance (pollutant), reference area. The measure is the quantity.	
Geographic regions code list	Open Data*		
Nace code	Open Data*		
Healthdata	Closed Data*		
Business registries	Open Data*		

\*All data is open after removing personal data from the dataset and removing documents that contain descriptions of specific industrial processes.

### Final Product

The target users of this product/service will be the citizens as they would see what pollutants in what amounts are emitted near where they live. The target users are also the legal entities as they would see the benchmark with other legal entities in the same sector (How ecological am I as a company compared with others). Finally, government is also target user of the final product/service as it can see if extra pollution be permitted in a certain area based on already permitted data and based on reported emissions. Ontop, tripple store, r-statistics and visualization library are needed to develop the final product/service.

## 7.5 Pilot 5: The Marine Institute

### Description

A search and rescue operation problem was described as a cross country problem of public administration, business and citizens. The rescue team wants to know the current conditions in the waters around the coastline. A member of the team wants to return information, such as geo-located photographs, to team's coordinator so he can be kept up to date of the search team's location and conditions. In addition to the public authorities, a member of the public involved in searching the coastline. The volunteers want to have access to the same apps and much of the same data as the authorities, but some information may not be available to them. The team's coordinator review the information collected by the app after each rescue to build up dataset which allows him to develop local search and rescue policies.

### Datasets

<i>Dataset</i>	<i>Description</i>	<i>Measure and dimensions</i>	<i>Available at</i>
Met ocean observations and forecasts	Open Data	<ul style="list-style-type: none"> <li>• temperature salinity</li> <li>• wind speed and direction</li> <li>• wave and current speed and direction at various locations and depths/heights across a bay</li> </ul>	erddap.marine.ie
Vessel locations in real-time across a bay	Closed Data		Commercial operators such as Marine Traffic
Location of a person through real-time	Closed Data		Telecommunications companies

### Final Product

The final product/service will provide a tool in which search and rescue personnel can identify the key areas to search for a casualty in the water for rescue or recovery. This may also include onshore search parties who can provide their location and coastal imagery. The target users of this product/service are search and rescue services. Statistical analysis of data on entry and rescue/recovery locations, visualization of forecast model outputs, visualization of traffic situation in the bay and predictive analysis of particle tracking are the tools are needed to develop the final product/service.

## 7.6 Pilot 6: The Estonian Ministry of Economics

### Description

A problem about buying or renting a flat was described as a national problem of public administration, business and citizens' problem. If someone wants to buy or rent a flat, a house, office premises or a land in Estonia, he usually needs to go to real estate websites to get the general information (total area, built year, building material, ownership, etc.). If he wants to know specific details or learn about restrictions concerning the flat, house, office premises or land as well as the area in which the real estate is located, he needs to visit numerous different databases owned by different public authorities. The problem is that an average citizen does not know those databases exist. However, if a web page/platform could provide the information someone needs would become available, this problem would not be exist.

### Datasets



<i>Dataset</i>	<i>Description</i>	<i>Measure and dimensions</i>	<i>Available at</i>
Building Register	*	<ul style="list-style-type: none"> <li>Buildings in communes in Estonia</li> <li>Building, address, commune</li> </ul>	Estonian state information system's secure data exchange layer (X-Road)
Land Register	*	<ul style="list-style-type: none"> <li>Land plots in communes in Estonia</li> <li>Plot, address, commune</li> </ul>	X-Road
Land Cadastre	*	<ul style="list-style-type: none"> <li>Cataster in commune in Estonia</li> <li>Cataster unit, description</li> </ul>	X-Road
Land tax	*	<ul style="list-style-type: none"> <li>land tax paid per plot</li> <li>Land tax per plot per year</li> </ul>	X-Road
Estonian Land Board registries	*	<ul style="list-style-type: none"> <li>Real estate valuation of building over time/history</li> <li>Building/land plot, price, date</li> </ul>	X-Road
Insurance costs per building	*	Cost, building	X-Road
Utilities in the house	*	Building, costs per sqm winter/summer	X-Road
City and local municipalities registries	*	<ul style="list-style-type: none"> <li>Planning and future building information</li> <li>Land plot</li> <li>Planned buildings</li> </ul>	X-Road
Technical Regulatory Authority registry	*	Internet, telephone, TV connection	X-Road

availability per building			
City and municipalities Environment Departments registries	*	Environmental information (noise, storm water, air pollution)	X-Road
City and local municipalities registries	*	Public transport, timetables, tickets	X-Road
Land Board Geographical Information System	*	Distance from kindergartens, schools, hobby education, pharmacies, hospitals, churches, dog parks, playgrounds, malls, etc.	X-Road
Police and Border Guard Board registries	*	Crimes in that area (physical violence, robberies, theft from residential premises, car theft, theft from cars)	X-Road
Tallinn Waste Centre	*	Closest recycling points, deposited packages kiosks, waste transport	X-Road
News about the nearby area	*	What has happened (example: festival)/is going to happen (new mall is going to be built or cultural happenings)	X-Road

\*most of them are closed data at the moment but access to data can rather easily be requested

## Final Product

The final product/service would be a search engine which will be able to show a large amount of useful information about the flat, house, office premises or land that different types of users are interested in. This information would normally have to be searched for from different registers, so the main benefit would be substantial time savings for users. The search engine would show different information on a map platform, providing visuals and information on one screen. It would be possible to zoom the map in and out, use different layers, filter information, click on additional links, etc. The target users of this product/service will be all citizens who need to rent or purchase new real estate or land, real estate broker, real estate developers, investors, notaries and government officials who are responsible for urban planning and “long range developments/plans” who need information on trends to improve the area (better living area – more taxes, more taxes – more stuff the city can do). The platform should foremost be able to visualize statistical and other kinds of data in one map solution. The development of this service also requires opening up and linking data from disparate sources.

## 8 Data Infrastructures: State-of-the-Art

In this section, we provide a review of the linked and open data infrastructure. We start by outlining the principles for opening data and discuss the various aspects of open data infrastructures. We conclude that infrastructures are diverse and can support a variety of activities during the open data life cycle.

### 8.1 Maturity Level

LOSD is in its infancies and organizations try to move towards higher levels of maturity. In psychological area, maturity means the ability to respond to the environment in an appropriate manner. Unlike the direct definition from maturity level, there is no LOD maturity model found in the literature review, although there are in related areas.

Kalampokis et al. (2011) proposed a model to evaluate the maturity level of OGD based on two main approaches: Technological and Organizational. The Technological approach was divided in two dimensions: Downloadable files and Linked Data. The factors to explain the Downloadable files were: Proprietary and desktop-centric formats, Machine-readable formats, and, Machine-readable formats using open standards. The LD dimensions was divided in two factors: LD principles and Linking available data. The Organisational approach was based on Direct and Indirect dimensions.

Using this model, it becomes possible to determine the existing LOD level of the organisation and provide a roadmap to "evolve and reach" the considered highest level of OGD Maturity Level, the Indirect Provision of Linked Data. To explain the model, Kalampokis et al. (2011) using an example of accessing RDF data via API. Further, they explain if the organisation offers OGD via downloadable files on a direct way, it was on a "Repository of downloadable files", also known as "dumps" (Jentzsch et al., 2009). If it was on Linked Data approach and direct, the level of "Direct Provision of Linked Data", such as a dump but on Linked Data formats such as RDF. If Still downloadable files but indirect, they considered a level of "Registry of Downloadable Files", what means a catalogue of datasets with URL for the owners' pages. Figure 23 summarises the OGD model of Kalampokis et al. (2011).

Technological Approach		Organizational Approach	
		Direct Data Provision	Indirect Data Provision
Downloadable Files	Proprietary and desktop-centric formats	Repository of Downloadable Files	Registry of Downloadable Files
	Machine-readable formats		
	Machine-readable formats using open standards		
Linked Data	Linked data principles	Direct Provision of Linked Data	Indirect Provision of Linked Data
	Linking available data		

Figure 23 - OGD Classification Scheme  
Source from Kalampokis et al. (2011)

Lee & Kwak (2012) proposed a model for open government that consists of five maturity levels: initial conditions (Level 1), data transparency (Level 2), open participation (Level 3), open collaboration (Level 4), and ubiquitous engagement (Level 5). The focus of their research was on transparency and collaborative public engagement on the health of United States of America (USA) based on the social media interactions. Indeed, this is not related on the topic, Figure 24 shows a trend that can be used for a Linked Data Maturity Level taking in consideration the scheme of benefits/challenges and organisational complexity.

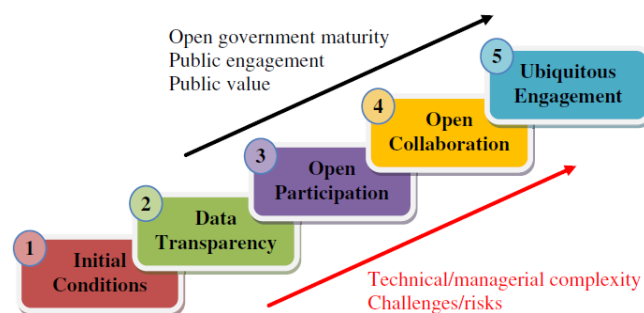


Figure 24 - Open Government Maturity Model (OGMM)  
Source from Lee and Kwak (2012)

Janssen et al. (2014) proposed a maturity level for interoperability of Big and Open Linked Data (BOLD). The maturity level takes in consideration stages, describing them with specific needs and requirements to evolve, such as a roadmap for implementation. Defining their data interoperability maturity level, expanding capabilities, and developing a data portfolio can help guide organizations as they move toward ubiquitous information sharing. This is helpful to create a Linked Data Maturity Model (LDML) using for example part of the benefits, risks and challenges found on literature review summarised on Figure 25.

	Stage 0: Independent	Stage 1: Ad hoc	Stage 2: Collaborative	Stage 3: Integrated	Stage 4: Unified
<b>Description</b>	An organization has no strategy and does not think about the impact of releasing or sharing its data.	Ad hoc arrangements are agreed upon for data use and interoperability within an organization.	Strategies about release and use of open data are identified, and roles and responsibilities are defined within an organization.	An organization implements shared goals and value systems with another organization based on common understandings and mutual desire for data interoperability.	An organization defines strategy, shared organizational goals, value systems, data portfolios, and knowledge bases.
<b>Levels of interoperability supported</b>	Technical	Technical and limited syntactic	Technical and syntactic	Technical, syntactic, and semantic	Technical, syntactic, semantic, and pragmatic
<b>Main characteristics</b>	Focus is on realizing benefits by developing certain applications; communications about data collection, linkage, and processing are unstructured and carried out through meetings, emails, and phone calls; no formal procedures or support are in place.	Information is used without formal governance procedures, standards, planning, or infrastructures; some overarching goals are in place, but a detailed strategy is lacking.	Governance mechanisms and roles and procedures are all in place for data acquisition, processing, and distribution; datasets can be related to other datasets to create value; strategy includes development of capabilities to ensure organizational readiness to interoperate.	Infrastructure, agreements with data providers, and assessments of information use and impact are in place; metadata is shared among organizations to enable linking and combining of data; data is shared in large volumes; manual processing is used for more complex operations, handling exceptions, and integration of heterogeneous data sources.	Governance for data portfolio use is defined; capabilities for the discovery, assessment, and integration of new data sources for a certain architecture within a short time frame are developed; data is viewed as an essential asset by the organization, and business value is created by quickly acquiring and using the data; data portfolio instruments are used to manage data quality, legal status, and permitted uses.

Figure 25 - Interoperability maturity stages for Big and Open Data  
Source from Janssen et al. (2014)

### 8.1.1 Linked Data Maturity Level

This LDML presented on this section is not part of the State-of-Art but it was decided to be written due the possibility of earning time and receiving inputs and review from partners to the WP2 – Framework Creation and the WP 4- Pilots Planning Evaluation.

To enable our evaluation on the OpenGovIntelligence pilots, we considered to propose a LDML observing related maturity models on OGD (Kalampokis et al., 2011; Lee & Kwak, 2012; Janssen et al., 2014). From the maturity levels identified on the literature review of OGD maturity level, the Figure 26 summarized the proposed Open Linked Data Maturity Level.

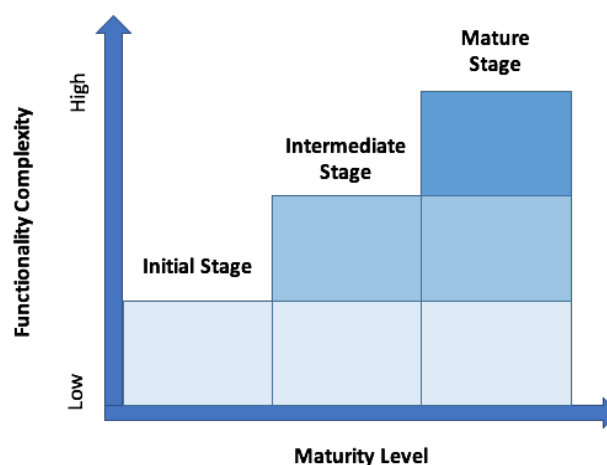


Figure 26 - Linked Data Maturity Level

The questionnaire presented on Table 3 can be used for evaluation of LDML from four main aspects: Data, Technical and Infrastructure, Organisational, Financial and People. Each of them have specific questions. Before the questionnaire, it is recommended to have some prerequisite questions for filtering cases and initiatives that are not yet Linked Data.

Using the questionnaire, the organisation can identify their current level of LOD, and create a roadmap to evolve, reach benefits and also overcome challenges or risks identified. For example, if the majority of the answers are on the initial stage then the organisation want to reach the intermediate and mature stage. This table will be improved and applied later in the project to diagnose each pilot in detail and suggest a strategy to overcome the challenges and risks identified in accordance with the objectives of pilots.

Taking consideration of IT governance maturity models already established on literature and market good practices we could bring some references to influence proposing questions for identify maturity level.

Table 3 - Proposed Questionnaire for Evaluating the Open Linked Data Maturity Level of Pilots

Dimensions / Stages	Questions	Initial Level	Intermediary Level	Mature Level
<b>Prerequisite</b>	<b>Which is the Organisation level of 15 Open Data principles?</b>	<b>At least half of the 8 first principles</b>	<b>At least 12 principles</b>	<b>All the principles</b>
	<b>Which is the Organisation level of Linked Open Data principles (Berners-Lee's 5-stars)?</b>	<b>At least 3rd Level (CSV)</b>	<b>At least 4th Level (RDF)</b>	<b>5th Level (Linked RDF)</b>
<b>Data</b>	Does the organisation use web ontology language or standards for Linked Data (LD) publishing?	YES	YES	YES
	Does the organisation use always the same ontology, web semantic and standards for LD publishing?	NO	50% or less of datasets	100% of datasets
	Does the organisation have metadata for the Linked Data?	NO	50% or less of datasets	100% of datasets
	Does the organization provide data license for each published datasets?	NO	YES	YES
	What is the level of datasets structure before Linked Data?	Unstructured	Semi-Structured	Structured
<b>Technical and Infrastructure</b>	Does the Organisation use Open Data Architecture?	NO	50% or less of systems	100% of system architectures
	Does the organisations use standardised and interoperable language of coding on systems and datasets?	NO	Mix of interoperable but no standardised languages	Java
	Does the Organisation have proper storage and processing computers (Big Data Environment)?	NO	NO	YES
	Does the Organisation have any data cube for Linked Data creation?	YES	YES	YES

	Do the organizations use common tools for the search functionality of their datasets?	YES	YES	YES
<b>Organisational, Financial and People</b>	Does the organisation have budget for evolve the collection, storage and processing toward Big Data?	NO proper budget.	Budget for partial project.	Budget for dull project.
	Does the organisation have a team of Data scientists for Linked Data publishing?	NO	YES	YES
	There is a diversity of backgrounds on data scientists team?	NO	No plurality of backgrounds and skills	Diverse of backgrounds and skills
	Which is the Level of governance for OGD and Linked Data implementation (Web-ontologies)?	Low	50% of time or less people use the OGD and LD properly	100% of time people use OGD and LD principles properly

## 8.2 Data Infrastructures

### 8.2.1 Principles for opening data

In 2007 a group of experts in Open Data had a meeting to decide upon the principles of Open Government Data (OpenGovData.org 2007). The eight initial principles were:

1. **Complete:** All public data are made available with no subject to valid privacy, security or privilege limitations.
2. **Primary:** Data are collected from the source, with the highest possible level of granularity, not in aggregate or modified forms.
3. **Timely:** Data are made available as quickly as necessary to preserve the value of the data.
4. **Accessible:** Data are available to the widest range of users for the widest range of purposes.
5. **Machine Process-able:** Data are reasonably structured to allow automated processing.
6. **Non Discriminatory:** Data are available to anyone, with no requirement of registration.
7. **Non-Proprietary:** Data are available in a format over which no entity has exclusive control.
8. **License-Free:** Data are not to be subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed.

Later, seven more principles were added on the eight principles list:

9. **Online and Free:** Data should be accessible on the Internet and also not a subject of any kind of costs.
10. **Permanent:** Data should be at the same place and with same format as long as possible.
11. **Trusted:** Data should be digitally signed or should include attestation of publication of datasets. This can bring more trustworthiness to the users of OGD.
12. **A Presumption of Openness:** Effects of the Freedom of Information Act (FOIA), data should have a comprehensive information management.
13. **Documented:** Datasets should have metadata which are the data that explain what the specifications of all the data on datasets are. An example would be the speed of buses measured in kilometres per hour rather than miles per hour.



14. **Safe to Open:** Avoid unsafe formats of data such as executable (.exe). Hackers can change datasets and attack vulnerable people via open data, thereby reducing trustiness on the OGD portals and governmental datasets.
15. **Design with Public Input:** The public should be able to choose the best format or way to access datasets. This offers the freedom of choice for users of the same datasets.

### 8.2.2 The rise of Open Data Infrastructures

The development of open data infrastructures is driven by open government policies. In Europe, the implementation of “Open Government Data” public policies started in 2003 with the publication of Public Sector Information Directive (PSI Directive). The Directive was finished in 2013 with the proper update of Open Government Data principles (Janssen, 2011).

In the United States of America (USA), the Open Data received a boost since the Obama Memorandum of Open Government, including Open Data format (Coglianese, 2009; Obama, 2009). In 2011, the Open Government Partnership (OGP) was created giving a push to the international discussion about OGD and the implementation of new projects around the world (Cretu & Manolea 2013; Manolea & Cretu, 2013). In order to participate in the OGP, an Action Plan was deemed mandatory. This plan should entail strategies for the open government and governmental data, further creating channels for public participation. All the plans were evaluated by the civil society with the aim of reducing bias of governmental evaluation of their own projects’ results.

Recently, in 2013, the Group of Eight 8 (G8) created the Open Data Charter, emphasizing the open government and open data format publishing (Chan, 2014) and following almost all the rules from the aforementioned described initiatives.

The implementation of OGD portals and their initiatives around the world prompted people, enterprises, journalists and the public sector itself to consume data, hence creating the so-called “OGD ecosystem” (Ding et al., 2011). This reuse of data has various objectives. For the civil society, the common usage is to promote transparency as “*infomediary*” and intermediary between government and people (Magalhaes et al., 2013; Janssen & Zuiderwijk, 2014), to create transparency portals and social web applications (Matheus & Janssen, 2013) or/and to provide accountability of governmental actions (Chun et al., 2010).

Enterprises are commonly using Open Data and Open Architectures (Linked Data) for better communication and data exchange with governments (Wood et al., 2014) or for creating applications (Ubaldi, 2013). Journalists also have been using data to create more trustiness and interactive stories concerning the government or the society with the use of OGD (Gray et al. 2012; Matheus et al., 2014).

### 8.2.3 Infrastructure Operating in an Open Data Ecosystem

Zuiderwijk et al. (2014) identified an ecosystem with multiple dimensions. The dimensions that should be taken into consideration, while creating an open data ecosystem, are (a) the tools and services that must be provided or used, (b) the contextual level of data producer, and (c) data users. The factors, influencing the three dimensions, help Public Sector to have better results on Open Data initiatives.

For example, the usage of Application programming Interface (API) for publishing data is faster than dumps of the datasets and more secure in comparison to direct access of government databases.

While considering other dimensions, Dawes et al. (2016) explained that, in order to create an open data ecosystem for the government, it is necessary to identify the beneficiaries and their objectives upon using the data. This will influence, via feedback and communication, the OGD providers. In addition, advocacy and interaction were identified. The factors considered in this initial design, that can influence the motivation of OGD development, are openness, data-driven products and service.

