opengovintelligence

OpenGovIntelligence

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

Deliverable 6.1

Data Management Plan

Editor(s):	Evangelos Kalampokis (CERTH)
Responsible Organisation:	CERTH
Version-Status:	V1.0 (1st Update)
Submission date:	31/01/2017
Dissemination level:	CO

Deliverable factsheet

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

Title of Deliverable:	D6.1 – Data Management Plan
Work package:	WP6 – Project Management
Due date according to contract:	31/01/2017 (update)

Editor(s):	Evangelos Kalampokis (CERTH)
Contributor(s):	Jamie Whyte (TRAF)
	Trevor Alcorn (MI)
	Ramūnas Čižas (EL)
	Eleni Chaniotaki (MAREG)
	Paul Hermans (ProXML)
	Marijn Janssen (TUDelft)
	Robert Krimmer (TUT)
	Ricardo Matheus (TUDelft)
	Maarja Toots (TUT)
	Efthimios Tambouris (CERTH)
	Konstantinos Tarabanis (CERTH)
Reviewer(s):	-
Approved by:	All Partners

Abstract:	This document provides the plan for managing the data generated and collected during the project. It covers: a) the handling of research data during and after the project, b) what data will be collected, processed or generated, c) what methodology and standards will be applied, d) whether data will be shared/made open and how and e) how data will be curated and preserved.
	The plan currently involves twenty five data sets from nine different CSA domains including *Environment*, *Transport*, *Population and housing censuses*, *Government finance, fiscal and public sector statistics*, *Science, technology and innovation* and others. Most data sets will be openly provided to the public and half of the data sets will be published as linked data using the RDF Data Cube vocabulary. DCAT and PROV vocabularies will be used to model metadata. Data sets will be preserved after the end of the project on the pilot's web sites, on web servers or other web-based solutions.
Keyword List:	Project, management, data, plan

Consortium

	Role	Name	Short Name	Country
1.	Coordinator	Centre for Research & Technology – Hellas	CERTH	Greece
2.	R&D partner	Delft University of Technology	TUDelft	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

Version	Date	Revised by	Reason
0.1	15/03/2016	CERTH	TOC
0.2	30/06/2016	CERTH	TOC updated
0.3	20/06/2016	ALL	Data management policy, input from pilots regarding data sets inserted
0.4	26/07/2016	CERTH	Introduction updated, abstract inserted
1.0	27/07/2016	CERTH	Final version
1.0 (1st Update)	25/01/2017	CERTH	Document updated to include new data set descriptions from the Trafford council (TRAF)

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

List of Tables

List of Abbreviations

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

Abbreviation	Description
DMP	Data Management Plan
DCAT	Data Catalogue Vocabulary
RDF	Resource Description Standard
WP	Work Package
PA	Public Administration
LOSD	Linked Open Statistical Data
ICT	Information and Communications Technology
EU	European Commission
CSA	Classification of Statistical Activities

Executive Summary

This is deliverable "D6.1 Data Management Plan" of the OpenGovIntelligence Project. The OpenGovIntelligence project aims at stimulating sustainable economic growth in Europe through fostering innovation in society and enterprises. Towards this end, OpenGovIntelligence suggests a holistic approach for the modernization of Public Administration (PA) by exploiting Linked Open Statistical Data (LOSD) technologies. This includes new business processes, policies, and tools that will enable the active participation of the society and enterprises in data sharing and in the co-production of innovative data-driven public services.

This deliverable documents activities that are performed in WP6 "Project Management". WP6 intends to:

- Perform strategic and day-to-day administrative, financial, scientific and technical management of the project.
- Ensure the sound management of project activities and the fulfilment of project objectives.
- Monitor resource usage, budget allocation and project cash flow.
- Ensure efficient communication within the consortium and reassure effective liaison with the EC, other projects, communities and other bodies as required.
- Deal with all ethical issues that might emerge during the project.