The strategies of publishing explained in Zuiderwijk et al. (2014) can be identified on the high-level Figure 27 of Dawes et al. (2016). The strategies identified are: legal framework, priorities of the government and resource allocation, or agenda setting (Jetzek, 2015; Kingdon 2003), and organisational relationships (internal and external).

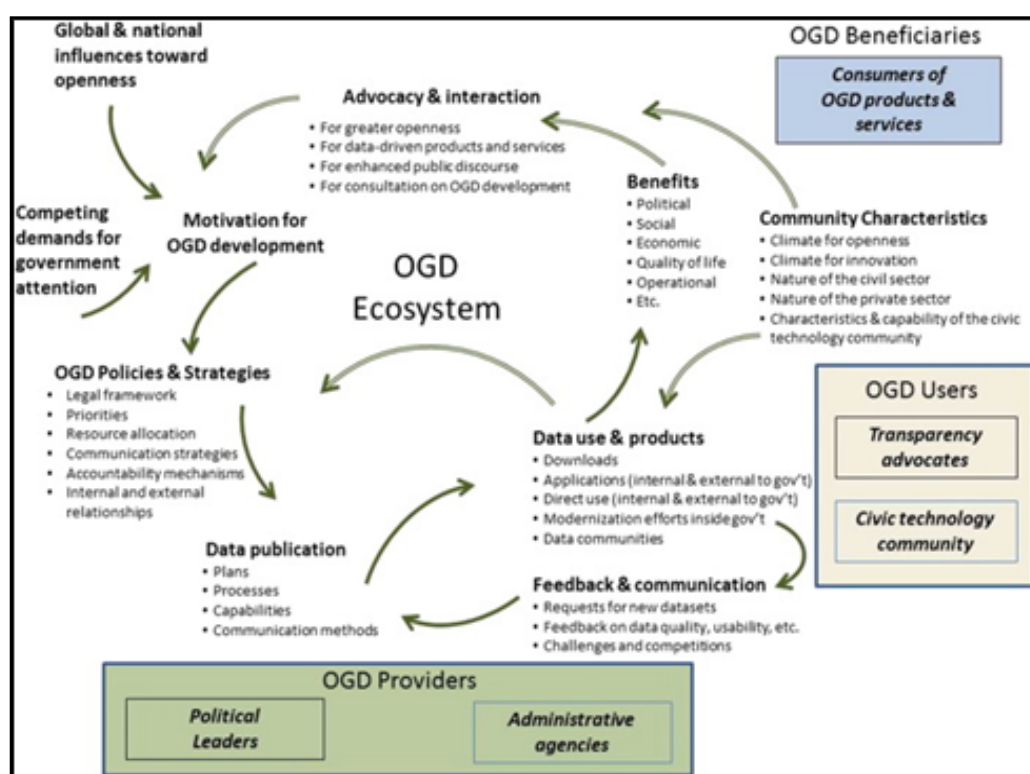


Figure 27 - Open Data Ecosystem  
Source from Dawes et al. (2016)

#### 8.2.4 Open Data Infrastructures (ODI)

Open Data are considered to bring several benefits. However, it is unlikely to create, maintain and reach all the potential benefits without the Open Data Infrastructures (ODI). This section enlists ODI and how they are used in governmental, societal and private sector initiatives. This section also provides an answer to the objectives 1 and 4 mentioned earlier at the section.

For Zuiderwijk et al. (2013) the ODI is seen as synonymous to the Open Data Portal (ODP). The categories and range of capabilities observed were general functionalities, such as the creation of

Open Data, Opening-up Data, Finding Open Data, Usage of Data and Discussing Open Data. For Zuiderwijk et al. (2013), there were other categories and characteristics of ODI at ODP, such as access, sharing, navigation of ODP, uploading, downloading, data quality, analysing, visualising, linking and combining open data, collaborating, support and help, and feedback. Matheus & Janssen (2013) provided insights for transparency via opening-up data and identifying the dimensions of interpretation and accessibility. The

Table 4 summarizes the others sub-dimensions identified and presented in Figure 28.

Table 4- Factors that influence Open Data Infrastructures

<b>Dimension</b>	<b>Sub-Dimension</b>	<b>Explanation</b>
<b>Interpretation</b>	<b>Interpretation of data</b>	Easier interpretation of data results in higher transparency
	<b>Examples</b>	Presence of examples of the website product, the higher has a positive influence on interpretation.
	<b>Simple Language</b>	Simple language has significant positive influence on transparency.
	<b>Data Quality</b>	Higher information quality has a significant influence on interpretation.
		Higher updated information has a significant influence on data quality.
		Higher data completeness has a significant influence on data quality.
		Higher data accuracy has a significant influence on data quality.
<b>Accessibility</b>	<b>Accessibility of data</b>	Higher accessibility has a significant positive influence on transparency.
	<b>Data Overload</b>	Data overload has a significant negative influence on accessibility.
	<b>Adhesion to Standards</b>	Adhesion to standards has a significant positive influence on accessibility.
	<b>Unified Technology</b>	Unified use of technology has positive influence on accessibility.

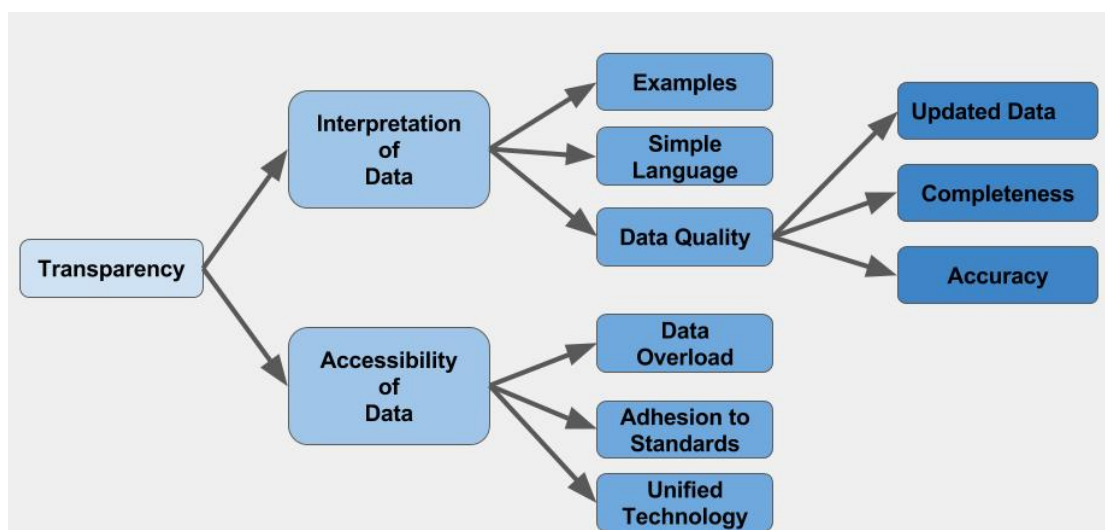


Figure 28 - Factors that influence Open Data Infrastructures  
Adapted from Matheus & Janssen (2013)

Furthermore, research, provided by Zuiderwijk et al. (2013), extends the ODI framework. For example, Jetzek (2015) pointed out that there is no “free data”, just “free and open access” as well as the costs for opening-up Open Data. The organization is paying, in any format, to collect, store, clean and publish the data sets. In addition, there are legal barriers to access the data, such as geographical maps of cities. To overcome those issues, Jetzek (2015) identified that a legal standardization of publishing data using creative common licenses is recommended. Next, Jetzek (2015) also identified the use of web-semantic and web-ontologies to address the interoperability issues between systems and data sets. According to Janssen et al. (2003), while discussing the interoperability between systems and data sets for the provision of electronic government (E-Gov) services, web-semantic and web-ontologies are enabled by Open Architecture. Last, Nushi et al. (2015) complemented this research with the need of spatial data to improve the connections and visualizations.

There is a variety of objectives of ODI. The outcomes identified for OD usage are commonly based on four axes (areas or objectives) that are summarized in Table 4:

Table 5 - Objectives of ODI

Outcome Axe of Open Data Usage	Sources Identified
<b>Transparency, Accountability and Anti-Corruption</b>	Peixoto (2013), Zuiderwijk-van Eijk & Janssen (2015), Matheus & Janssen (2016)
<b>Participation and Collaboration</b>	Zuiderwijk-van Eijk & Janssen (2015), Craveiro et al. (2016)
<b>Economic Growth and Innovation</b>	Janssen (2012), Janssen et al. (2012), Ubaldi (2013)
<b>Organisational Management and Improvement of Service Delivery</b>	Robinson et al. (2009), Chun et al. (2010), Shadbolt et al. (2012), Yang & Wu (2016)

ODI contain a mix of functionalities. Some portals have complex functionalities, whereas others are simple. We reviewed the literature and identified the functionalities of ODI as listed in the table below.

Table 6 - Overview of Requirements and Functionalities for Open Data Infrastructure

OGD Usage Categories	Functions	Description	Source
General, creation and opening-up of data	<b>Long-Term platform</b>	Platform should respect OD principle for long-term platforms (permanent).	OpenGovData.org (2007)
	<b>Support Multilingual</b>	Countries with more than one language need to provide multi-language support. Globalization also pushes English as second language.	Hillier (2003), Halb et al. (2010), Lehmann et al. (2015)
	<b>Registration Login</b>	Login system for users with possibility to add new features in accordance with needs.	Nevarez & White (1998)
	<b>Upload and edit new data sets</b>	After login, users of ODP and owners of data sets can freely upload and edit new data.	Charalabidis et al. (2014)
	<b>Support for publication (guidelines)</b>	Examples of how to use data sets and frequently asked questions (FAQ) can increase chances of usage.	Matheus & Janssen (2013)
	<b>Format of Data sets</b>	Data should be on the last level of 5 Stars Scheme (Linked RDF) and option for download and access on diverse granularity.	Berners-Lee (2009)
	<b>Metadata presence</b>	Respect to OD Principle (documented) and helps users on usage of data sets.	OpenGovData.org (2007)
	<b>Data creators' information</b>	Owners of data sets should be informed to enable openness of more data sets.	OpenGovData.org (2007)
	<b>Version Management</b>	Data should have a functionality showing the versions to users (history).	Matheus & Janssen (2013)
	<b>RDF high quality</b>	Reaching the last stage of 5 Star Scheme give flexibility and agility on usage of data.	Berners-Lee (2009)
	<b>Linked Open Vocabularies</b>	Following the common Linked Vocabularies influence on the quality of usage.	Bizer, Heath et al. (2009)
	<b>Linked Statistical Data Vocabulary</b>	Data sets have specific vocabularies. Statistic vocabulary influence on quality of usage.	Norway (1998)
	<b>Linked Geo Data Vocabulary</b>	Data sets enriched with standardised geolocation can help quality of usage.	Auer et al. (2009), Stadler et al. (2012), Nushi et al. (2015)
Finding, searching and accessing Open	<b>Tags based on web semantic</b>	Presence of tags can help quality of usage due organisation and visualisation at glance.	Sigurbjörnsson & Van Zwol (2008)
	<b>Data overview</b>	Providing overview of datasets can help usage of data sets.	Sigurbjörnsson & Van Zwol (2008), Zuiderwijk et al. (2013)

<b>Data</b>	<b>Structured Search Functionality</b>	Search functionality based on categories of data sets such as topic, owner, date, etc.	Tjondronegoro & Spink (2008)
	<b>High level of user experience design</b>	Design of website can help usage of data sets.	Garrett (2010)
	<b>FOIA channel for not found dataset</b>	FOIA channel for demanding officially datasets that were not found on the ODP.	Hunnius & Krieger (2014)
	<b>Channels for data consumption</b>	Option for channels to access data such as API, Dump, JavaScript Object Notation (JSON), Web Map Service (WMS) etc.	Berners-Lee (2009)
	<b>Free access</b>	Free of captcha features and usage of re-captcha	Von Ahn et al. (2008)
	<b>Free cost to access</b>	No cost to access data as principle of OD.	Bizer et al. (2009)
<b>Usage of Open Data</b>	<b>Simple Analysis features</b>	Regular analysis on the web browser (graphs, table and maps)	Winn (2013)
	<b>Data Cubes for Analysis</b>	Advanced analysis will require complex data cubes and their functionalities.	Klímek et al. (2016)
	<b>Linking and Combining Data</b>	Open or linked format enable combining data sets for several purposes (research, analysis, accountability, transparency, advocacy, etc.)	Janssen et al. (2012), Shadbolt et al. (2012)
	<b>Diverse options of visualisation</b>	Diverse options of visualisation, including dashboards at glance with graphs, maps and interactive tables.	Maheshwari & Janssen (2013), Maheshwari & Janssen (2014)
	<b>Cleansing and filtering Data</b>	Option for filtering and cleansing the data before analysis and visualisation.	Auer et al. (2007)
	<b>Curation of data sets</b>	Curation of data based on the cleansing process.	Janssen et al. (2012)
<b>Discussing and Feedback Process of OD and ODP</b>	<b>Channel for qualitative feedback</b>	Wiki with forum and mail box functionality to receive and discuss complains about data sets.	Zuiderwijk et al. (2013)
	<b>Forum for discussion of data</b>	Specific discussions can be done on the page of data set.	Zuiderwijk et al. (2013)
	<b>Channel for quantitative feedback</b>	Facebook like button feature.	Gerlitz & Helmond (2013)
	<b>Social Media share buttons</b>	Social media share functionality for Twitter, Facebook, etc. for data sets and analysis.	Kaplan & Haenlein (2010)
	<b>List of applications created with OD</b>	List with projects that used OD as source.	Matheus et al. (2015)
	<b>Upload and linking derived Data sets</b>	Creating new data sets can be uploaded and linked to original data.	Zuiderwijk et al. (2013)

### 8.2.5 Open Data Lifecycle support by infrastructures

Most infrastructures support only one part of the open data lifecycle. Attard et al. (2015) proposed that open data in government follow a lifecycle based on three main stages: pre-processing, exploiting and maintenance. Each stage has one or more steps that are represented at Figure 29 and described in Table 6 - Overview of Requirements and Functionalities for Open Data Infrastructure.

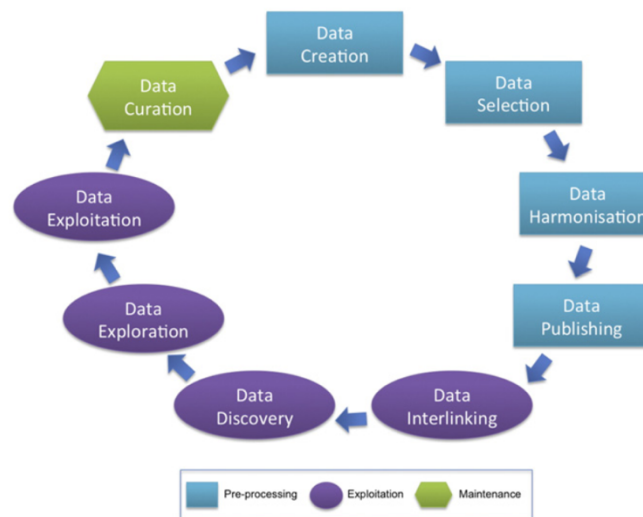


Figure 29 - Open Government Data Lifecycle (taken from Attard et al. (2015))

Lapi et al. (2012) identified a generic scenario for Open Data provision taking in consideration the process flow, the systems and the actor or domains involved. The Figure 30 summarises the generic scenario. Based on this scenario is possible to identify that Open Data has a flow with 7 steps: data collection; internal data management; data publication; data enrichment; data access; data integration into Apps; Data Usage. The Data usage will influence the Data collection, internal data management and data enrichment, as a “feedback” system for Open Data.

Hunnius and Krieger (2014) believe that there is no discussion a priori of which data will be open and formats will be chosen. Moreover, they highlighted the issues about licenses, that influence the usage of open data. They suggest that technical aspects and political aspects should be considered in the Open Data lifecycle.

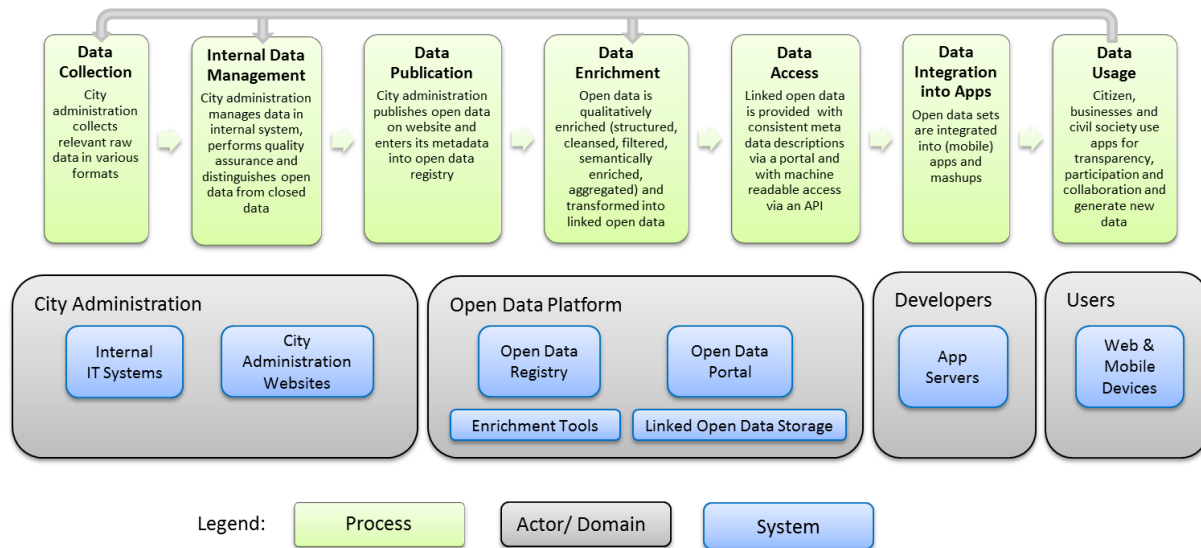


Figure 30 - Generic Scenario for Data Provision  
Source from Lapi, Tcholtchev et al. (2012)

In the open data lifecycle, there are many stakeholders involved. The decisions concerning the stages and identified in the Open Data lifecycle are summarized and described in Table 7. The steps and stages identified, have brought up the opportunities the OD offers to a variety of actors. Jetzek et al. (2014) pointed out that it is now possible to create any value from OD to civil society and governments. The common outcomes generated by OD are defined on economic or social values. Depending on the stakeholder and focus considered, the value generated can be transparency, participation, efficiency or innovation. The efficiency and innovation can be generated also by the ambidexterity influence onto opening-up data (Matheus & Janssen 2016), when is necessary to keep the provision of public services. However, constraints, such as a financial crisis, enforce the identification of innovative solutions in order to provide a better quality whilst using a smaller quantity.

For cases like the civil society, transparency and participation can be generated by non-governmental organizations. Magalhaes et al. (2013) and Janssen and Zuiderwijk (2014) observed that the creation of a group of “infomediaries”, that intermediate government data and people. Frequently, the initiatives created by infomediaries are based on websites and applications that bear value to people for enabling the better provision of services or for a specific reduction of information asymmetry. Figure 31 describes the enabling factors for reaching impact (values) on social and economic dimensions, in accordance with Jetzek et al. (2014). Apparently, the enabling factors can be divided into two dimensions: political and organizational, and, technical and infrastructural. These dimensions enable absorptive capacity, openness, governance and technical connectivity to reach the required impacts.