The purpose of this document is to provide the plan for managing the data generated and collected during the project; The Data Management Plan. Specifically, the DMP describes the data management life cycle for all datasets to be collected, processed and/or generated by a research project. It covers:

- the handling of research data during and after the project
- what data will be collected, processed or generated.
- what methodology and standards will be applied.
- whether data will be shared/made open and how
- how data will be curated and preserved

Following the EU's guidelines regarding the DMP, this document may be updated - if appropriate - during the project lifetime (in the form of deliverables).

The DMP currently involves twenty five data sets from nine different CSA domains namely *Environment*, *Transport*, *Population and housing censuses*, *Government finance, fiscal and public sector statistics*, *Science, technology and innovation*, *Human settlements and housing*, *Regional and small area statistics*, *Justice and crime*, and *Health*. All data sets except one will be openly provided to the public on Web servers, SPARQL endpoints or other, web-based solutions. Half of the data sets will be published as linked data using the RDF Data Cube vocabulary while the rest of them will be published as CSV, XLS and other formats. Moreover, for the linked data cubes the DCAT and PROV vocabularies will be used to model their metadata. Finally, data sets will be preserved after the end of the project on the pilot's web sites, on web servers or other web-based solutions.

1 Introduction

This document is the OpenGovIntelligence Data Management Plan (DMP). The consortium is required to create the DMP because the OpenGovIntelligence project participates in the Open Research Data pilot. The DMP describes the data management life cycle for all datasets to be collected, processed and/or generated by a research project. It covers:

- the handling of research data during and after the project
- what data will be collected, processed or generated.
- what methodology and standards will be applied.
- whether data will be shared/made open and how
- how data will be curated and preserved

1.1 Scope

The present document is the Deliverable 6.1 "D6.1 – Data Management Plant" (henceforth referred to as D6.1) of the OpenGovIntelligence project. The main objective of D6.1 is to to provide the plan for managing the data generated and collected during the project.

According to the EU's guidelines regarding the DMP (European Commission, 2016), the document may be updated - if appropriate - during the project lifetime (in the form of deliverables).

1.2 Audience

The intended audience for this document is the OpenGovIntelligence consortium and the European Commission.

1.3 Structure

The structure of the document is as follows:

- Section 2 presents the methodology that the consortium used to create the DMP.
- Section 3 presents the policy that the consortium of the OpenGovIntelligence project plans to use regarding the data sets collected, processed and/or generated within the project.
- Section 4 presents the analysis of the data sets included in the DMP.
- Section 5 draws conclusion and sets future goals.

2 Methodology

The Data Management Plan regards all the data sets that will be collected, processed and/or generated within the project. The methodology the consortium follows to create the DMP is as follows:

1. Create a data management policy. To this end, we describe a) the elements that the EU proposes to address for each data set and b) the strategy that is used by the consortium to address each of element. The elements were used to create a DMP template which was sent to the partners of the consortium in order to filled it in with information for each relative data set.
2. Analyse the completed by the project's partners DMP templates.

3 Data management policy

Being in line with the EU's guidelines regarding the DMP (European Commission, 2016), this document should address for each data set collected, processed and/or generated in the project the following elements:

1. Data set reference and name
2. Data set description
3. Standards and metadata
4. Data sharing
5. Archiving and preservation

To this end, the consortium develops a number of strategies that will be followed in order to address the above elements.

In this section, we provide a detailed description of these elements in order to ensure their understanding by the partners of the consortium. For each element, we also describe the strategy that will be used to address it.

3.1 Data set reference and name

In order to be able to distinguish and easily identify data sets, each data set is assigned with a unique name. This name can also be used as the identifier of the data sets.

In order to design the data set names, we use the following practice:

1. Each data set name consists of *four* different parts separated with a dot: *CSACode.CountryCode.PartnerName.DatasetName,* where
    a. The *CSACode* part describes the theme that the dataset fits. We use the Classification of Statistical Activities (CSA)[1] in order to represent different themes. Specifically, CSA uses five domains (namely (a) Demographic and social statistics, Economic statistics, (b) Economic statistics, (c) Environment and multi-domain statistics, (d) Methodology for data collection, processing dissemination and analysis, and (e) Strategic managerial issues of official statistics) to classify the statistical activities undertaken by national and international statistical organizations. Each domain is then further divided into a number of subdomains. To represent the theme in the name of the data sets, we use the number that corresponds to each sub-domain of CSA. For example, a data set that comes from the education domain will have *1.3* as theme which corresponds to "Education" in CSA.
    b. The *CountryCode* part represents the country associated with the dataset using ISO Alpha-2 country codes:

---

[1]

http://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=DSP_GEN_DESC_VIEW_NOHDR&StrNom=CSA&StrLanguageCode=EN

    i. BE for Belgium
    ii. EE for Estonia
    iii. GR for Greece
    iv. IE for Ireland
    v. LT for Lithuania
    vi. NL for the Netherlands
    vii. UK for the United Kingdom

  c. The *PartnerName* part represents the name of the organisation (e.g. data owner, data custodian) associated with the dataset:
    i. MarineInstitute for the Marine Institue
    ii. CERTH for the Center for Research and Technology, Hellas
    iii. TUDelft for the Technische Universiteit Delft
    iv. NUIG for the National University of Ireland, Galway
    v. TUT for Tallinn University of Technology
    vi. SWIRRL for Swirrl IT Limited
    vii. TRAF for Trafford council
    viii. VLO for the Flemish Government
    ix. MAREG for the Ministry of Interior and Administrative Reconstruction
    x. MKM for the Ministry of Economic Affairs and Communication
    xi. EnterpriseLithuania for the Public Institution Enterprise Lithuania

  d. The *DatasetName* represens the full name of the dataset.

An example of a data set's name could be the following:

<p align="center">2.4.4.GR.MAREG.GovernmentVehicles</p>

The above name indicates that the data set describes Government Vehicles from the Ministry of Interior and Administrative Reconstruction in Greece and that the data set regards the 2.4.4 (Economic Statistics→Sectoral Statistics→Transport) domain.

In addition, only the data sets from the Marine Institute will have a second name. This is because (a) these data sets are very environmental specific and may not be fully described by CSA and (b) the Marine Institute has already assigned to them a name in a previous effort so we decided to keep those names too. The practice used to design these data sets' names is the following:

1. Each data set name consists of *five* different parts separated with a dot *INSPIRECode.INSPIREDataSpecification.CountryCode.PartnerName.DatasetName,* where
  a. *INSPIRECode* consist of two letters for the appropriate INSPIRE theme the dataset fits.
  b. The *INSPIREDataSpecification* is the name of the dataset to be portrayed for INSPIRE European data harmonisation and this comes from INSPIRE Data Specifications.
  c. The *CountryCode* refers to the country associated with the dataset using ISO Alpha-2 country codes.
  d. The *PartnerName* is the full name of the organisation (e.g. data owner, data custodian) associated with the dataset.

e. The *DatasetName* is the full name of the dataset.

An example of such a data set's name is the following:

OF.PointTimeSeriesObservation.IE.MarineInstitute.IrishWeatherBuoyNetwork.Wave

The above name means that the 'Irish Weather Buoy Network' 'Wave' dataset from the Marine Institute in Ireland is a PointTimeSeriesObservation in the inspire theme 'Oceanographic geographical features'.

3.2 Data set description

Each data set that will be collected, processed or generated within the project will be accompanied by a brief description. The description will provide information regarding:

- The nature of the data set
- The scale of the data set
- To whom could the data set be useful
- Whether the data set underpins a scientific publication
- Information on the existence (or not) of similar data sets.
- Possibilities for integration with other data sets and reuse

In case the data set is collected, the origin of the data set will be also provided.

3.3 Standards and metadata

This field will describe suitable standards that will be used to describe the data as well as the metadata of the data sets. In OpenGovIntelligence, we identify two types of data sets: (a) linked data sets and (b) other datasets.

The first category includes data sets that will be transformed to linked data within the project's lifetime. To this end two types of standards will be mainly used: (a) the Resource Description Standard (RDF) and (b) the RDF data cube vocabulary. In some cases additional vocabularies will be used such as the Registered Organization Vocabulary[2] (regorg) for describing organizations and the ISA Programme Location Core Vocabulary[3] (lcon) for the description of names.