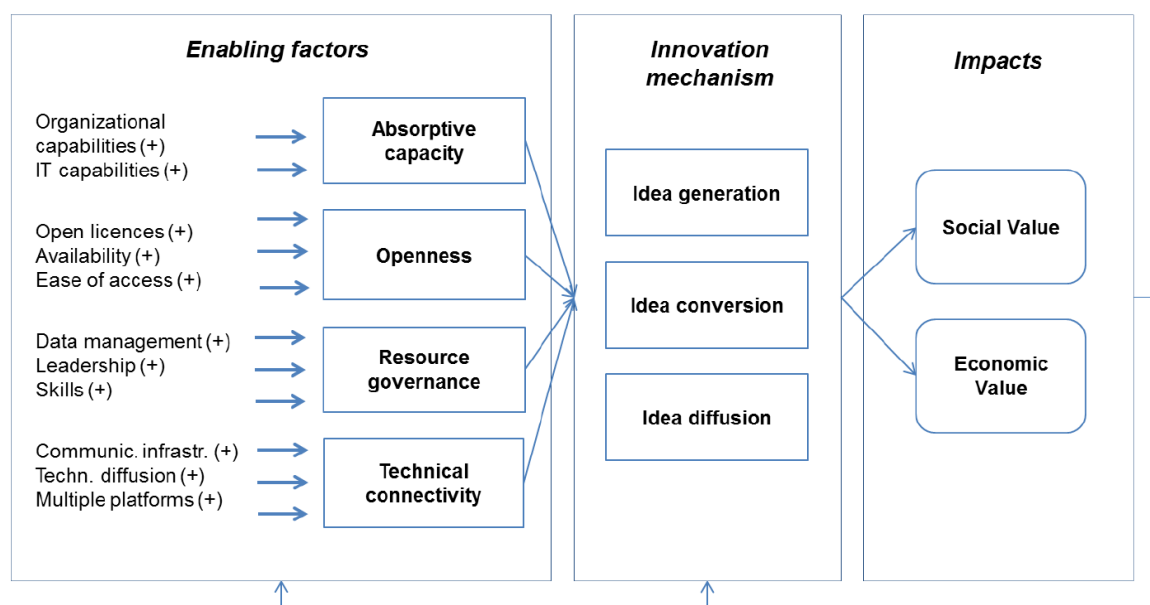


Figure 31 - Open Data Value generated based on focus and stakeholder  
Source from Jetzek et al. (2014)

Table 7 - Stages of Open Government Data Lifecycle  
Adapted from Attard et al. (2015)

Stage	Step	Description
<b>Pre-Processing</b>	<b>Data Creation</b>	Start of the cycle with the creation of data on the source.
	<b>Data Selection</b>	This step is selecting the data to be published taking in consideration the legal and technical needs.
	<b>Data Harmonisation</b>	Cleansing data to be in accordance to international standards such as Eight Principles of OGD.
	<b>Data Publishing</b>	Step that OGD is opened on the Open Data Portal.
<b>Exploiting</b>	<b>Data Interlinking</b>	Following the final stage of Five Star Scheme for LOD all the data sets are linked bringing value.
	<b>Data Discovery</b>	Publishing data is part of enabling usage. Another step considered is the discovery of data sets to users consume them and creating proper connections for analysis.
	<b>Data Exploration</b>	Step that highlight the most trivial way of consuming data, browsing and examining characteristics of data sets by simple visualisation for example or summary tables.
	<b>Data Exploitation</b>	Advanced way of consuming data, creating complex analysis and combining other data sets bringing innovation upon data usage.
<b>Maintenance</b>	<b>Data Curation</b>	Only step on the maintenance stage and is considered vital to give sustainability to the data published. Update and metadata enrichment are examples of curation.

### 8.3 Infrastructure functionalities in the Linked Open Data Lifecycle

This LOD Lifecycle presented on this section is not part of the State-of-Art, however, it was decided to be written due the possibility of earning time and receiving inputs and review from partners to the WP2 – Framework Creation and the WP 4- Pilots Planning Evaluation.

#### 8.3.1 Comparing open data life cycles

In general, there are many LOD lifecycles provided by prior research, for instance from Auer et al. (2012), Hyland & Wood (2011), or Villazón-Terrazas et al. (2011). These research becoming even longer if research in Open Government Data lifecycle (already explained in section 8.2.5), for instance Attard et al. (2015), or Charalabidis et al. (2016) are included. Most of the approaches have similarities on a core set of stages and mainly focus in two dimensions, the supply of data and demand point of view. Some of prior life-cycles were presented are presented at Table 8. Several steps from those LOD life-cycles were grouped based on its similarities into four stages of life-cycle: Data Creation, Data Publishing, Data Usage and Data Curation.

Table 8 –Comparing Linked Open Data Lifecycle approaches

Hyland and Wood (2011)	Villazón-Terrazas et al. (2011)	Attard et al. (2015)	Charalabidis et al. (2016)	Heusenblas (2011) <sup>33</sup>	Proposed Lifecycle
Identify	Specify	Create	Create	Data awareness	Data Creation
Model	Model	Select	Pre-Process	Modelling	
Name			Store/Obtain		
Describe	Generate	Harmonise	Publish	Publishing	Data Publishing
Convert	Publish	Publish			
Publish					
	Exploit	Interlink	Retrieve/Acquire	Discover	Data Usage
		Discover	Process	Integrate	
		Explore	Use	Use	
		Exploit	Collaborate		
Maintain		Curate	Curate		Data Curation

#### 8.3.2 An integrated open data life cycle

Figure 27 illustrates the life-cycle of LOD. As mentioned previously, there are four stages in the LOD life-cycle:

<sup>33</sup> Acquired from [https://www.w3.org/2011/gld/wiki/GLD\\_Life\\_cycle](https://www.w3.org/2011/gld/wiki/GLD_Life_cycle), accessed on 06/07/2016.

1. Data Creation, including data collection, data storage, and data selection.
2. Data Publishing, including pre-processes such as harmonising, and cleansing data, metadata creation and opening-up data.
3. Data Usage, including linking, discovering, retrieving, exploring and exploiting data.
4. Data Curation, including maintaining data and data governance.

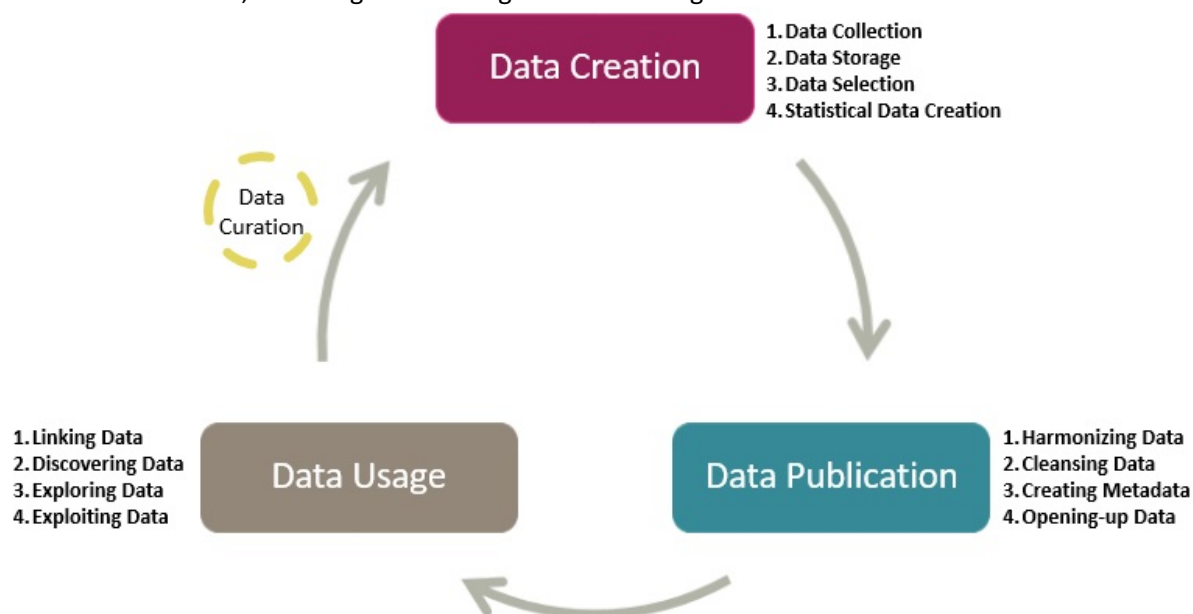


Figure 9 - Stages of LOD Lifecycle

Table 10 – LOD lifecycles Stages, Steps and Descriptions provides a summarised description of steps in the proposed lifecycle.

Table 10 – LOD lifecycles Stages, Steps and Descriptions

Stage	Step	Description
Data Creation	<b>Start of the cycle with the creation of data on the source.</b>	
	<b>Data Set Collection</b>	In this step, data is collected from various sources. Those data can be either structured data or unstructured data.
	<b>Data Storage</b>	Usually, the collection of data from previous step is stored in databases. However, there is also possible to directly processing data without storing data.
	<b>Data Set Selection</b>	This step is selecting the data to be published taking in consideration the legal and technical needs.
Data Publication	<b>Data Harmonisation</b>	In this step, data is translated in accordance to international standards such as Eight Principles of OGD.
	<b>Cleansing Data</b>	Before publishing data, data provider should check any errors on data, and inform the limitation of the data in metadata.
	<b>Metadata Creation</b>	Metadata creation
	<b>Data Publishing</b>	Data can be published in two ways: via API or Dump. Step that OGD is opened on the Open Data Portal via API.
Data Usage	<b>Linking Data</b>	Following the final stage of Five Star Scheme for LOD all the data sets are linked bringing value.
	<b>Data Discovery</b>	Publishing data is part of enabling usage. Another step considered is the discovery of data sets to users consume them and creating proper connections for analysis.
	<b>Data Exploration</b>	Step that highlight the most trivial way of consuming data, browsing and examining characteristics of data sets by simple visualisation for example or summary tables.
	<b>Data Exploitation</b>	Advanced way of consuming data, creating complex analysis and combining other data sets bringing innovation upon data usage.
Data Curation	<b>Only step on the maintenance stage and is considered vital to give sustainability to the data published. Update and metadata enrichment are examples of curation.</b>	

### 8.3.3 LOD Creation

The first stage of the LOD lifecycle is data creation. In any organization, especially in government agencies, creating data is part of their daily business process (Attard, Orlandi et al. 2015). In this stage, raw data from many sources need to be collected, and stored, before it can be published via the Internet.

#### 8.3.3.1 Data Collection

Data collection defined as a systematic way to gather information that can be used to explain phenomena, evaluate a process, decision making or problem solving (McLean, Mark et al. 1998, Anokwa, Hartung et al. 2009). Data provided by various sources, such as sensors, RFIDs, other information technologies and human, are gathered automatically by system or entered manually (Charalabidis, Alexopoulos et al. 2016) into several formats.

#### 8.3.3.2 Data Storage

In this process, the data loaded to secure data warehouse or databases. The way data being stored is also crucial, for example, in what kind of data format data is deposited (that should be align for whole life-cycle process), or how to create agile processes to load and store data in a scalable and flexible architecture; this will benefit for the next stage in the life-cycle. With an increase in the volume of data, data storage might be challenging because the amount of storage needed to process and store will also increase (Krishnan 2013).

#### 8.3.4 LOD Publication

This stage including pre-processing data and opening up data. In the pre-processing data, data already stored in repository should be selected, harmonized, and refined following legal and technical requirements; these pre-processes steps can be also act as the data quality assurance.

##### 8.3.4.1 Data Selection

Before being published, data should be going through selection process. In the case of LOD, this process includes removing any private data or personal data, as well as identifying under which conditions will this data be published, in term of the specification of LOD policies (Attard, Orlandi et al. 2015).

##### 8.3.4.2 Data Harmonisation

The purpose of this process is to compare similar conceptual and logical data models to create the common data elements, filter similar and different data elements to produce a unified data model that can be used consistently in the next stages of data-cycle (Thanos 2013). In term of LOD, data need to conform with publishing standards, such as the Eight Open Government Data Principles (Attard, Orlandi et al. 2015).

##### 8.3.4.3 Data Cleansing

When related data is provided by several sources, it is highly likely suffered with inconsistency or error, for example misspellings, be missing in recent changes, or be in a wrong order. For these kind of errors, data need to be refined. An example of data cleansing is entity resolution, or deduplication, which identifies and merges information from multiple sources that actually refer to the same entity (Bernstein and Haas 2008).

#### 8.3.4.4 The role of Metadata in Linked Open Data Infrastructures

Public and private organisations increasingly opening their data to gain benefits such as transparency, public accountability or economic growth. The use of this activity can be supported and stimulated by providing considerable metadata, including discovery, contextual and detailed metadata (Zuiderwijk, Jeffery et al. 2012). Metadata are simply defined as "data about data" (Dempsey and Heery 1998, Sheth 1999, NISO 2004, Vardaki, Papageorgiou et al. 2009). "Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource" (NISO 2004). Metadata should also describe the relationship between resources, and characterise attributes of people, services, software components and data (Dempsey and Heery 1998). For example, metadata can be information about the sampling subjects of a conducted survey, the method used for this survey, and the population studied (Vardaki, Papageorgiou et al. 2009).

Ideally, there are three different types of metadata should be provided for LOSD (Zuiderwijk, Jeffery et al. 2012) as shown at Figure 32.

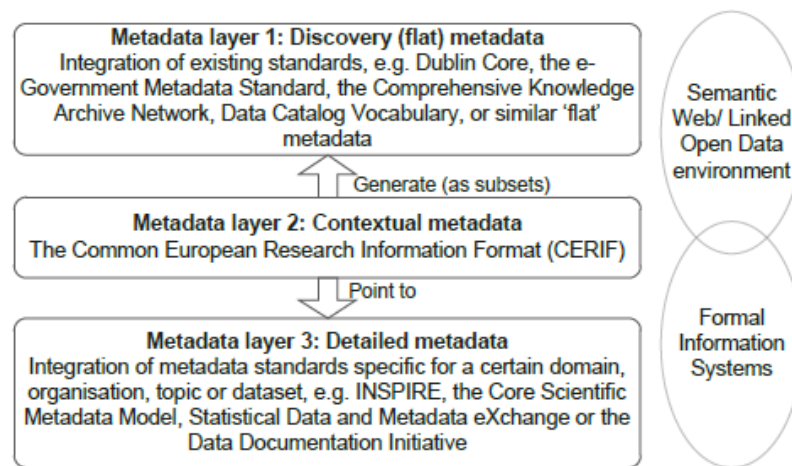


Figure 32 - Metadata system design  
Source from (Zuiderwijk, Jeffery et al. 2012)

- 1) Discovery (flat) metadata. These metadata are useful for discovery functionality of datasets by browsing and querying, for example, identifier, title, creator, publisher, country, source, format, language, keywords, validity date (from – to), audience, legal framework, status, relevant resources and linked data sets.
- 2) Contextual metadata. These metadata contain information about the context of datasets, in which datasets has been created and in which it can be reused. Examples of these metadata include data about organizations, persons, projects, funding, facilities, equipment, services and pointers to detailed metadata.
- 3) Detailed metadata. These metadata are additional information about discovery and contextual metadata, for example, quality and domain or dataset-specific parameters that are used by software processing the dataset.

(Zuiderwijk, Jeffery et al. 2012) present several benefits of metadata for LOD as shown in the table X. For this table, they categorized the benefits of metadata into 5 aspects: accessibility of data, discovery of data, interpretation of data, linking data and other purposes.

Table 11 - Benefits of Metadata creation for LOD  
Adapted from (Zuiderwijk, Jeffery et al. 2012)

Functionality	Benefits	Sources
<b>Accessibility of data</b>	Metadata improve storing and preservation of LOD	(NISO 2004, King, Liakata et al. 2011, Tenopir, Allard et al. 2011)
	Metadata improve the accessibility of LOD by describing, locating and retrieving the data efficiently	(Bertot, Jaeger et al. 2009, Tenopir, Allard et al. 2011, Pallickara, Pallickara et al. 2012)
<b>Discovery of data</b>	Metadata improve the ability to find LOD and the chance to be found.	(Schuurman, Deshpande et al. 2008, Borgman 2010)
	Metadata make potential users aware of the existence of certain datasets.	(Schuurman, Deshpande et al. 2008)
<b>Interpretation of data</b>	Metadata make sense of LOD by creating order within datasets	(NISO 2004, Berners-Lee 2009)
	Metadata improve easily analysing, finding patterns, comparing, reproducing and finding inconsistencies in LOD.	(King, Liakata et al. 2011)
	Metadata improve chances of a correct interpretation of LOD	(Jeffery 2000; Foulonneau and Cole 2005)
	Metadata may make it possible to assess and rank the quality and reliability of LOD.	(Honle et al. 2005; Dawes 2010)
	Metadata allow bringing similar resources together and distinguishing dissimilar resources.	(NISO 2004)
	Metadata provide a link between the creator of the data and the person who reuses these data.	(Burt and Taylor 2003)
	Metadata may improve visualizing LOD. Metadata can, for instance, improve accuracy of mapping.	(Park, Kim et al. 2011)
	Metadata enable detecting changes of LOD and therefore they can help in version management.	(Sen 2004, Liu and Li 2011)
	Metadata allow a dataset to be understood by both humans and machines in ways that promote interoperability.	(NISO 2004)
	Metadata allow giving location information.	(NISO 2004)
<b>Linking Data</b>	Metadata make linking data easier.	(Rahm and Do 2000)
	Metadata facilitate legacy resource integration.	(NISO 2004)
	Metadata are essential for integrating and linking data from heterogeneous sources.	(Jeffery 2000)
<b>Other purposes</b>	Metadata avoid unnecessary duplication of LOD.	(King, Liakata et al. 2011)

	Metadata can increase the visibility of researchers and hereby stimulate collaboration among researchers from various organizations.	(Nonthakarn and Wuwongse 2012)
	Metadata can help to create and maintain a common understanding of business objects and business processes.	(Vardaki, Papageorgiou et al. 2009, Hüner, Otto et al. 2011)

However, despite the many potential benefits of metadata, providing metadata for open data is often complicated (Martin 2014). While the literature posits that it is essential for the correct interpretation and use of open data to offer sufficient metadata at the same time (Jeffery 2000; Braunschweig et al. 2012), open government initiatives in general have been criticised for providing insufficient metadata (Jurisch et al. 2015). The provision of inadequate metadata was also found for OGD initiatives in particular (Dawes & Helbig 2010; Dawes 2010). These situations may lead to ambiguous semantics of the data (Conradie & Choenni 2012) and unawareness of datasets by users (Schoorman et al. 2008). Table Y below shows the disadvantage of metadata found from literature provided by (Zuiderwijk, Jeffery et al. 2012).

Table 12 - Disadvantages of Metadata for LOD  
Adapted from (Zuiderwijk, Jeffery et al. 2012)

	Disadvantages	Sources
<b>Costs of metadata</b>	Metadata may be sensitive and may be spread with the data unwillingly.	(Castiglione et al. 2007)
	Adding metadata is very time-consuming, as metadata operations consume over 60 per cent of the operations in typical workloads.	(Xiong et al. 2011)
	Requires high investments and is costly.	(Duval et al. 2002; Vardaki et al. 2009)
<b>Interpretation of data</b>	The provision of considerable metadata makes it difficult to create consistency between metadata.	(Duval et al. 2002)
	When metadata contain assumptions for the use of open data, they could point at certain choices and interpretations. This may unconsciously exclude certain ways of reusing data.	(Zuiderwijk et al. 2012)

According to the Guidelines for Statistical metadata on the Internet<sup>34</sup>, metadata should assist the user in particularly 3 functionalities: searching for statistical data, interpreting its content and post-processing statistical applications. These functionalities are essential in addressing some particular features of statistical data, specifically in term of the broad variety of users and the quality of statistical data.

For the searching functionality, basically there is no specific distinction between metadata for statistical data and other data. However, based on the Guidelines for Statistical metadata on the

<sup>34</sup> "Guidelines for Statistical Metadata on the Internet - unece." 2002. 7 Jun. 2016  
<<http://www.unece.org/stats/publications/metadata.pdf>>



Internet<sup>35</sup> there are several additional information needed in statistical metadata to search and discovery of data:

- Descriptions of statistical subject areas;
- Description of the statistical institution (legal framework, organisational structure, etc.);
- Description of the statistical system (role of different partners, etc.);
- Reference publications, product overview, general publications;
- Contact persons or e-mail addresses for more information;
- Release (and update) date(s);
- Links to other statistical sites;

Users of statistical data mainly require data related to a specific problem or general interest. They may also have different ability and understanding regarding statistics. Based on this, there are 12 minimum set of metadata required for the correct interpretation of statistics (ibid):

- Title/content description; often including statistical population, geographical coverage, observation unit and classification and standards applied.
- Labels for rows/columns in tables and elements of graph.
- Definitions of labels.
- Measurement unit.
- Time reference/period of data retrieval.
- Regional units.
- Comparability over time (break in series, missing data).
- Footnotes highlighting specific precautions.
- Source of the data.
- Explanation of standard symbols in tables.
- Any information on copyright or restriction of usage.
- Contact points for additional information.

In addition, there are 5 recommended metadata for better interpretation:

- 1) Comparability with alternative sources.
- 2) Links to summary of findings.
- 3) Description of methods used in collection, revision, calculation and estimation of the statistics.
- 4) Information on error sources and accuracy of the statistics.
- 5) Description of background and purpose of the statistics, concepts, variables and standards used.

---

<sup>35</sup> "Guidelines for Statistical Metadata on the Internet - unece." 2002. 7 Jun. 2016  
<<http://www.unece.org/stats/publications/metadata.pdf>>

---

For the post-processing purpose, the minimum set of metadata should either be attached to downloadable data or easy to access and download in separate file. The metadata should also allow further processing using suitable tools. When putting statistical data on the Internet, providers should be aware of the possibilities and limitations of the different formats with regard to downloading of statistical data (ibid).

#### 8.3.4.5 Data Publishing

This stage is the actual act of opening up the data. There are two ways to opening-up data: First, dump the data to organization websites, 3<sup>rd</sup> party sites or via FTP servers; however, by using this approach, it might be very difficult for an outsider to discover where to find updated information. Second, via API, which allows users to select specific portions of the data, rather than providing all of the data in bulk as a large file. APIs are typically connected to a database which is being updated in real-time. This means that making information available via an API can ensure that it is up to date<sup>36</sup>.

#### 8.3.5 LOD Usage

According to (Davies 2010), there are 5 types of data use in LOD:

- 1) data to fact, includes extracting particular facts from datasets;
- 2) data to information, includes generating a representation of a dataset, interpreting it, and reporting on the interpretation;
- 3) data to data, includes extending an original dataset by combining it with other data, changing its format or manipulating it otherwise, and subsequently sharing the extended dataset;
- 4) data to interface, includes providing an interface to interactively access and explore data;
- 5) data to service, includes integrating datasets to produce new products or services.

##### 8.3.5.1 Linking Data

Interlinking defined as “the degree to which entities that represent the same concept are linked to each other” (Zaveri et al. 2013). Data Interlinking is the highest level in the Five Star Scheme for LOD. This process allows published data to create additional values by giving context to its interpretation. Linking data is a derivative from the semantic web. The basic idea of semantic web is to convert or map the data on at least three different levels: used syntax, schemas and vocabularies used to deliver meaningful information (Bauer and Kaltenböck 2011). Detail information about this process already presented in chapter 8.3.

##### 8.3.5.2 Data Discovery

In order to gain benefits from LOD, users must discover the existence of open data. According to (Zuiderwijk 2015), there are four influencing factors of data discovery in LOD:

1. Data fragmentation Data; deal with difficulties to locate the datasets that they want to use, since the data are offered at many different places.
2. Terminology heterogeneity; deal with heterogeneous terminologies or keywords are used to describe datasets. Because of this problem, users often do not know which terms they should use to search for the data that they need.

---

<sup>36</sup> <http://opendatahandbook.org/guide/en/how-to-open-up-data/>

3. Search support; deal with issues regarding search functionalities, lack of more advanced multilingual, or data query functionalities.
4. Information overload; deal with amounts of OGD can at a certain point become overwhelming which complicates finding the OGD that a user needs.

Further, (Krishnan 2013) defined several steps in the process of data discovery for analytics:

- Data acquisition, includes gathering the files and preparing the data for import using a tool.
- Data tagging, includes creating an identifying link on the data for metadata integration.
- Data classification, includes creating subsets of value pairs for data processing and integration.
- Data modelling, includes creating a model for data visualization or analytics. The output from this step can be combined into an extraction exercise.

#### 8.3.5.3 Data Exploration

This step is the basic way in using data. There are (Zuiderwijk 2015) summarised several activities included in this step that are data manipulation and contextualisation, online view, statistical analysis, data download, and data visualisation.

Those activities can be done by tools provided by LOD infrastructure, or users may need to download all needed data and explore it using their own tools. For example, data can be visualised using mashup tools and software such as Datameer, Karmasphere, Tableau, and Spotfire. These mashup tools provide the capability for the user to integrate multiple data into one picture, by linking common data between the different data sets. The use of statistical software like R, SAS, and KXEN, may also useful for visualising the data (Krishnan 2013).

#### 8.3.5.4 Data Exploitation

This is a more advanced way of using data, including several activities such as data integration, data cleansing, data enrichment, and data transformation. Exploitation enables users to use, reuse or distribute the open data by leading out analysis, creating mashups, or innovating upon the open data in more active manner (Attard et al. 2015).

#### 8.3.6 Data Curation

Data curation defined as “the act of discovering a data source(s) of interest, cleaning and transforming the new data, semantically integrating it with other local data sources, and de-duplicating the resulting composite” (Stonebraker et al. 2013). From this definition, data curation presents in all stages in LOD life-cycle as an act to manage and govern data in proper and systematic way.

## 9 Conclusion

This document is the first deliverable, entitled D1.1 “OpenGovIntelligence Challenges & Needs”, of the first work package of the OpenGovIntelligence project. The objective of the OpenGovIntelligence project is to provide a holistic approach for the modernisation of Public Administration (PA) by exploiting Linked Open Statistical Data (LOSD) technologies thus stimulate sustainable economic growth in Europe through fostering innovation in societies and enterprises.

Work package one (WP1) is responsible for eliciting the challenges and needs regarding political, legal, institutional, social and technical issues in opening-up and exploiting LOSD for the co-production of innovative data-driven public services. WP1 comprises four tasks with the following objectives: Task 1.1 aims to review existing government and other data infrastructures, Task 1.2 aims to elicit needs and expectations for service co-production, Task 1.3 aims to identify LOSD challenges and needs, and Task 1.4 aims to understand how PA can be transformed to innovation based on better use of data.

The methodology that was followed in this deliverable comprises the following actions:

- Identifying *PA’s challenges and needs* related to open data-driven innovation and public service co-creation. Towards this end, a thorough literature review was performed and an online survey aiming at both public and private sector was organised.
- Identifying *technical challenges and needs* of LOSD. Towards this end, the LOSD scientific literature was studied and 11 LOSD top experts from around the globe were involved.
- Identifying *users’ challenges and needs* regarding the exploitation of Open Statistical Data that are provided by Open Data Portals. Towards this end, we first play the role of a user and we studied the UK Open Data Portal in order to discover statistical data about a specific phenomenon. Moreover, we elicited the opinion of developers who have used Open Statistical Data to create data-driven applications.
- Identifying the *needs of the OpenGovIntelligence pilot partners* regarding data-driven innovation. Towards this end, a number usage scenarios were specified in order the pilots to describe the challenges they want to address, the available datasets, and the final data-driven public service they want to co-produce.
- Identifying the *State-of-the-art of data infrastructures*.

These actions resulted in a set of outcomes that will feed future work in the project. More specifically, the results of this deliverable include:

- *PA’s challenges and needs* regarding data-driven innovation (Section 4). This includes challenges such as lack of political and administrative priority and leadership, incompatibility of existing administrative and organisational cultures with the idea of co-creation and resistance of public administrators to fundamental changes.
- *Technical challenges and needs of LOSD* that are presented in the following way (Section 5):
  - A mapping and analysis of the LOSD state-of-the-art. This includes a documentation of existing software tools, vocabularies, architectures, and use cases that can be re-used and re-purposed in the frame of the project.
  - A set of LOSD publishing practices that cause LOSD interoperability challenges. These technical challenges are accompanied by a list of advantages and disadvantages per

practice based on top experts' opinions. These results can feed project's standardisation activities in WP5, so that the community to agree on a list of best practices for LOSD publishing.

- *Users' challenges regarding the exploitation of statistical data* as they are provided through major Open Data Portals (Section 6). These challenges are presented in the form of:
  - An empirical description of the "Open Statistical Data Fragmentation" challenge based on the analysis of data.gov.uk. In summary, our research revealed that searching this portal for useful data on unemployment produces 122 results that provide access to 56 files and 610 links to 18 other portals (such as the Office for National Statistics and the National Archives) and by following the relevant links to more than 2,000 other files.
  - Analysis of the responses of 24 developers that have exploited Open Statistical Data in order to produce data-driven applications. Some of the challenges that they have faced include interoperability among datasets, the lack of metadata, and the quality of metadata.
- *Usage scenarios that reflect pilot's needs* regarding data-driven innovation in PA (Section 7). These scenarios will feed into WP4 in order the evaluation scenarios to be developed.
- *The state-of-the-art regarding open data infrastructures* (Section 8). This comprises a maturity model for LOSD and a lifecycle for data infrastructures.

## References

1. C. Abdallah, S. and I.-S. Fan (2012). "Framework for e-government assessment in developing countries: case study from Sudan." *Electronic Government, an International Journal* 9(2): 158 - 177.
2. Ackerman, J. M. and I. E. Sandoval-Ballesteros (2006). "The global explosion of freedom of information laws." *Administrative Law Review*: 85-130.
3. Ackoff, R. L. (1989). "From data to wisdom." *Journal of applied systems analysis* 16(1): 3-9.
4. Akerman, A. and J. Tyree (2006). "Using ontology to support development of software architectures." *IBM Systems Journal* 45(4): 813-825.
5. Albury, D. (2005). *Fostering Innovation in Public Services*. Public Money & Management, Vol. 25, pp. 51-56.
6. Álvarez-Rodríguez J. M., Labra-Gayo J. E., & de Pablos P. O. (2013, November). Leveraging Semantics to Represent and Compute Quantitative Indexes: The RDFIndex Approach. In *Research Conference on Metadata and Semantic Research* (pp. 175-187). Springer International Publishing
7. Anokwa, Y., C. Hartung, W. Brunette, G. Borriello and A. Lerer (2009). "Open source data collection in the developing world." *Computer* 42(10): 97-99.
8. Aracri, R., De Francisci, S., Pagano, A., Scannapieco, M., Tosco, L., & Valentino, L. (2011). Publishing the 15 th Italian Population and Housing Census in Linked Open Data.
9. Aracri R.M., Francisci, S., Pagano, A., Scannapieco, M. (2015). Official Statistics meets the Semantic Web: How SDMX and RDF can live together. In *New Techniques and Technologies for Statistics (NTTS) 2015*.
10. Attard, J., F. Orlandi, S. Scerri and S. Auer (2015). "A systematic review of open government data initiatives." *Government Information Quarterly* 32(4): 399-418.
11. Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z. Ives (2007). *Dbpedia: A nucleus for a web of open data*, Springer.
12. Auer, S., Demter, J., Martin, M., & Lehmann, J. (2012, October). LODStats—an extensible framework for high-performance dataset analytics. In *International Conference on Knowledge Engineering and Knowledge Management* (pp. 353-362). Springer Berlin Heidelberg
13. Auer, S., J. Lehmann and S. Hellmann (2009). *Linkedgeodata: Adding a spatial dimension to the web of data*. International Semantic Web Conference, Springer.
14. Auer, S., L. Bühmann, C. Dirschl, O. Erling, M. Hausenblas, R. Isele, J. Lehmann, M. Martin, P. N. Mendes and B. Van Nuffelen (2012). *Managing the life-cycle of linked data with the LOD2 stack*. International semantic Web conference, Springer.
15. Bandholtz, T., Schulte-Coerne, T., & Rüther, M. (2009). *Linked environment data: Scovo-fying the environment specimen bank*. ESWC.

16. Banisar, D. (2006). "Freedom of information around the world 2006: A global survey of access to government information laws." Privacy International.
17. Barry, E., Bannister, F. (2013). Barriers to open data release: A view from the top. European Group for Public Administration, Edinburgh.
18. Batini A., Lenzerini M., Navathe S. B., A comparative analysis of methodologies for database schema integration, ACM computing surveys (CSUR) 18 (4) (1986) pp.323-364.
19. Bauer, F. and M. Kaltenböck (2011). "Linked open data: The essentials." Edition mono/monochrom, Vienna.
20. Bayerl, S., & Granitzer, M. (2015, July). bacon: Linked Data Integration based on the RDF Data Cube Vocabulary. In Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics (p. 14). ACM.
21. Bayerl, S., & Granitzer, M. (2015, October). Data-transformation on historical data using the RDF data cube vocabulary. In Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business (p. 15). ACM.
22. Becker, K., Jahangiri, S., & Knoblock, C. A. (2015a). A Quantitative Survey on the Use of the Cube Vocabulary in the Linked Open Data Cloud. SemStats 2015.
23. Becker, K., Tan, X., Jahangiri, S., & Knoblock, C. A. (2015b). Finding, Assessing, and Integrating Statistical Sources for Data Mining. In KNOW@ LOD.
24. Bekkers, V.J.J.M., Tummers, L.G., Voorberg, W.H. (2013). From public innovation to social innovation in the public sector: A literature review of relevant drivers and barriers. Rotterdam: Erasmus University Rotterdam.
25. Berger S., Schrefl M., Feddw global schema architect: Uml-based design tool for the integration of data mart schemas., in: I.-Y. Song, M. Golfarelli (Eds.), DOLAP, ACM, Maui, Hawaii, USA, 2012, pp. 33-40.
26. Berners-Lee, T. (2009). "5-Star Deployment Scheme for Linked Data." from <http://5stardata.info/en/>.
27. Bernier L., Hafsi T. (2007). The Changing Nature of Public Entrepreneurship. Public Administration Review, Vol. 67, pp. 488-503.
28. Bernstein, P. A. and L. M. Haas (2008). "Information integration in the enterprise." Communications of the ACM 51(9): 72-79.
29. Bertot, J. C., P. T. Jaeger, J. A. Shuler, S. N. Simmons and J. M. Grimes (2009). "Reconciling government documents and e-government: Government information in policy, librarianship, and education." Government Information Quarterly 26(3): 433-436.
30. Bharosa, N., M. Janssen, R. van Wijk, N. de Winne, H. van der Voort, J. Hulstijn and Y.-h. Tan (2013). "Tapping into existing information flows: The transformation to compliance by design in business-to-government information exchange." Government Information Quarterly 30: S9-S18.

31. Bizer, C., T. Heath, D. Ayers and Y. Raimond (2007). Interlinking open data on the web. Demonstrations track, 4th european semantic web conference, innsbruck, austria.
32. Bizer, C., T. Heath and T. Berners-Lee (2009). "Linked data-the story so far." *Semantic Services, Interoperability and Web Applications: Emerging Concepts*: 205-227.
33. Blakemore, M. and M. Craglia (2006). "Access to public-sector information in Europe: Policy, rights, and obligations." *The Information Society* 22(1): 13-24.
34. Borgman, C. L. (2010). *Scholarship in the digital age: Information, infrastructure, and the Internet*, MIT press.
35. Bosch, T., Cyganiak, R., Gregory, A., & Wackerow, J. (2013). DDI-RDF Discovery Vocabulary: A Metadata Vocabulary for Documenting Research and Survey Data. In LDOW.
36. Brasoveanua, A. M., Saboub, M., Scharla, A., Hubmann-Haidvogela, A., & Fischla, D. (2016) Visualizing statistical linked knowledge for decision support. *Semantic Web*, pp. 1-25.
37. Bratt, S. (2007). "Semantic web and other W3C technologies to watch." Talks at W3C, January.
38. Braunschweig, K., J. Eberius, M. Thiele and W. Lehner (2012). "The State of Open Data Limits of Current Open Data Platforms."
39. Bruckner R. M., Ling T. W., Mangisengi O., et al., A framework for a multidimensional olap model using topic maps, in: *Web Information Systems Engineering, 2001. Proceedings of the Second International Conference on*, Vol. 2, IEEE, 2001, pp. 109-118.
40. Burt, B. and J. Taylor (2003). "Constructing new ways of living together: Government relationships with the voluntary sector in the information polity." *Information Polity: The International Journal of Government & Democracy in the Information Age* 8(3/4): 181-192.
41. Capadisli, S., Auer, S., & Ngonga Ngomo, A. C. (2013). Linked SDMX Data: Path to high fidelity Statistical Linked Data for OECD, BFS, FAO, and ECB. *Semantic Web*.
42. Capadisli, S., Auer, S., & Riedl, R. (2013). Linked statistical data analysis. *Semantic Web Challenge*.
43. Capadisli, S., Meroño-Peñuela, A., Auer, S., & Riedl, R. (2014). Semantic similarity and correlation of linked statistical data analysis. In *Proceedings of the 2nd International Workshop on Semantic Statistics (SemStats 2014)*, ISWC. CEUR.
44. Castiglione, A., A. De Santis and C. Soriente (2007). "Taking advantages of a disadvantage: Digital forensics and steganography using document metadata." *Journal of Systems and Software* 80(5): 750-764.
45. Celino, I., & Calegari, G. R. (2014). Geo-statistical exploration of Milano datasets. In *Second International Workshop for Semantic Statistics SemStats*.
46. Ceolin, D., Nottamkandath, A., & Fokink, W. (2012, May). Automated evaluation of annotators for museum collections using subjective logic. In *IFIP International Conference on Trust Management* (pp. 232-239). Springer Berlin Heidelberg.



47. Chamberlayne, R., B. Green, M. L. Barer and C. Hertzman (1998). "Creating a population-based linked health database: a new resource for health services research." *Canadian Journal of Public Health* 89(4): 270.
48. Chan, J. K.-s. (2014). "G8 Open Data Charter Action Plan: Open data by default, but you may have to pay." The Sunlight Foundation.
49. Channah N., Aris O., A classification of semantic conflicts in heterogeneous database systems, *Journal of Organizational Computing* 5 (2) (1995) pp.167-193.
50. Charalabidis, Y., C. Alexopoulos and E. Loukis (2016). "A taxonomy of open government data research areas and topics." *Journal of Organizational Computing and Electronic Commerce* 26(1-2): 41-63.
51. Charalabidis, Y., E. Loukis and C. Alexopoulos (2014). Evaluating second generation open government data infrastructures using value models. 2014 47th Hawaii International Conference on System Sciences, IEEE.
52. Chun, S. A., S. Shulman, R. Sandoval and E. Hovy (2010). "Government 2.0: Making connections between citizens, data and government." *Information Polity* 15(1): 1.
53. Coglianese, C. (2009). "The transparency president? The Obama administration and open government." *Governance* 22(4): 529-544.
54. Conradie, P., Choenni, S. (2012). Exploring process barriers to release public sector information in local government. In: *Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance*, ACM, pp. 5-13.
55. Conradie, P. and S. Choenni (2012). Exploring Process Barriers to Release Public Sector Information in Local Government. 6th International Conference on Theory and Practice of Electronic Governance (ICEGOV 2012), Albany, New York.
56. Cox, P. and G. Alemanno (2003). "Directive 2003/98/EC of the european parliament and of the council of 17 november 2003 on the re-use of public sector information." *Official Journal of the European Union* 46: 1-156.
57. Craveiro, G. S., J. A. Machado and J. S. Machado (2016). The Use of Open Government Data to Citizen Empowerment. *Proceedings of the 9th International Conference on Theory and Practice of Electronic Governance*, ACM.
58. Cretu, V. and B. Manolea (2013). The Influence of the Open Government Partnership (OGP) on the Open Data discussions, EPSI, Topic Report.
59. Cyganiak, R., Field, S., Gregory, A., Halb, W., & Tennison, J. (2010). Semantic Statistics: Bringing Together SDMX and SCOVO. LDOW, 628.
60. Cyganiak, R., Hausenblas, M., & McCuirc, E. (2011). Official statistics and the practice of data fidelity. In *Linking Government Data* (pp. 135-151). Springer New York.
61. Cyganiak R. and D. Reynolds, The RDF Data Cube vocabulary, <http://www.w3.org/TR/vocab-data-cube/> (2014)

62. Daga, E., d'Aquin, M., Gangemi, A., & Motta, E. (2014). Early analysis and debugging of linked open data cubes. *International Semantic Web Conference (ISWC)*.
63. Davies, T. (2010). "Open data, democracy and public sector reform."
64. Dawes, S. (2010). "Stewardship and usefulness: Policy principles for information-based transparency." *Government Information Quarterly* 27(4): 377-383.
65. Dawes, S., L. Vidasova and O. Parkhimovich (2016). "Planning and designing open government data programs: An ecosystem approach." *Government Information Quarterly* 33(1): 15-27.
66. Dawes and N. Helbig (2010). *Information strategies for open government: Challenges and prospects for deriving public value from government transparency*. Electronic Government, Springer: 50-60.
67. Debattista, J., Lange, C., & Auer, S. (2014, September). Representing dataset quality metadata using multi-dimensional views. In *Proceedings of the 10th International Conference on Semantic Systems*(pp. 92-99). ACM.
68. de Leon, A., Wisniewski, F., Villazón-Terrazas, B. and Corcho, O (2012). *Map4rdf - Faceted Browser for Geospatial Datasets*. In *PMOD workshop*, 2012.
69. Dempsey, L. and R. Heery (1998). "Metadata: a current view of practice and issues." *Journal of documentation* 54(2): 145-172.
70. Deutsch, A., et al. (2004). Specification and verification of data-driven web services. *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. Paris, France, ACM: 71-82.
71. De Vries, H., Bekkers, V. and Tummers, L. (2016). *Innovation in the Public Sector: A Systematic Review and Future Research Agenda*. *Public Administration*, Vol. 94 (1), pp. 146–166.
72. Diamantini, C., & Potena, D. (2008, October). Semantic enrichment of strategic datacubes. In *Proceedings of the ACM 11th international workshop on Data warehousing and OLAP* (pp. 81-88). ACM.
73. Diamantini A., Potena D., Storti E., *Data mart reconciliation in virtual innovation factories* 178 (2014) pp. 274-285. doi:10.1007/978-3-319-07869-4\_26.
74. Ding, L., T. Lebo, J. S. Erickson, D. DiFranzo, G. T. Williams, X. Li, J. Michaelis, A. Graves, J. G. Zheng and Z. Shangquan (2011). "TWC LOGD: A portal for linked open government data ecosystems." *Web Semantics: Science, Services and Agents on the World Wide Web* 9(3): 325-333.
75. Do, B. L., Aryan, P. R., Trinh, T. D., Wetz, P., Kiesling, E., & Tjoa, A. M. (2015a). Toward a framework for statistical data integration. In *Proceedings of the 3rd International Workshop on Semantic Statistics co-located with 14th International Semantic Web Conference (ISWC 2015)*.
76. Do, B. L., Trinh, T. D., Aryan, P. R., Wetz, P., Kiesling, E., & Tjoa, A. M. (2015b, September). Toward a statistical data integration environment: the role of semantic metadata. In *Proceedings of the 11th International Conference on Semantic Systems*(pp. 25-32). ACM.