RDF is a W3C standard model since 1999 for interchanging data on the Web. RDF models data as triples (subject-predicate-object). At the same time, the RDF data cube vocabulary is also a W3C standard since 2015 that is used to model statistical data as linked data cubes using RDF. Specifically, the RDF data cube vocabulary allows the modelling of statistical data as cubes using dimensions, attributes and measures. A measure represents the phenomenon that is being observed in the data set. A dimension defines what an observation in the data set applies to (e.g. time, area). Attributes are used to specify units of measure, scaling factors and other.

---

[2] https://www.w3.org/TR/vocab-regorg/
[3] https://www.w3.org/ns/locn

As we believe that metadata are crucial for the long-term use of the data sets, we decided to accompany (whenever possible) linked data set with a set of metadata. Specifically, we will use existing vocabularies to describe metadata such as the Data Catalogue Vocabulary (DCAT)[4] and the PROV[5] ontology.

DCAT is a W3C vocabulary that facilitates the interoperability between data catalogs published on the Web. Properties of DCAT that can be used to describe open data sets in the project include:

- *dct:title* to give a title to the dataset
- *dct:descripton* to provide a brief description for the data set
- *dct:issued* to provide the date of the publication of the data set
- *dct:language* for the language of the data set
- *dct:publisher* for the entity that publishes the data set
- *dcat:keyword* to provide a keyword that describes the data set

At the same time, the PROV ontology is also a W3C standard that can be used to provide provenance descriptions to the data sets. In this context, the PROV ontology can be used to model the activities, entities and actors involved in the data set production process. Specifically, according to the PROV ontology:

- *prov:Entity* can be used to describe a physical, digital, conceptual, or other kind of thing. A prov:Entity
- *prov:Activity* represents an activity that occurs over a period of time and acts upon or with entities; it may include consuming, processing, transforming, modifying, relocating, using, or generating entities
- *prov:Agent* is something that bears some form of responsibility for an activity taking place, for the existence of an entity, or for another agent's activity

Finally, in the second category comply all the data sets published in other formats (e.g. excel, csv, pdf, txt etc.). These data sets will also by accompanied, is some cases, by metadata.

3.4 Data Sharing

The partner responsible for drafting the DMP along with all work package leaders, will define how data will be shared and more specifically the access procedures, the embargo periods, the necessary software and other tools for enabling re-use, for all datasets that will be collected, generated, or processed in the project. In case the dataset cannot be shared, the reasons for this will be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).

---

[4] https://www.w3.org/TR/vocab-dcat
[5] https://www.w3.org/TR/prov-o/

3.5   Archiving and preservation

The partners will decide and describe the procedures that will be used in order to ensure long-term preservation of the data sets. This field will provide information regarding the duration of the data preservation, the approximate end volume, the associated costs and the plans of the consortium to cover the costs.

4 Data sets analysis

The DMP currently includes twenty five data sets proposed by the partners of the consortium. This section presents the results of the analysis of these data sets.

4.1 Data set reference and name

Table 1 presents the data sets recorded by the partners of the consortium along with the CSA domain they belong to. Data sets come from thirteen different domains. Specifically, seven data sets come from the *Environment* domain, three from the *Transport* domain, two from the *Population and housing censuses* domain, two from the *Regional and small area statistics* domain, two from the *Government finance, fiscal and public sector statistics* domain, one from the *Science, technology and innovation* domain, one from the *Human settlements and housing* domain, two from the *Justice and crime* domain, one from the *Business statistics* domain, two from the *Labour* domain, one from the *Population and migration* domain, one from the *Social protection* domain, and one from the *Health* domain.