77. Do, B. L., Trinh, T. D., Wetz, P., Kiesling, E., Anjomshoaa, A., & Tjoa, A. M. Multiscale Exploration of Spatial Statistical Datasets: A Linked Data Mashup Approach.
78. Doan A., Halevy A. Y., Semantic integration research in the database community: A brief survey, *AI magazine* 26 (1) (2005) pp.83-94.
79. Duval, E., W. Hodgins, S. Sutton and S. L. Weibel (2002). "Metadata principles and practicalities." *D-lib Magazine* 8(4): 16.
80. Dwivedi, Y.K., Ravichandran, K., Williams, M.D., Miller, S., Lal, B., Antony, G.V., Kartik M. (2013). IS/IT Project Failures: A Review of the Extant Literature for Deriving a Taxonomy of Failure Factors. In: Grand Successes and Failures in IT: Public and Private Sectors, Y.K. Dwivedi, H.Z. Henriksen, D. Wastell, R. De (Eds). *Proceedings of IFIP WG 8.6 International Working Conference on Transfer and Diffusion of IT, TDIT*. Bangalore, India, 27-29 June 2013. Springer, pp. 73-88.
81. Eaves, D. (2011). "The three laws of open government."
82. Ermilov, I., Auer, S., & Stadler, C. (2013, September). Csv2rdf: User-driven csv to rdf mass conversion framework. In *Proceedings of the ISEM* (Vol. 13, pp. 04-06).
83. Ermilov, I., Martin, M., Lehmann, J., & Auer, S. (2013, October). Linked open data statistics: Collection and exploitation. In *International Conference on Knowledge Engineering and the Semantic Web* (pp. 242-249). Springer Berlin Heidelberg.
84. Etcheverry, L., & Vaisman, A. A. (2012a). Enhancing OLAP analysis with web cubes. In *The Semantic Web: Research and Applications* (pp. 469-483). Springer Berlin Heidelberg.
85. Etcheverry, L., & Vaisman, A. A. (2012b). QB4OLAP: a new vocabulary for OLAP cubes on the semantic web. In *Proceedings of the Third International Conference on Consuming Linked Data - Volume 905 (COLD'12)*, Juan F. Sequeda, Andreas Harth, and Olaf Hartig (Eds.), Vol. 905. CEUR-WS.org, Aachen, Germany, Germany, 27-38.
86. Etcheverry, L., Gomez, S. S., & Vaisman, A. (2015). Modeling and Querying Data Cubes on the Semantic Web. *arXiv preprint arXiv:1512.06080*.
87. Etcheverry, L., Vaisman, A., & Zimányi, E. (2014, September). Modeling and querying data warehouses on the semantic web using QB4OLAP. In *International Conference on Data Warehousing and Knowledge Discovery* (pp. 45-56). Springer International Publishing.
88. European Commission. (2013). Powering European Public Sector Innovation: Towards A New Architecture. Report of the Expert Group on Public Sector Innovation. Available at: [https://ec.europa.eu/research/innovation-union/pdf/psi\\_eg.pdf](https://ec.europa.eu/research/innovation-union/pdf/psi_eg.pdf)
89. Fernández, A. V., & Zarrabeitia, A. S. (2013, August). Implementation of a linked open data solution for the statistics agency of Cantabria's metadata and data bank. In *Proceedings of the 2013 International Conference on Dublin Core and Metadata Applications* (pp. 1-8). Dublin Core Metadata Initiative.

90. Fidan, G., I. Dikmen, A. M. Tanyer and M. T. Birgonul (2011). "Ontology for relating risk and vulnerability to cost overrun in international projects." *Journal of Computing in Civil Engineering* 25(4): 302-315.
91. Fielding, R. T. and R. N. Taylor (2002). "Principled design of the modern Web architecture." *ACM Transactions on Internet Technology (TOIT)* 2(2): 115-150.
92. Follenfant, C., Trastour, D., & Corby, O. (2011, October). A model for assisting business users along analytical processes. In *SPIM-2nd Workshop on Semantic Personalized Information Management: Retrieval and Recommendation-2012* (Vol. 781, pp. 38-41). CEUR-WS.
93. Foulonneau, M. and T. W. Cole (2005). *Strategies for reprocessing aggregated metadata*. International Conference on Theory and Practice of Digital Libraries, Springer.
94. Frischmuth, P., Martin, M., Tramp, S., Riechert, T., & Auer, S. (2015). OntoWiki—an authoring, publication and visualization interface for the data web. *Semantic Web*, 6(3), 215-240.
95. Gamage, P. (2016). "New development: Leveraging 'big data' analytics in the public sector." *Public Money & Management* 36(5): 385-390.
96. Ganapati, S., Reddick, C.G. (2012). Open e-government in US state governments: Survey evidence from Chief Information Officers. *Government Information Quarterly*, Vol. 29(2), pp. 115-122.
97. García, R. and R. Gil (2009). Publishing xbrl as linked open data. *CEUR Workshop Proceedings*.
98. García, R. and R. Gil (2010). *Linking XBRL financial data*. Linking enterprise data, Springer: 103-125.
99. Garrett, J. J. (2010). *Elements of user experience, the: user-centered design for the web and beyond*, Pearson Education.
100. Gayo, J. E. L., & Rodríguez, J. M. A. (2013, September). Validating statistical index data represented in RDF using SPARQL queries. In *RDF Validation Workshop. Practical Assurances for Quality RDF Data*, Cambridge, Ma, Boston.
101. Gayo, J. E. L., Farhan, H., Fernández, J. C., & Rodríguez, J. M. Á. (2014, October). Representing verifiable statistical index computations as linked data. In *Second International Workshop for Semantic Statistics SemStats*.
102. Gerlitz, C. and A. Helmond (2013). "The like economy: Social buttons and the data-intensive web." *New Media & Society*: 1461444812472322.
103. Gottron, T., Hachenberg, C., Harth, A., & Zapolko, B. (2011). Towards a Semantic Data Library for the Social Sciences. In *SDA* (pp. 48-59).
104. Gray, J., L. Chambers and L. Bounegru (2012). *The data journalism handbook*, " O'Reilly Media, Inc."
105. Gupta, M. K., Lambhate, P., & Emmanuel, M. Processing Linked Multidimensional Data on the Semantic Web (2016), *International Conference on Computing, Communication, and Energy Systems (ICCCES-16)* In Association with IET, UK & Sponsored by TEQIP.

106. Halb, W., A. Stocker, H. Mayer, H. Mülner and I. Ademi (2010). Towards a commercial adoption of linked open data for online content providers. Proceedings of the 6th International Conference on Semantic Systems, ACM.
107. Hallo, M., Luján-Mora, S., & Maté, A. (2015, July). Publishing a scorecard for evaluating the use of open-access journals using linked data technologies. In Computer, Information and Telecommunication Systems (CITS), 2015 International Conference on(pp. 1-5). IEEE.
108. Hallo, M., Luján-Mora, S., & Maté, A. (2016). Evaluating open access journals using Semantic Web technologies and scorecards. *Journal of Information Science*, 0165551515624353.
109. Hallo, M., Luján-Mora, S., Trujillo, J. (2015) An Approach to Publish Statistics from Open-Access Journals Using Linked Data Technologies, INTED2015 Proceedings, pp. 5940-5948.
110. Halpin, H., P. J. Hayes, J. P. McCusker, D. L. McGuinness and H. S. Thompson (2010). When owl: sameas isn't the same: An analysis of identity in linked data. *The Semantic Web–ISWC 2010*, Springer: 305-320.
111. Harris, H., S. Murphy and M. Vaisman (2013). Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work, " O'Reilly Media, Inc."
112. Hartley J. (2005). Innovation in Governance and Public Services: Past and Present. *Public Money & Management*, Vol. 25, pp. 27-34.
113. Hartmann, T., Zapilko, B., Wackerow, J., & Eckert, K. (2015a). Constraints to Validate RDF Data Quality on Common Vocabularies in the Social, Behavioral, and Economic Sciences. *arXiv preprint arXiv:1504.04479*.
114. Hartmann, T., Zapilko, B., Wackerow, J., & Eckert, K. (2015b). Evaluating the Quality of RDF Data Sets on Common Vocabularies in the Social, Behavioral, and Economic Sciences. *arXiv preprint arXiv:1504.04478*.
115. Hasapis, P., E. Fotopoulou, A. Zafeiropoulos, S. Mouzakitis, S. Koussouris, M. Petychakis, B. Kapourani, N. Zanetti, F. Molinari and S. Virtuoso (2014). Business value creation from Linked Data analytics: The LinDA approach. *eChallenges e-2014 Conference Proceedings*, IEEE.
116. Hausenblas, M. (2009). "Exploiting linked data to build web applications." *IEEE Internet Computing* 13(4): 68.
117. Hausenblas, M., Halb, W., Raimond, Y., Feigenbaum, L., & Ayers, D. (2009). Scovo: Using statistics on the web of data. In *The Semantic Web: Research and Applications* (pp. 708-722). Springer Berlin Heidelberg.
118. Heath, T. (2008). "How will we interact with the web of data?" *Internet Computing*, IEEE 12(5): 88-91.
119. Heath, T. and C. Bizer (2011). "Linked data: Evolving the web into a global data space." *Synthesis lectures on the semantic web: theory and technology* 1(1): 1-136.

120. Heeks, R. (2003). Most eGovernment-for-Development Projects Fail: How Can Risks be Reduced? iGovernment working paper series, Institute for Development Policy and Management University of Manchester.
121. Helmich, J., Klímek, J., & Nečaský, M. (2014, May). Visualizing RDF data cubes using the linked data visualization model. In European Semantic Web Conference (pp. 368-373). Springer International Publishing.
122. Hepp, M. (2008). Goodrelations: An ontology for describing products and services offers on the web. Knowledge Engineering: Practice and Patterns, Springer: 329-346.
123. Hillier, M. (2003). "The role of cultural context in multilingual website usability." Electronic Commerce Research and Applications 2(1): 2-14.
124. Hoefler, P., Granitzer, M., Veas, E. E., & Seifert, C. (2014, April). Linked Data Query Wizard: A Novel Interface for Accessing SPARQL Endpoints. In LDOW.
125. Höffner, K., & Lehmann, J. (2014, September). Towards question answering on statistical linked data. In Proceedings of the 10th International Conference on Semantic Systems (pp. 61-64). ACM.
126. Höffner, K., & Lehmann, J. (2014, September). Towards question answering on statistical linked data. In Proceedings of the 10th International Conference on Semantic Systems (pp. 61-64). ACM.
127. Höffner, K., Martin, M., & Lehmann, J. (2015). LinkedSpending: OpenSpending becomes Linked Open Data. Semantic Web, 7(1), 95-104.
128. Honle, N., U.-P. Kappeler, D. Nicklas, T. Schwarz and M. Grossmann (2005). Benefits of integrating meta data into a context model. Pervasive Computing and Communications Workshops, 2005. PerCom 2005 Workshops. Third IEEE International Conference on, IEEE.
129. Huijboom, N., Broek, T. (2011). Open data: an international comparison of strategies. European journal of ePractice, Vol. 12(1), pp. 4-16.
130. Hüner, K. M., B. Otto and H. Österle (2011). "Collaborative management of business metadata." International Journal of Information Management 31(4): 366-373.
131. Hunnius, S. and B. Krieger (2014). The Social Shaping of Open Data through Administrative Processes. Proceedings of The International Symposium on Open Collaboration, ACM.
132. Hyland, B. and D. Wood (2011). The joy of data-a cookbook for publishing linked government data on the web. Linking government data, Springer: 3-26.
133. Hsu C.-C., Sandford B. A., The delphi technique: making sense of consensus, Practical assessment, research & evaluation 12 (10) (2007) pp.1-8.
134. Ibragimov, D., Hose, K., Pedersen, T. B., & Zimányi, E. (2015). Towards exploratory OLAP over linked open data—a case study. In Enabling Real-Time Business Intelligence (pp. 114-132). Springer Berlin Heidelberg.

135. Jakobsen, K. A., Andersen, A. B., Hose, K., & Pedersen, T. B. (2015). Optimizing RDF data cubes for efficient processing of analytical queries. In Ceur Workshop Proceedings. CEUR-WS. org.
136. Janev, V., Mijovic, V., MiloSevic, U., & Vranes, S. (2014). Supporting the linked data publication process with the LOD2 statistical workbench. *Semantic Web Journal* (Submitted) <http://www.semantic-web-journal.net/content/supporting-linked-datapublication-process-lod2-statistical-workbench>.
137. Janev, V., Mijović, V., & Vraneš, S. (2013, September). Lod2 tool for validating rdf data cube models. In *Web Proceedings of the 5th ICT Innovations Conference* (pp. 12-15).
138. Janev, V., Mijović, V., Paunović, D., & Milošević, U. (2014, September). Modeling, fusion and exploration of regional statistics and indicators with linked data tools. In *International Conference on Electronic Government and the Information Systems Perspective* (pp. 208-221). Springer International Publishing.
139. Janev, V., Milošević, U., Spasić, M., Vraneš, S., Milojković, J., & Jireček, B. (2012, September). Integrating serbian public data into the LOD cloud. In *Proceedings of the Fifth Balkan Conference in Informatics* (pp. 94-99). ACM.
140. Janssen, K. (2011). "The influence of the PSI directive on open government data: An overview of recent developments." *Government Information Quarterly* 28(4): 446-456.
141. Janssen, K. (2011). The influence of the PSI directive on open government data: An overview of recent developments. *Government Information Quarterly*, Vol. 28(4), pp. 446-456.
142. Janssen, K. (2012). "Open government data and the right to information: Opportunities and obstacles." *The Journal of Community Informatics* 8(2).
143. Janssen, M., Charalabidis, Y., Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, Vol. 29(4), pp. 258-268.
144. Janssen, M., E. Estevez and T. Janowski (2014). "Interoperability in Big, Open, and Linked Data-Organizational Maturity, Capabilities, and Data Portfolios." *Computer*(10): 44-49.
145. Janssen, M., R. Wagenaar and J. Beerens (2003). Towards a flexible ICT-architecture for multi-channel e-government service provisioning. *System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on*, IEEE.
146. Janssen, M., Y. Charalabidis and A. Zuiderwijk (2012). "Benefits, adoption barriers and myths of open data and open government." *Information Systems Management* 29(4): 258-268.
147. Janssen, M. and A. Zuiderwijk (2014). "Infomediary business models for connecting open data providers and users." *Social Science Computer Review* 32(5): 694-711.
148. Janssen, M. and J. van den Hoven (2015). "Big and Open Linked Data (BOLD) in government: A challenge to transparency and privacy?" *Government Information Quarterly* 32(4): 363-368.
149. Jeffery, K. G. (2000). "Metadata: The future of information systems." *World Meteorological Organization*.

150. Jentzsch, A., J. Zhao, O. Hassanzadeh, K.-H. Cheung, M. Samwald and B. Andersson (2009). Linking Open Drug Data. I-SEMANTICS.
151. Jetzek, T. (2015). "Managing complexity across multiple dimensions of liquid open data: The case of the Danish Basic Data Program." *Government Information Quarterly*.
152. Jetzek, T., M. Avital and N. Bjorn-Andersen (2014). "Data-driven innovation through open government data." *Journal of theoretical and applied electronic commerce research* 9(2): 100-120.
153. Jifa, G. and Z. Lingling (2014). "Data, DIKW, Big data and Data science." *Procedia Computer Science* 31: 814-821.
154. Johnson, J.A. (2014). From open data to information justice. *Ethics and Information Technology*, Vol. 16(4), pp. 263-274.
155. Jurisch, M. C., M. Kautz, P. Wolf and H. Krcmar (2015). An international survey of the factors influencing the intention to use open government. *System Sciences (HICSS)*, 2015 48th Hawaii International Conference on, IEEE.
156. Kalampokis, E., A. Karamanou, A. Nikolov, P. Haase, R. Cyganiak, B. Roberts, P. Hermans, E. Tambouris and K. Tarabanis (2014). Creating and utilizing linked open statistical data for the development of advanced analytics services. *Second international workshop for semantic statistics, SemStats2014*. CEUR-WS. org.
157. Kalampokis, E., E. Tambouris and K. Tarabanis (2011). "A classification scheme for open government data: towards linking decentralised data." *International Journal of Web Engineering and Technology* 6(3): 266-285.
158. Kalampokis, E., Nikolov, A., Haase, P., Cyganiak, R., Stasiewicz, A., Karamanou, A., Zotou, M., Zeginis, D., Tambouris, E., Tarabanis, K. (2014) Exploiting Linked Data Cubes with OpenCube Toolkit, *Proc. of the ISWC 2014 Posters and Demos Track a track within 13th International Semantic Web Conference (ISWC2014)*, 19-23 October 2014, Riva del Garda, Italy, CEUR-WS Vol.1272
159. Kalampokis, E., Roberts, B., Karamanou, A., Tambouris, E., & Tarabanis, K. (2015, October). Challenges on Developing Tools for Exploiting Linked Open Data Cubes. In *Proceedings of the 3rd International Workshop on Semantic Statistics (SemStats2015) within the 14th International Semantic Web Conference (ISWC2015)* (Vol. 1551).
160. Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013, September). Linked open government data analytics. In *International Conference on Electronic Government* (pp. 99-110). Springer Berlin Heidelberg.
161. Kalampokis, E., Tambouris, E., & Tarabanis, K. (2016). ICT Tools for Creating, Expanding, and Exploiting Statistical Linked Open Data. *Statistical Journal of the IAOS*.
162. Kalampokis, E., Tambouris, E., Tarabanis, K. (2016). Linked Open Cube Analytics Systems: Potential and Challenges, *IEEE Intelligent Systems*, [in press]



163. Kamateri, E., Kalampokis, E., Tambouris, E., & Tarabanis, K. (2014). The linked medical data access control framework. *Journal of biomedical informatics*, 50, 213-225.
164. Kämpgen, B. (2011, October). DC proposal: online analytical processing of statistical linked data. In *International Semantic Web Conference* (pp. 301-308). Springer Berlin Heidelberg.
165. Kämpgen, B., & Harth, A. (2011, September). Transforming statistical linked data for use in OLAP systems. In *Proceedings of the 7th international conference on Semantic systems* (pp. 33-40). ACM.
166. Kämpgen, B., & Harth, A. (2013, May). No size fits all—running the star schema benchmark with SPARQL and RDF aggregate views. In *Extended Semantic Web Conference* (pp. 290-304). Springer Berlin Heidelberg.
167. Kämpgen, B., & Harth, A. (2014, May). OLAP4LD—A framework for building analysis applications over governmental statistics. In *European Semantic Web Conference* (pp. 389-394). Springer International Publishing.
168. Kämpgen, B., O’Riain, S., & Harth, A. (2012, May). Interacting with statistical linked data via OLAP operations. In *Extended Semantic Web Conference* (pp. 87-101). Springer Berlin Heidelberg.
169. Kämpgen, B., Stadtmüller, S., & Harth, A. (2014, November). Querying the global cube: integration of multidimensional datasets from the web. In *International Conference on Knowledge Engineering and Knowledge Management* (pp. 250-265). Springer International Publishing.
170. Kämpgen, B., Weller, T., O’Riain, S., Weber, C., & Harth, A. (2014, May). Accepting the xbrl challenge with linked data for financial data integration. In *European Semantic Web Conference* (pp. 595-610). Springer International Publishing.
171. Kaplan, A. M. and M. Haenlein (2010). "Users of the world, unite! The challenges and opportunities of Social Media." *Business horizons* 53(1): 59-68.
172. Karamanou, A. (2015). "Towards Exploiting Linked Statistical Open Government Data." *Innovation and the Public Sector*: 319.
173. Kautz, K., Cecez-Kecmanovic, D. (2013). Sociomateriality and Information Systems Success and Failure. In: *Grand Successes and Failures in IT: Public and Private Sectors*, Y.K. Dwivedi, H.Z. Henriksen, D. Wastell, R. De (Eds). *Proceedings of IFIP WG 8.6 International Working Conference on Transfer and Diffusion of IT, TDIT*. Bangalore, India, 27-29 June 2013. Springer, pp. 1-20.
174. Khan, Y., Saleem, M., Iqbal, A., Mehdi, M., Hogan, A., Ngomo, A. C. N., ... & Sahay, R. (2014, December). SAFE: Policy Aware SPARQL Query Federation Over RDF Data Cubes. In *SWAT4LS*.
175. Kim W. , Seo J., Classifying schematic and data heterogeneity in multidatabase systems, *Computer* 24 (12) (1991) pp.12-18. doi:10.1109/2. 116884.
176. King, R. D., M. Liakata, C. Lu, S. G. Oliver and L. N. Soldatova (2011). "On the formalization and reuse of scientific research." *Journal of the Royal Society Interface* 8(63): 1440-1448.
177. Kingdon, J. W. (2003). *Agendas, alternatives, and public policies*, Longman Pub Group.

178. Klímek, J., P. Škoda and M. Nečaský (2016). Requirements on Linked Data Consumption Platform. WWW2016 Workshop: Linked Data on the Web (LDOW2016).
179. Koho, M., Hyvönen, E., & Lehtikainen, A. (2014, May). Ornithology Based on Linking Bird Observations with Weather Data. In European Semantic Web Conference (pp. 75-85). Springer International Publishing.
180. Krishnan, K. (2013). Data warehousing in the age of big data, Newnes.
181. Lapi, E., N. Tcholtchev, L. Bassbouss, F. Marienfeld and I. Schieferdecker (2012). Identification and utilization of components for a linked open data platform. Computer Software and Applications Conference Workshops (COMPSACW), 2012 IEEE 36th Annual, IEEE.
182. Latif, A., A. U. Saeed, P. Höfler, A. Stocker and C. Wagner (2009). The Linked Data Value Chain: A Lightweight Model for Business Engineers. I-SEMANTICS, Citeseer.
183. Lee A., Chen C.-J., Lu H., An aspect of query optimization in multidatabase systems, SIGMOD Rec. 24 (3) (1995) pp. 28-33. doi:10.1145/211990.212011.
184. Lee, G. and Y. H. Kwak (2012). "An open government maturity model for social media-based public engagement." Government Information Quarterly 29(4): 492-503.
185. Lee K.-H., Kim M.-H., Lee K.-C., Kim B.-S., Lee M.-Y., Conflict classification and resolution in heterogeneous information integration based on xml schema, in: TENCON'02. Proceedings. 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, Vol. 1, IEEE, 2002, pp. 93-96.
186. Lefort, L., & Leroux, H. (2013). Design and generation of linked clinical data cubes. In Proceedings of 1st International Workshop on Semantic Statistics (SemStats 2013). Sydney, Australia.
187. Lefort, L., Bobruk, J., Haller, A., Taylor, K., & Woolf, A. (2012, November). A linked sensor data cube for a 100 year homogenised daily temperature dataset. In Proceedings of the 5th International Conference on Semantic Sensor Networks-Volume 904 (pp. 1-16). CEUR-WS. org.
188. Leforta, Laurent, Armin Hallera, Kerry Taylora, and Andrew Woolfb. "The ACORN-SAT linked climate dataset." (2013).
189. Lehmann, J., R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef and S. Auer (2015). "DBpedia—a large-scale, multilingual knowledge base extracted from Wikipedia." Semantic Web 6(2): 167-195.
190. Leroux, H., & Lefort, L. (2012, November). Using CDISC ODM and the RDF Data Cube for the Semantic Enrichment of Longitudinal Clinical Trial Data. In SWAT4LS.
191. Leroux, H., & Lefort, L. (2015). Semantic enrichment of longitudinal clinical study data using the CDISC standards and the semantic statistics vocabularies. Journal of biomedical semantics, 6(1), 1.

192. Liu, F. and X. Li (2011). Using metadata to maintain link integrity for linked data. Internet of Things (iThings/CPSCoM), 2011 International Conference on and 4th International Conference on Cyber, Physical and Social Computing, IEEE.
193. Llaves, A., Corcho, O., & Fernandez-Carrera, A. (2014). Map4rdf-ios: a tool for exploring linked geospatial data. In Proceedings of Workshop on Linked Geospatial Data.
194. Lodi, G., Maccioni, A., Scannapieco, M., Scanu, M., & Tosco, L. (2014). Publishing official classifications in linked open data. In Proceedings of the 2nd International Workshop on Semantic Statistics (SemStats2014) in conjunction with the 13th International Semantic Web Conference (ISWC). Springer, Riva del Garda, Italy.
195. Machado, A. L. and J. M. P. de Oliveira (2011). DIGO: An open data architecture for e-government. 2011 IEEE 15th International Enterprise Distributed Object Computing Conference Workshops, IEEE.
196. Mader, C., Martin, M., & Stadler, C. (2014). Facilitating the exploration and visualization of linked data. In Linked Open Data--Creating Knowledge Out of Interlinked Data (pp. 90-107). Springer International Publishing.
197. Magalhaes, G., C. Roseira and S. Strover (2013). Open government data intermediaries: A terminology framework. Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance, ACM.
198. Maheshwari, D. and M. Janssen (2013). "Measurement and benchmarking foundations: Providing support to organizations in their development and growth using dashboards." Government Information Quarterly 30, Supplement 1: S83-S93.
199. Maheshwari, D. and M. Janssen (2014). Dashboards for supporting organizational development: principles for the design and development of public sector performance dashboards. Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance, ACM.
200. Mangisengi O., Huber J., Hawel C., Essmayr W., A framework for supporting interoperability of data warehouse islands using xml, Data Warehousing and Knowledge Discovery 2114 (2001) pp.328-338. doi:10.1007/3-540-44801-2\_32.
201. Manolea, B. and V. Cretu (2013). Topic Report No. 2013/10: The influence of the Open Government Partnership (OGP) on the Open Data discussions, European Public Sector Information Platform, <http://epsiplatform.eu/topicreports>.
202. Martin, C. (2014). "Barriers to the Open Government Data Agenda: Taking a Multi-Level Perspective." Policy & Internet 6(3): 217-240.
203. Martin, M., Abicht, K., Stadler, C., Ngonga Ngomo, A. C., Soru, T., & Auer, S. (2015, May). Cubeviz: Exploration and visualization of statistical linked data. In Proceedings of the 24th International Conference on World Wide Web (pp. 219-222). ACM.
204. Martin, M., van Nuffelen, B., Abruzzini, S., & Auer, S. (2012). The digital agenda scoreboard: A statistical anatomy of europe's way into the information age. Semantic Web Journal.

205. Martin, S., Foulonneau, M., Turki, S., Ihadjadene, M., Paris, U., Tudor, P.R.C.H. (2013). Risk Analysis to Overcome Barriers to Open Data. *Electronic Journal of e-Government*, Vol. 11(1), pp. 348-359.
206. Maté, A., Llorens, H., & de Gregorio, E. (2012, October). An integrated multidimensional modeling approach to access big data in business intelligence platforms. In *International Conference on Conceptual Modeling* (pp. 111-120). Springer Berlin Heidelberg.
207. Matei, A., Chao, K. M., & Godwin, N. (2015). OLAP for Multidimensional Semantic Web Databases. In *Enabling Real-Time Business Intelligence* (pp. 81-96). Springer Berlin Heidelberg.
208. Matheus, R., F. Angélico and M. I. Atoji (2014). Dados Abertos no Jornalismo: Os Limites e os Desafios das Estratégias de Uso e Criação de Cadeia de Valor Social incentivando a transparência e controle social na América Latina. OD4D, 2014.
209. Matheus, R., M. M. Ribeiro and J. C. Vaz (2015). Brazil Towards Government 2.0: Strategies for Adopting Open Government Data in National and Subnational Governments. *Case Studies in e-Government 2.0*, Springer: 121-138.
210. Matheus, R. and M. Janssen (2013). Transparency of civil society websites: towards a model for evaluation websites transparency. *Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance*, ACM.
211. Matheus, R. and M. Janssen (2016). Exploitation and Exploration Strategies to Create Data Transparency in the Public Sector. *Proceedings of the 9th international conference on theory and practice of electronic governance*.
212. McCusker, J. P., McGuinness, D. L., Lee, J., Thomas, C., Courtney, P., Tatalovich, Z., ... & Shaikh, A. (2013, January). Towards next generation health data exploration: a data cube-based investigation into population statistics for tobacco. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on* (pp. 2725-2732). IEEE.
213. McGuinness, D. L. and F. Van Harmelen (2004). "OWL web ontology language overview." *W3C recommendation 10(10)*: 2004.
214. McLean, T., L. Mark, M. Loper and D. Rosenbaum (1998). Applying temporal databases to HLA data collection and analysis. *Proceedings of the 30th conference on Winter simulation*, IEEE Computer Society Press.
215. Mehdi, M., Sahay, R., Derguech, W., & Curry, E. (2013, October). On-the-fly generation of multidimensional data cubes for web of things. In *Proceedings of the 17th International Database Engineering & Applications Symposium* (pp. 28-37). ACM.
216. Meimaris, M., & Papastefanatos, G. (2014). Containment and complementarity relationships in multidimensional linked open data. *Semantic Statistics (SEMSTATS)*.
217. Meimaris, M., Papastefanatos, G., Vassiliadis, P., & Anagnostopoulos (2016) I. Efficient Computation of Containment and Complementarity in RDF Data Cubes, *International Conference on Extending Database Technology (EDBT)*.

218. Meroño-Peñuela, A. (2014, November). LSD dimensions: Use and reuse of linked statistical data. In *International Conference on Knowledge Engineering and Knowledge Management* (pp. 159-163). Springer International Publishing.
219. Meroño-Peñuela, A., Ashkpour, A., & Guéret, C. (2014). From Flat Lists to Taxonomies: Bottom-up Concept Scheme Generation in Linked Statistical Data. In *Proceedings of the 2nd International Workshop on Semantic Statistics (Sem-Stats 2014)*. International Semantic Web Conference (ISWC). CEUR Workshop Proceedings.
220. Meroño-Peñuela, A., Ashkpour, A., Rietveld, L., Hoekstra, R., & Schlobach, S. (2012). Linked Humanities Data: The Next Frontier?. In *A Case-study in Historical Census Data. Proceedings of the 2nd International Workshop on Linked Science* (Vol. 951, p. 2012).
221. Meroño-Peñuela, A., Guéret, C., Ashkpour, A., & Schlobach, S. (2015a). Cedar: The dutch historical censuses as linked open data. *Semantic Web—Interoperability, Usability, Applicability*.
222. Meroño-Peñuela, A., Guéret, C., Hoekstra, R., & Schlobach, S. (2013). Detecting and reporting extensional concept drift in statistical linked data. In *1st International Workshop on Semantic Statistics (SemStats 2013)*, ISWC. CEUR.
223. Meroño-Peñuela, A., Guéret, C., Schlobach, S., Capadislí, S., Cotton, F., Haller, A., ... & Troncy, R. (2015b). Linked Edit Rules: A Web Friendly Way of Checking Quality of RDF Data Cubes. In *Third International Workshop on Semantic Statistics (SemStats 2015)*. CEUR-WS. org.
224. Miloevi, U., Janev, V., Spasi, M., Milojkovi, J., & Vrane, S. (2012). Publishing statistical data as linked open data. In *Proceedings of the 2nd International Conference on Information Society Technology, Information Society of the Republic of Serbia*.
225. Moore M., Hartley J. (2008). Innovations in governance. *Public Management Review*, Vol. 10, pp. 3-20.
226. Munné, R. (2013). BIG Work in Progress: Big Data Public Private Forum and Public Sector. *European Conference on e-Government, Academic Conferences International Limited*.
227. Mutlu, B., Hoefler, P., Sabol, V., Tschinkel, G., & Granitzer, M. (2013). Automated Visualization Support for Linked Research Data. *I-SEMANTICS (Posters & Demos)*, 1026, 40-44.
228. Mutlu, B., Hoefler, P., Tschinkel, G., Veas, E., Sabol, V., Stegmaier, F., & Granitzer, M. (2014, January). Suggesting visualisations for published data. In *Information Visualization Theory and Applications (IVAPP), 2014 International Conference on* (pp. 267-275). IEEE.
229. Mynarz, J., Cyganiak, R., Hausenblas, M., & Iqbal, A. (2011). Modelling of statistical linked data. *Proceedings of Znalosti 2011*.
230. Nasi, G., Cucciniello, M., Mele, V., Valotti, G., Bazurli, R., Vries, H., Bekkers, V., Tummers, L., Gascó, M., Ysa, T., Fernández, C., Albareda, A., Matei, A., Savulescu, C., Antonie, C., Balaceanu, E.B., Nemec, J., Svidroňová, M., Mikusova Merickova B., Oviska, M., Mendes, C., Eymeri-Douzans M, Montheubert, E.M. (2015). Determinants and Barriers of Adoption, Diffusion and Upscaling of ICT-driven Social Innovation in the Public Sector: A Comparative Study Across 6 EU Countries. LIPSE research report. Available at:

[http://www.lipse.org/userfiles/uploads/Research%20Report%20LIPSE%20WP5\\_final\\_20150530.pdf](http://www.lipse.org/userfiles/uploads/Research%20Report%20LIPSE%20WP5_final_20150530.pdf)

231. Neumayr A., Schre M., Thalheim B., Hetero-homogeneous hierarchies in data warehouses, in: Song I.-Y., Golfarelli M. (Eds.), Proc. 7th Asia-Pacific Conference on Conceptual Modelling, Brisbane, Australia, 2010
232. Nevarez, C. A. and K. P. White (1998). Method and system for integrating additional functionality into a login system, Google Patents.
233. Nguyen, T. B., & Ngo, S. N. (2014, December). Semantic cubing platform enabling interoperability analysis among cloud-based linked data cubes. In Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services (pp. 547-553). ACM.
234. NISO (2004). "Understanding metadata." National Information Standards 20.
235. Nonthakarn, C. and V. Wuwongse (2012). Linked OpenScholar: A Researcher Network Using Linked Open Data. International Conference on Asian Digital Libraries, Springer.
236. Norway, S. (1998). "Guidelines for statistical metadata on the Internet." Statistical Journal of the United Nations Economic Commission for Europe 15(2): 169-176.
237. Nushi, B., B. Van Loenen and J. Cromptoets (2015). "The STIG: A new SDI assessment method." International Journal of Spatial Data Infrastructures Research 10, 2015.
238. O'Hara, K. (2012). Transparency, open data and trust in government: shaping the infosphere. In: Proceedings of the 4th Annual ACM Web Science Conference, ACM, pp. 223-232.
239. O'Riain, S., E. Curry and A. Harth (2012). "XBRL and open data for global financial ecosystems: A linked data approach." International Journal of Accounting Information Systems 13(2): 141-162.
240. Obama, B. (2009). "Transparency and open government." Memorandum for the heads of executive departments and agencies.
241. OECD. (1997). Measuring Public Employment in OECD Countries: Sources, Methods and Results, Paris: OECD.
242. OECD. (2005). Guidelines for collecting and interpreting innovation data. Paris: OECD.
243. OECD. (2011). Innovation in public service delivery. Context, Solutions and Challenges. Paris: OECD.
244. OECD. (2014). Recommendation of the Council on Digital Government Strategies. Paris: OECD.
245. Ojo, A. and M. Janssen (2013). Aligning core stakeholders' perspectives and issues in the open government data community. Proceedings of the 14th Annual International Conference on Digital Government Research, ACM.
246. OKF. (2015). "Open definition." from <http://opendefinition.org/od/2.1/en/>.
247. OpenGovData.org (2007). "Principles of open Government data." OpenGovData. org.

248. Pallickara, S. L., S. Pallickara and M. Zupanski (2012). "Towards efficient data search and subsetting of large-scale atmospheric datasets." *Future Generation Computer Systems* 28(1): 112-118.
249. Park, Y. R., H. H. Kim, H. J. Seo and J. H. Kim (2011). "CDISC Transformer: a metadata-based transformation tool for clinical trial and research data into CDISC standards." *TIIS* 5(10): 1830-1840.
250. Patton, E. W., Brown, E., Poegel, M., De Los, H., Santos, C. F., Bennett, K. P., & McGuinness, D. L. *SemNExT: A Framework for Semantically Integrating and Exploring Numeric Analyses*.
251. Peixoto, T. (2013). "The Uncertain Relationship between Open Data and Accountability: A Response to Yu and Robinson's 'The New Ambiguity of Open Government'."
252. Perakis, K., Bouras, T., Ntalaperas, D., Hasapis, P., Georgousopoulos, C., Sahay, R., ... & Usurelu, D. (2013, October). Advancing patient record safety and EHR semantic interoperability. In *2013 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 3251-3257). IEEE.
253. Petrou, I., Meimaris, M., & Papastefanatos, G. (2014). Towards a methodology for publishing Linked Open Statistical Data. *JeDEM-eJournal of eDemocracy and Open Government*, 6(1), 97-105.
254. Petrou, I., Papastefanatos, G., & Dalamagas, T. (2013, June). Publishing census as linked open data: a case study. In *Proceedings of the 2nd International Workshop on Open Data* (p. 4). ACM.
255. Petticrew, M. and H. Roberts (2008). *Systematic reviews in the social sciences: A practical guide*, John Wiley & Sons.
256. Pinsker, R. and S. Li (2008). "Costs and benefits of XBRL adoption: Early evidence." *Communications of the ACM* 51(3): 47-50.
257. Pirrotta, G. (2010, September). Linking Italian university statistics. In *Proceedings of the 6th International Conference on Semantic Systems* (p. 2). ACM.
258. Pollitt, C., Bouckaert, G. (2011). *Public Management Reform: A Comparative Analysis: New Public Management, Governance, and the Neo-Weberian State*. Oxford: Oxford University Press.
259. Prat, N., Akoka, J., & Comyn-Wattiau, I. (2012, May). Transforming multidimensional models into OWL-DL ontologies. In *2012 Sixth International Conference on Research Challenges in Information Science (RCIS)* (pp. 1-12). IEEE.
260. Prat, N., Akoka, J., & Comyn-Wattiau, I. (2012b). Transforming multidimensional models into OWL-DL ontologies. In *2012 Sixth International Conference on Research Challenges in Information Science (RCIS)* (pp. 1-12). IEEE.
261. Prat, N., Megdiche, I., and Akoka, A. (2012a). Multidimensional models meet the semantic web: defining and reasoning on OWL-DL ontologies for OLAP. In *Proceedings of the fifteenth international workshop on Data warehousing and OLAP (DOLAP '12)*. ACM, New York, NY, USA, 17-24. DOI=<http://dx.doi.org/10.1145/2390045.2390049>

- 262. Ram S., Park J., Semantic conflict resolution ontology (scrol): An ontology for detecting and resolving data and schema-level semantic conflicts, *IEEE Transactions on Knowledge and Data Engineering* 16 (2) (2004) pp.189-202.
- 263. Rahm, E. and H. H. Do (2000). "Data cleaning: Problems and current approaches." *IEEE Data Eng. Bull.* 23(4): 3-13.
- 264. Reddy M., Prasad B. E., Reddy P., Gupta A., A methodology for integration of heterogeneous databases, *IEEE Transactions on Knowledge and Data Engineering* 6 (6) (1994) pp.920-933.
- 265. Richardson, L., M. Amundsen and S. Ruby (2013). *RESTful Web APIs*, " O'Reilly Media, Inc."
- 266. Richardson, L. and S. Ruby (2008). *RESTful web services*, " O'Reilly Media, Inc."
- 267. Ristoski, P., Bizer, C., & Paulheim, H. (2015). Mining the web of linked data with rapidminer. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35, 142-151.
- 268. Robinson, D. G., H. Yu, W. P. Zeller and E. W. Felten (2009). "Government data and the invisible hand." *Yale Journal of Law & Technology* 11: 160.
- 269. Rochet, C., Peignot, J., Peneranda, A. (2012). Digitizing the Public Organization: Information System Architecture as a Key Competency to Foster Innovation Capabilities in Public Administration. *Halduskultuur – Administrative Culture*, Vol. 13, No. 1, pp. 49-66.
- 270. Rodríguez, J. M. A., Clement, J., Gayo, J. E. L., Farhan, H., & De Pablos, P. O. (2013). Publishing statistical data following the linked open data principles: The web index project. *IGI Global*, 199-226.
- 271. Roussakis, Y., Chrysakis, I., Stefanidis, K., Flouris, G., & Stavarakas, Y. (2015). A flexible framework for defining, representing and detecting changes on the data web. *arXiv preprint arXiv:1501.02652*.
- 272. Rowley, J. E. (2007). "The wisdom hierarchy: representations of the DIKW hierarchy." *Journal of information science*.
- 273. Ruback, L., Pesce, M., Manso, S., Ortiga, S., Salas, P. E. R., & Casanova, M. A. (2013, March). A mediator for statistical linked data. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing* (pp. 339-341). ACM.
- 274. Saad, R., TROJAHN, C., & TESTE, O. (2013). OLAP manipulations on RDF data following a constellation model. In *First International Workshop on Semantic Statistics, collocated with the 12th International Semantic Web Conference, Sydney*, page (on line). DataLift.
- 275. Sabol, V., Tschinkel, G., Veas, E., Hoefler, P., Mutlu, B., & Granitzer, M. (2014, October). Discovery and visual analysis of linked data for humans. In *International Semantic Web Conference* (pp. 309-324). Springer International Publishing.
- 276. Sabou, M., Braşoveanu, A. M., & Önder, I. (2015). Linked data for cross-domain decision-making in tourism. In *Information and Communication Technologies in Tourism 2015* (pp. 197-210). Springer International Publishing.