Table 1. Data sets and CSA domains

Dataset	CSA Domain
3.1.IE.MarineInstitute.IrishWeatherBuoyNetwork.Wave	Environment
3.1.IE.MarineInstitute.WaveBuoyNetwork	Environment
3.1.IE.MarineInstitute.EastAtlanticSWANWaveModel	Environment
3.1.IE.MarineInstitute.IrishNationalTideGaugeNetwork	Environment
2.9.EE.TUT.OGIServiceCoproductionNeeds	Science, technology and innovation
3.1.EE.TUT.EstonianLandCadastre	Regional and small area statistics
1.7.EE.TUT.LandRegister	Human settlements and housing
4.3.1.EE.TUT.EstonianBuildingRegister	Population and housing censuses; registers of population, dwellings and buildings
4.3.1.EE.TUT.KVee	Population and housing censuses; registers of population, dwellings and buildings
1.8.EE.TUT.CrimeStatisticalData	Justice and crime
2.5.GR.MAREG.GreekPublicAgenciesPersonnel	Government finance, fiscal and public sector statistics
2.4.4.GR.MAREG.GovernmentVehiclesOperationMaintenace	Transport

3.2.GR.MAREG.GreekMunicipalitiesPrefecturesRegions	Regional and small area statistics
2.4.4.GR.MAREG.GovernmentVehiclesLifecycle	Transport
2.4.4.GR.MAREG.GovernmentVehicles	Transport
3.1.BE.VLO.AirWaterPollutionMetadata	Environment
3.1.BE.VLO.AirWaterPollutionData	Environment
2.5. BE.VLO.CBB	Environment
1.4.LT.EnterpriseLithuania.PermitsHygienePassports	Health
2.5.LT.EnterpiseLithuania.AdvertisementPermitsData	Business statistics
1.2.UK.TRAF.NEET	Labour
1.6.UK.TRAF.Benefits	Social protection
1.8.UK.TRAF.Crime	Justice and crime
1.1.UK.TRAF.Census	Population and migration
1.2.UK.TRAF.JCPlus	Labour

4.2 Standards and metadata

Table 2 presents information regarding the datasets, the standards that will be used to model them (as linked data or not) and also the standards that will be used to model their metadata (if any). Most data sets will be published as linked data cubes using the RDF data cube vocabulary and, in one case, the Dublin Core vocabulary. The rest of the data sets will be only used as CSV, XLS, PDF and similar formats. In addition, from the data sets that will be published as linked data cubes, five will be accompanied by PROV metadata, five with Dublin Core metadata, while the rest of them will be accompanied by DCAT and/or PROV metadata.

Table 2. Datasets, standards and metadata

Dataset	Data standard	Linked Data standard	Metadata standard
3.1.IE.MarineInstitute.IrishWeatherBuoyNetwork.Wave	CSV, JSON	-	ISO 19139
3.1.IE.MarineInstitute.WaveBuoyNetwork	CSV, JSON	-	ISO 19139
3.1.IE.MarineInstitute.EastAtlanticSWANWaveModel	CSV, esriAscii, JSON, GeoTiff, PNG, PDF, the ocean data format NetCDF	-	-

3.1.IE.MarineInstitute.IrishNationalTideGauge Network	CSV, JSON	-	ISO 19139
2.9.EE.TUT.OGIServiceCoproductionNeeds	XLS, CSV	-	-
3.1.EE.TUT.EstonianLandCadastre	CSV	-	-
1.7.EE.TUT.LandRegister	N/A	-	-
4.3.1.EE.TUT.EstonianBuildingRegister	N/A	-	-
4.3.1.EE.TUT.KVee	CSV	-	-
1.8.EE.TUT.CrimeStatisticalData	CSV	-	-
2.5.GR.MAREG.GreekPublicAgenciesPersonnel	CSV	RDF, RDF data cube	DCAT, PROV
2.4.4.GR.MAREG.GovernmentVehiclesOperationMaintenace	CSV	RDF, RDF data cube	DCAT, PROV
3.2.GR.MAREG.GreekMunicipalitiesPrefecturesRegions	CSV	RDF, RDF data cube	DCAT, PROV
2.4.4.GR.MAREG.GovernmentVehiclesLifecycle	CSV	RDF, RDF data cube	DCAT, PROV
2.4.4.GR.MAREG.GovernmentVehicles	CSV	RDF, RDF data cube	DCAT, PROV
3.1.BE.VLO.AirWaterPollutionMetadata		Dublin Core	PROV
3.1.BE.VLO.AirWaterPollutionData	XML	RDF, RDF data cube	PROV
2.5. BE.VLO.CBB	-	org, regorg & locn	PROV
1.4.LT.EnterpriseLithuania.PermitsHygienePassports	CSV	RDF, RDF data cube	PROV
2.5.LT.EnterpiseLithuania.AdvertisementPermitsData	-	RDF, RDF data cube	PROV
1.2.UK.TRAF.NEET	N/A	N/A	Dublin Core
1.6.UK.TRAF.Benefits	XLS, CSV, TSV	RDF, RDF data cube	Dublin Core
1.8.UK.TRAF.Crime	CSV	RDF, RDF data cube	Dublin Core
1.1.UK.TRAF.Census	XLS	RDF, RDF	Dublin Core