- 277. Salas, P. E. R., Martin, M., Da Mota, F. M., Auer, S., Breitman, K., & Casanova, M. A. (2012a, September). Publishing statistical data on the web. In *Semantic Computing (ICSC), 2012 IEEE Sixth International Conference on* (pp. 285-292). IEEE.
- 278. Salas, P. E. R., Martin, M., Da Mota, F. M., Auer, S., Breitman, K. K., & Casanova, M. A. (2012b, June). Olap2datacube: An ontowiki plug-in for statistical data publishing. In *Proceedings of the Second International Workshop on Developing Tools as Plug-Ins* (pp. 79-83). IEEE Press.
- 279. Sato, H., & Wen, W. Towards Easy Matching Between Statistical Linked Data: Dimension Patterns. *Semstats 2013*.
- 280. Sauer, C. (1993). *Why Information Systems Fail: A Case Study Approach*. Oxfordshire: Alfred Waller Publishers.
- 281. Sboui T., Bedard Y. , Brodeur J., Badard T., A conceptual framework to support semantic interoperability of geospatial datacubes 4802 (2007) pp. 378-387. doi:10.1007/978-3-540-76292-8\_44.
- 282. Schlegel, K., Bayerl, S., Zwicklbauer, S., Stegmaier, F., Seifert, C., Granitzer, M., & Kosch, H. (2013, May). Trusted Facts: Triplifying Primary Research Data Enriched with Provenance Information. In *Extended Semantic Web Conference* (pp. 268-270). Springer Berlin Heidelberg.
- 283. Schuurman, N., A. Deshpande and D. M. Allen (2008). "Data integration across borders: a case study of the Abbotsford-Sumas aquifer (British Columbia/Washington State) 1." *JAWRA Journal of the American Water Resources Association* 44(4): 921-934.
- 284. Schütz, C., Neumayr, B., & Schrefl, M. (2013, June). Business model ontologies in OLAP cubes. In *International Conference on Advanced Information Systems Engineering* (pp. 514-529). Springer Berlin Heidelberg.
- 285. Seifert, C., Granitzer, M., Höfler, P., Mutlu, B., Sabol, V., Schlegel, K., ... & Kern, R. (2013). Crowdsourcing fact extraction from scientific literature. In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data* (pp. 160-172). Springer Berlin Heidelberg.
- 286. Sen, A. (2004). "Metadata management: past, present and future." *Decision Support Systems* 37(1): 151-173.
- 287. Shadbolt, N., K. O'Hara, T. Berners-Lee, N. Gibbins, H. Glaser and W. Hall (2012). "Linked open government data: Lessons from data. gov. uk." *IEEE Intelligent Systems* 27(3): 16-24.
- 288. Shadbolt, N. and K. O'Hara (2013). "Linked data in government." *IEEE Internet Computing*(4): 72-77.
- 289. Sheth, A. P. (1999). Changing focus on interoperability in information systems: from system, syntax, structure to semantics. *Interoperating geographic information systems*, Springer: 5-29.
- 290. Sheth A.P., Kashyap V., So far (schematically) yet so near (semantically), in: *Proceedings of the IFIP WG2: Conference on Semantics of Interoperable Database Systems*, Lorne, Victoria, Australia, 1992, pp. 283-312.

291. Sigurbjörnsson, B. and R. Van Zwol (2008). Flickr tag recommendation based on collective knowledge. Proceedings of the 17th international conference on World Wide Web, ACM.
292. Sørensen E., Torfing J. (2011). Enhancing Collaborative Innovation in the Public Sector. Administration & Society, Vol. 43, pp. 842-868.
293. Southall, H. R., & Stoner, M. J. (2015). Creating a spatio-temporal “Data Feed” API for a large and diverse library of historical statistics for areas within Britain.
294. Spaccapietra S. , Parent C. , Dupont Y., Model independent assertions for integration of heterogeneous schemas, The VLDB Journal 1 (1) (1992) 81-126.
295. Staab, S. and R. Studer (2013). Handbook on ontologies, Springer Science & Business Media.
296. Stadler, C., J. Lehmann, K. Höffner and S. Auer (2012). "Linkedgeodata: A core for a web of spatial open data." Semantic Web 3(4): 333-354.
297. Stegmaier, F., Seifert, C., Kern, R., Höfler, P., Bayerl, S., Granitzer, M., ... & Schlegel, K. (2014). Unleashing semantics of research data. In Specifying Big Data Benchmarks (pp. 103-112). Springer Berlin Heidelberg.
298. Stonebraker, M., D. Bruckner, I. F. Ilyas, G. Beskales, M. Cherniack, S. B. Zdonik, A. Pagan and S. Xu (2013). Data Curation at Scale: The Data Tamer System. CIDR.
299. Tambouris, E., Kalampokis, E., & Tarabanis, K. (2015, August). Processing Linked Open Data Cubes. In International Conference on Electronic Government (pp. 130-143). Springer International Publishing.
300. Tarasova, T., Argenti, M., & Marx, M. (2013, October). Semantically-Enabled Environmental Data Discovery and Integration: Demonstration Using the Iceland Volcano Use Case. In International Conference on Knowledge Engineering and the Semantic Web (pp. 289-297). Springer Berlin Heidelberg.
301. Tenopir, C., S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, M. Manoff and M. Frame (2011). "Data sharing by scientists: practices and perceptions." PloS one 6(6): e21101.
302. Thanos, C. (2013). "A vision for global research data infrastructures." Data Science Journal 12(0): 71-90.
303. Thellmann, K., F. Orlandi and S. Auer (2014). LinDA-Visualising and Exploring Linked Data. Proceedings of the Posters and Demos Track of 10th International Conference on Semantic Systems-SEMANTICS2014, Leipzig, Germany, Citeseer.
304. Thomas, S., Marath, B. (2013). An Integrative Model Linking Risk, Risk Management and Project Performance: Support from Indian Software Projects. In “Grand Successes and Failures in IT: Public and Private Sectors”, Y.K. Dwivedi, H.Z. Henriksen, D. Wastell, R. De (Eds), Proceedings of IFIP WG 8.6 International Working Conference on Transfer and Diffusion of IT, TDIT 2013. Bangalore, India, 27-29 June 2013. Springer, pp. 326-342.

305. Tilahun, B., Kauppinen, T., Keßler, C., & Fritz, F. (2014). Design and development of a linked open data-based health information representation and visualization system: Potentials and preliminary evaluation. *JMIR medical informatics*, 2(2).
306. Tjondronegoro, D. and A. Spink (2008). "Web search engine multimedia functionality." *Information Processing & Management* 44(1): 340-357.
307. Torlone R., Interoperability in data warehouses, in: *Encyclopedia of Database Systems*, Springer, 2009, pp. 1560-1564.
308. Trinh, T. D., Do, B. L., Wetz, P., Anjomshoa, A., & Tjoa, A. M. (2013, December). Linked widgets: An approach to exploit open government data. In *Proceedings of International Conference on Information Integration and Web-based Applications & Services* (p. 438). ACM.
309. Tseng A.S. , Chen C.-W., Integrating heterogeneous data warehouses using xml technologies, *Journal of Information Science* 31 (3) (2005) pp.209-229. doi:10.1177/0165551505052467
310. Tschinkel, G., Veas, E., Mutlu, B., & Sabol, V. (2014, October). Using semantics for interactive visual analysis of linked open data. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272* (pp. 133-136). CEUR-WS. org.
311. Ubaldi, B. (2013). "Open Government Data."
312. Vaisman, A., & Zimányi, E. (2014). Data Warehouses and the Semantic Web. In *Data Warehouse Systems* (pp. 539-576). Springer Berlin Heidelberg.
313. van der Waal, S., Węcel, K., Ermilov, I., Janev, V., Milošević, U., & Wainwright, M. (2014). Lifting open data portals to the data web. In *Linked Open Data--Creating Knowledge Out of Interlinked Data* (pp. 175-195). Springer International Publishing.
314. Van Nuffelen, B., Janev, V., Martin, M., Mijovic, V., & Tramp, S. (2014). Supporting the linked data life cycle using an integrated tool stack. In *Linked Open Data--Creating Knowledge Out of Interlinked Data* (pp. 108-129). Springer International Publishing.
315. Vardaki, M., H. Papageorgiou and F. Pentaris (2009). "A statistical metadata model for clinical trials' data management." *Computer methods and programs in biomedicine* 95(2): 129-145.
316. Veenstra, A. F., Broek, T.A. (2013). Opening moves--drivers, enablers and barriers of open data in a semi-public organization. In *Electronic Government*. Berlin Heidelberg: Springer, pp. 50-61.
317. Verhoest, K., Verschuere, B., Bouckaert, G., Peter, G.B. (2006). Innovative Public Sector Organizations. In: C. Campell et al., *Comparative Trends in Public Management. Smart Practices Toward Blending Policy and Administration*. Ottawa: Canada School of Public Service, pp. 106-118.
318. Vilches-Blázquez, L. M., Villazón-Terrazas, B., Corcho, O., & Gómez-Pérez, A. (2014). Integrating geographical information in the Linked Digital Earth. *International Journal of Digital Earth*, 7(7), 554-575.
319. Villazón-Terrazas, B., L. M. Vilches-Blázquez, O. Corcho and A. Gómez-Pérez (2011). Methodological guidelines for publishing government linked data. *Linking government data*, Springer: 27-49.

- 320. Von Ahn, L., B. Maurer, C. McMillen, D. Abraham and M. Blum (2008). "recaptcha: Human-based character recognition via web security measures." *Science* 321(5895): 1465-1468.
- 321. Voorberg, W. H., V. J. J. M. Bekkers, L. G. Tummers (2014) A Systematic Review of Co-Creation and Co-Production: Embarking on the social innovation journey. *Public Management Review*, published online 30 June 2014.
- 322. Vrandečić, D., Lange, C., Hausenblas, M., Bao, J., & Ding, L. (2010). Semantics of Governmental Statistics Data. *Proceedings of WebSci': Extending the Frontiers of Society On-Line..*: <http://journal.webscience.org/400/>.(Cit. on pp.,,).
- 323. W3C. (2011). "GLD Life cycle." Retrieved 06/07/2016, 2016, from [https://www.w3.org/2011/gld/wiki/GLD\\_Life\\_cycle](https://www.w3.org/2011/gld/wiki/GLD_Life_cycle).
- 324. Wagner, A., Haase, P., Rettinger, A., & Lamm, H. (2013). Discovering related data sources in data-portals. In *First International Workshop on Semantic Statistics*.
- 325. Webster J., Watson R. T., Analyzing the past to prepare for the future: Writing a literature review, *Management Information Systems Quarterly* 26 (2) (2002) 3.
- 326. Winn, J. (2013). "Open data and the academy: An evaluation of CKAN for research data management."
- 327. Wood, D., M. Zaidman, L. Ruth and M. Hausenblas (2014). *Linked Data*, Manning Publications Co.
- 328. XBRL-International (2013). *XBRL Specification 2.1*.
- 329. Xiong, J., Y. Hu, G. Li, R. Tang and Z. Fan (2011). "Metadata distribution and consistency techniques for large-scale cluster file systems." *IEEE Transactions on Parallel and Distributed Systems* 22(5): 803-816.
- 330. Yang, T.-M. and Y.-J. Wu (2016). "Examining the socio-technical determinants influencing government agencies' open data publication: A study in Taiwan." *Government Information Quarterly*.
- 331. Yu, H., Robinson, D.G. (2012). The New Ambiguity of 'Open Government. *UCLA Law Review Discourse* 59, pp. 178-208. Available at <http://www.uclalawreview.org/the-new-ambiguity-of-%E2%80%9Copen-government%E2%80%9D/>
- 332. Zancanaro, A., Pizzol, L. D., de Moura Speroni, R., Todesco, J. L., & Gauthier, F. O. (2013). Publishing multidimensional statistical linked data. In *Proceedings of the Fifth International Conference on Information, Process, and Knowledge Management* (pp. 290-304).
- 333. Zapolko, B., & Mathiak, B. (2011, September). Performing statistical methods on linked data. In *International conference on dublin core and metadata applications* (pp. 116-125).
- 334. Zapolko, B., & Mathiak, B. (2014, May). Object property matching utilizing the overlap between imported ontologies. In *European Semantic Web Conference* (pp. 737-751). Springer International Publishing.

- 335. Zaveri, A., A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, S. Auer and P. Hitzler (2013). "Quality assessment methodologies for linked open data." Submitted to Semantic Web Journal.
- 336. Zaveri, A., Lehmann, J., Auer, S., Hassan, M. M., Sherif, M. A., & Martin, M. (2013). Publishing and interlinking the global health observatory dataset. *Semantic Web*, 4(3), 315-322. (pp. 108-129). Springer International Publishing.
- 337. Zaveri, A., Pietrobon, R., Auer, S., Lehmann, J., Martin, M., & Ermilov, T. (2011, August). ReDD-Observatory: Using the web of data for evaluating the research-disease disparity. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01* (pp. 178-185). IEEE Computer Society.
- 338. Zuiderwijk, A. (2015). Open data infrastructures: The design of an infrastructure to enhance the coordination of open data use, TU Delft, Delft University of Technology.
- 339. Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., Alibaks, R.S. (2012). Socio-technical impediments of open data. *Electronic Journal of e-Government*, Vol. 10(2), pp. 156-172.
- 340. Zuiderwijk, A., K. Jeffery and M. Janssen (2012). "The potential of metadata for linked open data and its value for users and publishers." *JeDEM-e-Journal of e-Democracy and Open Government*, 4 (2) 2012.
- 341. Zuiderwijk, A., M. Janssen, S. Choenni and R. Meijer (2014). "Design principles for improving the process of publishing open data." *Transforming Government: People, Process and Policy* 8(2): 185-204.
- 342. Zuiderwijk, A., M. Janssen and A. Parnia (2013). The complementarity of open data infrastructures: an analysis of functionalities. *Proceedings of the 14th Annual International Conference on Digital Government Research*, ACM.
- 343. Zuiderwijk, A., M. Janssen and K. Jeffery (2013). Towards an e-infrastructure to support the provision and use of open data. *Conference for E-Democracy and Open Government*.
- 344. Zuiderwijk, A. and M. Janssen (2014). The negative effects of open government data- investigating the dark side of open data. *Proceedings of the 15th Annual International Conference on Digital Government Research*, ACM.
- 345. Zuiderwijk-van Eijk, A. and M. Janssen (2015). Participation and Data Quality in Open Data use: Open Data Infrastructures Evaluated. *Proceedings of the 15th European Conference on e-Government*, Portsmouth, UK, 18-19 June 2015; Authors version, ACPI.

## Appendix A: Questionnaire for PA's challenges

### Questionnaire for Public Administration Representatives

#### General information

*The information you provide in this section will only be used as background information for the research team at Tallinn University of Technology and will not be published nor made available to any other parties.*

**Country:**

- ☐ Belgium
- ☐ Estonia
- ☐ Lithuania
- ☐ Greece
- ☐ Ireland
- ☐ UK

**Name of your organisation:** \_\_\_\_\_**Level of government:**

- ☐ Central/federal
- ☐ Regional
- ☐ Local/municipal

**Your position/function in the organisation:** \_\_\_\_\_**E-mail:** \_\_\_\_\_**Phone number (optional):** \_\_\_\_\_

#### II Survey questions

1. In what ways and for what purposes (if at all) has your organisation used open data in its work? By open data we refer to data that is presented in a machine-readable format and can be freely used, re-used and redistributed by anyone. Open data can come from different sources and comprise different kinds of data (financial, statistical, scientific, cultural, etc.)
2. To your knowledge, has your organisation initiated or participated in the co-creation of public services using open data? By co-creation we mean the direct involvement of individual users, groups of citizens and other stakeholders in the planning and delivery of public services.
  - ☐ Yes (please go to Question 3.a)
  - ☐ No (please go to Question 3.b)
3. (a) What are the main reasons why your organisation decided to initiate or participate in the co-creation of public services using open data? (Please skip this question and go to Question 3.(b) if you answered „No“ to Question 2)  
(b) What are the main reasons why your organisation has not initiated or participated in the co-creation of public services using open data? (Please skip this question and go to Question 3.(a) if you answered „Yes“ to Question 2)

4. In your view, what are the main barriers that hinder the use of open data for the co-creation of public services? Name up to 3 barriers that you consider the most important.
5. In your view, what are the main drivers that enable and stimulate the use of open data for the co-creation of public services? Name up to 3 drivers that you consider the most important.
6. In your view, what are the main needs (missing capacities) in your organisation that would help you open up data and build innovative services based on open data?
7. How would you assess the current level of involvement of citizens and businesses (and their representative organisations) in open data-driven co-creation of public services in your country?
8. In your opinion, in what ways could the public administration use open data to come up with better decisions?
9. Do you know of any successful policies or initiatives (in your country or elsewhere) that have been used to promote open data-driven innovation in public services and decision-making? Please give a brief description. What do you think were the reasons why these measures were successful?
10. Do you know of any unsuccessful policies or initiatives that have been tried out to promote open data-driven innovation in public services and decision-making? Please give a brief description. What do you think were the reasons these initiatives failed?
11. In your opinion, what policy instruments or initiatives are currently missing (in your country or at the EU level) that would help advance the use of open data for the co-creation of innovative services?

### **Questionnaire for Non-Governmental Stakeholders**

#### **General information**

*The information you provide in this section will only be used as background information for the research team at Tallinn University of Technology and will not be published nor made available to any other parties.*

#### **Country:**

- ☐ Belgium
- ☐ Estonia
- ☐ Lithuania
- ☐ Greece
- ☐ Ireland
- ☐ UK

**Name of your organisation:** \_\_\_\_\_

#### **Type of organisation:**

- ☐ University/research institution
- ☐ Private company
- ☐ Non-governmental/civil society organisation
- ☐ Other (please specify): \_\_\_\_\_

**Your position/function in the organisation:** \_\_\_\_\_

E-mail: \_\_\_\_\_

Phone number (optional): \_\_\_\_\_

## II Survey questions

1. In what ways and for what purposes (if at all) has your organisation used open data in its work?  
By open data we refer to data that is presented in a machine-readable format and can be freely used, re-used and redistributed by anyone. Open data can come from different sources and comprise different kinds of data (financial, statistical, scientific, cultural, etc.)
2. To your knowledge, has your organisation initiated or participated in the co-creation of public services using open data? By co-creation we mean the direct involvement of individual users, groups of citizens and other stakeholders in the planning and delivery of public services.  
  
☐ Yes (please go to Question 3.a)  
  
☐ No (please go to Question 3.b)
3. (a) What are the main reasons why your organisation decided to initiate or participate in the co-creation of public services using open data? (Please skip this question and go to Question 3.(b) if you answered „No“ to Question 2)  
(b) What are the main reasons why your organisation has not initiated or participated in the co-creation of public services using open data? (Please skip this question and go to Question 3.(a) if you answered „Yes“ to Question 2)
4. In your view, what are the main barriers that hinder the use of open data for the co-creation of public services? Name up to 3 barriers that you consider the most important.
5. In your view, what are the main drivers that enable and stimulate the use of open data for the co-creation of public services? Name up to 3 drivers that you consider the most important.
6. In your view, what are the main needs (missing capacities) in your organisation that would help you open up data and build innovative services based on open data?
7. How would you assess the current level of involvement of citizens and businesses (and their representative organisations) in open data-driven co-creation of public services in your country?
8. In your opinion, in what ways could the public administration use open data to come up with better decisions?
9. Do you know of any successful policies or initiatives (in your country or elsewhere) that have been used to promote open data-driven innovation in public services and decision-making? Please give a brief description. What do you think were the reasons why these measures were successful?
10. Do you know of any unsuccessful policies or initiatives that have been tried out to promote open data-driven innovation in public services and decision-making? Please give a brief description. What do you think were the reasons these initiatives failed?
11. In your opinion, what policy instruments or initiatives are currently missing (in your country or at the EU level) that would help advance the use of open data for the co-creation of innovative services?



## Appendix B: Questionnaire for LOSD technical challenges

The questionnaire used for identifying needs of experts regarding Linked Open Statistical Data.

**Q1: A cube contains a set of measures that represent the phenomena being observed. In your opinion what is the best approach to define a measure?**

1. Re-use sdmx-measure:obsValue e.g.:

```
eg:obs1 a qb:Observation;  
    sdmx-measure:obsValue "0.17"^^xsd:double.
```

2. Define a new measure based on types of units (e.g. count, ratio). This measure is also defined as subproperty of sdmx-measure:obsValue. Use the sdmx-attribute:unitMeasure to specify the type of measure e.g.:

```
eg:ratio a qb:MeasureProperty;  
    rdfs:subPropertyOf sdmx-measure:obsValue.  
eg:obs1 a qb:Observation;  
    eg:ratio "0.17"^^xsd:double;  
    sdmx-attribute:unitMeasure eg:unemployment.
```

3. Define a new measure. This measure is also defined as subproperty of sdmx-measure:obsValue e.g.:

```
eg:unemployment a qb:MeasureProperty;  
    rdfs:subPropertyOf sdmx-measure:obsValue.  
eg:obs1 a qb:Observation;  
    eg:unemployment "0.17"^^xsd:double.
```

4. Other (please describe)

**Q2: If you have data with more than one measure, which approach would you consider the best?**

1. Create several data cubes with one measure each
2. Create one data cube with multiple measure

**Q3: In case of multiple measures per cube which is the best approach to define them?**

1. *Multi-measure observations* (proposed by QB vocabulary). Define multiple qb:MeasureProperty at the data structure definition and use all measures to every observation e.g.:

```
eg:unemployment a qb:MeasureProperty.  
eg:poverty a qb:MeasureProperty.  
eg:obs1 a qb:Observation;  
    eg:unemployment "0.17"^^xsd:double;  
    eg:poverty "0.25"^^xsd:double.
```

2. *Measure dimension* (proposed by QB vocabulary): Define multiple qb:MeasureProperty at the data structure definition and use an extra dimension, the qb:measureType, to denote which particular qb:MeasureProperty is being conveyed by the observation e.g.:

```
eg:unemployment a qb:MeasureProperty.  
eg:poverty a qb:MeasureProperty.  
eg:obs1 a qb:Observation;  
    eg:unemployment "0.17"^^xsd:double;  
    qb:measureType eg:unemployment.  
eg:obs2 a qb:Observation;  
    eg:poverty "0.25"^^xsd:double;  
    qb:measureType eg:poverty.
```

3. *Indicator dimension*. Use sdmx-measure:obsValue along with a qb:DimensionProperty that indicates the measure being conveyed by the observation e.g.:

```
eg:indicator a qb:DimensionProperty.  
eg:obs1 a qb:Observation;  
    sdmx-measure:obsValue "0.17"^^xsd:double;  
    eg:indicator eg:unemployment.  
eg:obs2 a qb:Observation;  
    sdmx-measure:obsValue "0.25"^^xsd:double;  
    eg:indicator eg:poverty.
```

**Q4: An official portal defines a dimension that contains non-associated values (i.e. eg:female, eg:age18-25 eg:large-enterprises). In this case multiple dimensions are embraced together. Are there any advantages in this practice?**

**Q5: The time dimension is very common among the published cubes. In your opinion which is the best practice to define the time dimension?**

1. Re-use sdmx-dimension:refPeriod e.g.:

```
eg:obs1 a qb:Observation;  
    sdmx-dimension:refPeriod eg:2016;  
    eg:unemployment "0.17"^^xsd:double.
```

2. Define a new dimension property. This property is also a subproperty of the sdmx-dimension:refPeriod e.g.:

```
eg:time a qb:DimensionProperty;  
    rdfs:subPropertyOf sdmx-dimension:refPeriod.  
eg:obs1 a qb:Observation;  
    eg:time eg:2016;  
    eg:unemployment "0.17"^^xsd:double.
```

3. Other (please describe)

**Q6: The geographical dimension is also very common among the published cubes. In your opinion which is the best practice to define the geographical dimension?**

1. Re-use sdmx-dimension:refArea e.g.:

```
eg:obs1 a qb:Observation;  
    sdmx-dimension:refArea eg:Greece;  
    eg:unemployment "0.17"^^xsd:double.
```

2. Define a new property. This property is also a subproperty of the sdmx-dimension:refArea e.g.:

```
eg:geo a qb:DimensionProperty;  
    rdfs:subPropertyOf sdmx-dimension:refArea.  
eg:obs1 a qb:Observation;  
    eg:geo eg:Greece;  
    eg:unemployment "0.17"^^xsd:double.
```

3. Other (please describe)

**Q7: The sex dimension is also very common among the published cubes. In your opinion which is the best practice to define the sex dimension?**

1. Re-use sdmx-dimension:sex e.g.:

```
eg:obs1 a qb:Observation;  
    sdmx-dimension:sex sdmx-code:sex-F;  
    eg:unemployment "0.17"^^xsd:double.
```

2. Define a new property. This property is also a subproperty of the sdmx-dimension:sex e.g.:

```
eg:gender a qb:DimensionProperty;  
    rdfs:subPropertyOf sdmx-dimension:sex.  
eg:obs1 a qb:Observation;  
    eg:gender eg:female;  
    eg:unemployment "0.17"^^xsd:double.
```

3. Define a new property. This property is NOT a subproperty of the sdmx-dimension:sex e.g.:

```
eg:gender a qb:DimensionProperty.  
eg:obs1 a qb:Observation;  
    eg:gender eg:female;  
    eg:unemployment "0.17"^^xsd:double.
```

4. Other (please describe)

**Q8: The age is another very common dimension among the published cubes. In your opinion which is the best approach to define the age dimension?**

1. Re-use sdmx-dimension:age e.g.:

```
eg:obs1 a qb:Observation;  
    sdmx-dimension:age eg:18-25;  
    eg:unemployment "0.17"^^xsd:double.
```

2. Define a new property. This property is also a subproperty of the sdmx-dimension:age e.g.:

```
eg:age a qb:DimensionProperty;
    rdfs:subPropertyOf sdmx-dimension:age.
eg:obs1 a qb:Observation;
    eg:age eg:18-25;
    eg:unemployment "0.17"^^xsd:double.
```

3. Define a new property. This property is NOT a subproperty of the sdmx-dimension:age e.g.:

```
eg:age a qb:DimensionProperty.
eg:obs1 a qb:Observation;
    eg:age eg:18-25;
    eg:unemployment "0.17"^^xsd:double.
```

4. Other (please describe)

**Q9:The attribute properties (i.e. qb:AttributeProperty) enable the specification of the units of measure. In your opinion which is the best practice to define the unit of measure?**

1. Re-use sdmx-attribute:unitMeasure e.g.:

```
eg:obs1 a qb:Observation;
    eg:unemployment "0.17"^^xsd:double;
    sdmx-attribute:unitMeasure eg:Percent.
```

2. Define a new attribute. This property is also a subproperty of the sdmx-attribute:unitMeasure e.g.:

```
eg:unit a qb:AttributeProperty;
    rdfs:subPropertyOf sdmx-attribute:unitMeasure.
eg:obs1 a qb:Observation;
    eg:unemployment "0.17"^^xsd:double;
    eg:unit eg:Percent.
```

3. Other (please describe)

**Q10:The QB vocabulary enables the declaration of the unit of measure at different levels. In your opinion which level is the best to declare the unit of measure?**

1. Declare the unit at qb:DataSet level e.g.:

```
eg:dataset1 a qb:DataSet;
    sdmx-attribute:unitMeasure eg:Percent.
eg:unemployment a qb:MeasureProperty.
eg:obs1 a qb:Observation;
    qb:dataSet eg:dataset1;
    eg:unemployment "0.17"^^xsd:double.
```

2. Declare the unit at qb:MeasureProperty level e.g.:

```
eg:dataset1 a qb:DataSet.
```

```

eg:unemployment a qb:MeasureProperty;
    sdmx-attribute:unitMeasure eg:Percent.
eg:obs1 a qb:Observation;
    qb:dataSet eg:dataset1;
    eg:unemployment "0.17"^^xsd:double.

```

3. Declare the unit at qb:Observation levellevel e.g.:

```

eg:dataset1 a qb:DataSet.
eg:unemployment a qb:MeasureProperty.
eg:obs1 a qb:Observation;
    qb:dataSet eg:dataset1;
    eg:unemployment "0.17"^^xsd:double;
    sdmx-attribute:unitMeasure eg:Percent.

```

**Q11: Time dimension values can be represented either as URIs e.g. <http://example.com/2016> or as xsd:date e.g. "2016-01-01"^^xsd:date. Which practice do you consider the best and why?**

**Q12: The values of the time dimension can be drawn from a code list (in the case that URIs are used). In your opinion which code list should be used for the time dimension?**

1. Use reference.data.gov.uk e.g.:

```
http://reference.data.gov.uk/id/year/2015
```

2. Use DBpedia e.g.:

```
http://dbpedia.org/resource/2015
```

3. Define a new code list  
4. Other (please describe)

**Q13: Some datasets contain data for a fixed time period (e.g. census data). In this case should the cube contain a time dimension with a fixed value? e.g. for 2011 census data should the cube contain a time dimension with fixed value "2011"? Which are the advantages/disadvantages of this approach?**

**Q14: The values of the sex dimension can be drawn from a code list. In your opinion which code list should be used for the sex dimension?**

1. Use SDMX code list e.g.:

```
sdmx-code:sex-F, sdmx-code:sex-M
```

2. Define a new code list  
3. Other (please describe)

**Q15: The values of the measure units can be drawn from a code list. In your opinion which code list should be used for the measure units? Should different code lists be used for diverse units?**

1. Use QUDT (<http://qudt.org/>) e.g.:

<http://qudt.org/vocab/unit#Euro>

2. Use DBpedia e.g.:

<http://dbpedia.org/resource/Euro>

3. Define a new code lists  
4. Other (please describe)

**Q16: The QB vocabulary allows two different practices for defining the values of a dimension. In your opinion which practice should be followed?**

1. Use the property `qb:codeList` to associate a `skos:ConceptScheme` to the `qb:DimensionProperty`

`sdmx-code:sex a skos:ConceptScheme.`  
`eg:sex a qb:DimensionProperty, qb:CodedProperty;`  
`qb:codeList sdmx-code:sex.`

2. Use the property `rdfs:range` to associate a `skos:Concept` to the `qb:DimensionProperty`

`sdmx-code:Sex a rdfs:Class.`  
`eg:sex a qb:DimensionProperty, qb:CodedProperty;`  
`rdfs:range sdmx-code:Sex.`

3. Use both

`sdmx-code:sex a skos:ConceptScheme.`  
`sdmx-code:Sex a rdfs:Class.`  
`eg:sex a qb:DimensionProperty, qb:CodedProperty;`  
`qb:codeList sdmx-code:sex;`  
`rdfs:range sdmx-code:Sex.`

**Q17: In some cases aggregated values (e.g:total) are used at the dimensions. For example the sex dimension may have values `sdmx-code:sex-F`, `sdmx-code:sex-M` and `eg:total`. Which are the advantages/disadvantages of this approach? Should the relationship between `eg:total` and `sdmx-code:sex-F`, `sdmx-code:sex-M` be defined explicitly?**

**Q18: The dimension values often have hierarchical relations i.e. generalization and specialization e.g. Greece is part of Europe. Which approach to follow to express such relations?**

1. Use SKOS e.g.:

`eg:Greece skos:broader eg:Europe`

2. Use XKOS e.g.:

eg:Greece xkos:isPartOf eg:Europe

3. Define new properties e.g.:

eg:Greece eg:within eg:Europe

4. Other (please describe)

**Q19: The dimension values are often organized into hierarchical levels e.g. region, district. Which is the best practice to define the hierarchical levels?**

1. Use the rdf:type e.g.:

eg:Country rdf:type rdfs:Class.  
eg:Greece rdf:type eg:Country.

2. Define new properties e.g.

eg:Country rdf:type rdfs:Class.  
eg:Greece eg:hasLevel eg:Country.

3. Use SKOS/XKOS e.g.:

eg:Country rdf:type xkos:ClassificationLevel.  
eg:Greece skos:member eg:Country.

4. Other (please describe)

**Q20: An official portal defines separate cubes for each hierarchical level e.g. define a cube that contains the regions and another that contains the districts. Which are the advantages of this approach?**

**Q21: Please indicate any other issues regarding the publishing practices of Linked Data Cubes.**

## Appendix C: Questionnaire for developers

### General Questions

- How important are the open statistical data that are produced by Governments and organizations, according to your opinion?

Not at all                      1            2            3            4            5                      Very much

☐                      ☐            ☐            ☐            ☐

- What percentage of the data that provides national portals of open data is statistical according to you?

0%                      <20%                      <50%                      <75%                      100%

☐                      ☐                      ☐                      ☐                      ☐

- Which of the following open statistical data sources are you aware of?

☐ World Bank    ☐ Eurostat

☐ Global Health Observatory (GHO data)                      ☐ European Data Portal

☐ OECD    ☐ Google Finance

### Open Data

- Have you ever used open data for the development of an application?

Yes            No

☐            ☐

### General questions about your Open Data application

- What was your application about?

☐ Economics    ☐ Security

☐ Agricultural    ☐ Entertainment

☐ Education    ☐ Healthcare

☐ Legislation

- What was the theme of the data you used for the application?

☐ Economics    ☐ Security

☐ Agricultural    ☐ Entertainment

☐ Education    ☐ Healthcare

☐ Legislation

- During the development of your application, how many portals did you use in total, to collect your data?

Not at all                      1            2            3            4            5+                      Very much

☐                      ☐            ☐            ☐            ☐



- How many datasets did you collect in total?
- How many of them were open statistical data?

### Open Data Questions

- How satisfied are you with the quality of the open data that you used in your application?

	Not at all					Very Much
	1	2	3	4	5	
Occupancy	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Validity	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Comprehensiveness	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Accuracy	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Understandable	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Up-to-dated	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Format	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

- How difficult was it for you to find the open data that you used for your application?

Not at all	1	2	3	4	5	Very much
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

- Which was the greatest difficulty you encountered during finding the necessary open data that you needed for your application?

- How difficult was it for you to combine the open data that you used for your application?

Not at all	1	2	3	4	5	Very much
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

- Which was the greatest difficulty that you encountered during the combination of the open data that you used for your application?

### Open Statistical Data

- Have you used Open and Statistical Data for the development of your application?

Yes	No
<input type="checkbox"/>	<input type="checkbox"/>

### Open Statistical Data Questions

- How satisfied are you with the quality of the open statistical data that you used in your application?

	Not at all					Very Much
	1	2	3	4	5	
Occupancy	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
Validity	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

Comprehensiveness	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Accuracy	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Understandable	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Up-to-dated	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Format	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

- How difficult was it for you to find the open statistical data that you used for your application?

Not at all	1	2	3	4	5	Very much
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

- Which was the greatest difficulty you encountered during finding the necessary open and statistical data that you needed for your application?

- How difficult was it for you to combine the open statistical data that you used for your application?

Not at all	1	2	3	4	5	Very much
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

- Which was the greatest difficulty that you encountered during the combination of the open statistical data that you used for your application?

### General Questions

- Which programming language used to develop your application?

<input type="checkbox"/> Python	<input type="checkbox"/> C++
<input type="checkbox"/> Java	<input type="checkbox"/> HTML5
<input type="checkbox"/> PHP	<input type="checkbox"/> JavaScript
<input type="checkbox"/> Ruby	<input type="checkbox"/> Other

- Where did you store the open data for usage by the application?

<input type="checkbox"/> Relational database	<input type="checkbox"/> RDF Store
<input type="checkbox"/> NOSQL database	<input type="checkbox"/> other

- Did you make any sort of statistical analysis in these data?

<input type="checkbox"/> Correspondence analysis	<input type="checkbox"/> Descriptive Statistical Analysis
<input type="checkbox"/> Linear Regression	<input type="checkbox"/> Timeline
<input type="checkbox"/> Logistic Regression	<input type="checkbox"/> Other

- Was there any visualization of the data? If yes, what kind of visualization did you make?

<input type="checkbox"/> Visualization on a graph	<input type="checkbox"/> There was no visualization
<input type="checkbox"/> Visualization on a map	

**Personal Characteristics**

- Country
- How many individuals did you collaborate as a team in order to conclude your application?

1	2	3	4	5	6	7+
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
- Thinking about yourself, on a scale of 1 to 3, to what extent do you associate with the following labels?

	I do not associate with this label at all	I slightly associate with this label	I associate myself with this label
Web Developer	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Data analyst	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Business Analyst	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Policy Maker	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
IT Specialist	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Entrepreneur	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Designer	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Open Data advocate	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
- Employment Status

<input type="checkbox"/> Student	<input type="checkbox"/> Employed
<input type="checkbox"/> Voluntary Work	<input type="checkbox"/> Unemployed
<input type="checkbox"/> Self Employed	<input type="checkbox"/> Retired
<input type="checkbox"/> Business owner	<input type="checkbox"/> Other

## Appendix D: Interview form for pilot partners

### Instructions

- This interview form was designed for the European project "OpenGovIntelligence" which aims at fostering innovation in society and enterprises through the modernization of Public Administration by exploiting Linked Open Statistical Data technologies.
- The form will support the interviews that will be conducted by CERTH with the pilot partners of the project in the frame of task T1.3.
- The objective of the interviews is to identify challenges and needs of users that aim to exploit statistical open data.
- Pilot partners of the project will describe problems of public administration, businesses and citizens that can be solved through the exploitation of statistical data. National or cross-country problems can be mentioned.
- The present interview form will not specify the final version of the pilot scenarios.

### Problems

- Please describe an existing problem that can be solved through the exploitation of statistical data (in 200 words)
- Please select the category that your problem falls into

	National	Cross Country
Public Administration	<input type="checkbox"/>	<input type="checkbox"/>
Business	<input type="checkbox"/>	<input type="checkbox"/>
Citizens	<input type="checkbox"/>	<input type="checkbox"/>

### Datasets

- Please list all the datasets that are necessary to solve the problem
- For each dataset please specify: (a) the measured variable and the dimensions (e.g. count of unemployed people per month in municipalities of Greece) (b) Open/Closed data (if data are open which open data portal provide them and if data are closed who is the owner and what are the access constraints)

### Final Product/Service

- Please describe the target users of the final product/service
- Please describe the functionalities the final product/service will support

Please describe methods and tools (statistical analysis, visualization etc.) that are needed to develop the final product/service