		data cube	
1.2.UK.TRAF.JCPlus	N/A	World Geodetic System 1984 (WGS84), RDF, RDF data cube	Dublin Core

4.3 Data Sharing

Table 3 describes the access policy and sharing medium for each data set. All data sets except one will be openly offered to the public. Moreover, all data sets apart from the one with the restricted access will be published in a Web server, SPARQL endpoint or through a web –based solution.

Table 3. Datasets, access policy and sharing medium

Dataset	Open / Restricted	Sharing medium
3.1.IE.MarineInstitute.IrishWeatherBuoyNetwork.Wave	Open.	Web server
3.1.IE.MarineInstitute.WaveBuoyNetwork	Open	Web server
3.1.IE.MarineInstitute.EastAtlanticSWANWaveModel	Open	Web server
3.1.IE.MarineInstitute.IrishNationalTideGaugeNetwork	Open	Web server
2.9.EE.TUT.OGIServiceCoproductionNeeds	Restricted	-
3.1.EE.TUT.EstonianLandCadastre	Open	Web-based solution
1.7.EE.TUT.LandRegister	Open	Web-based solution
4.3.1.EE.TUT.EstonianBuildingRegister	Open	Web-based solution
4.3.1.EE.TUT.KVee	Open	Web-based solution
1.8.EE.TUT.CrimeStatisticalData	Open	Web-based solution
2.5.GR.MAREG.GreekPublicAgenciesPersonnel	Open	Web server
2.4.4.GR.MAREG.GovernmentVehiclesOperationMaintenace	Open	Web server
3.2.GR.MAREG.GreekMunicipalitiesPrefecturesRegions	Open	Web server

2.4.4.GR.MAREG.GovernmentVehiclesLifecycle	Open	Web server
2.4.4.GR.MAREG.GovernmentVehicles	Open	Web server
3.1.BE.VLO.AirWaterPollutionMetadata	Open	SPARQL endpoint
3.1.BE.VLO.AirWaterPollutionData	Open	SPARQL endpoint
2.5. BE.VLO.CBB	Open	SPARQL endpoint
1.4.LT.EnterpriseLithuania.PermitsHygienePassports	Open	SPARQL endpoint
2.5.LT.EnterpiseLithuania.AdvertisementPermitsData	Open	SPARQL endpoint
1.2.UK.TRAF.NEET	Open	SPARQL endpoint
1.6.UK.TRAF.Benefits	Open	SPARQL endpoint
1.8.UK.TRAF.Crime	Open	SPARQL endpoint
1.1.UK.TRAF.Census	Open	SPARQL endpoint
1.2.UK.TRAF.JCPlus	Open	SPARQL endpoint

4.4   Archiving and preservation

Long-term preservation of data sets will be ensured by archiving them for a number of years after the end of the project. Table 4 summarizes the data sets, their planned preservation duration, the medium that they will be archived to and the planned costs.

Table 4. Datasets, preservation and costs

Dataset	Preservation Duration	Preservation Medium	Costs
3.1.IE.MarineInstitute.IrishWeatherBuoyNetwork.Wave	N/A	ERDDAP data server	None
3.1.IE.MarineInstitute.WaveBuoyNetwork	N/A	ERDDAP data server	None
3.1.IE.MarineInstitute.EastAtlanticSWANWaveModel	N/A	ERDDAP data server	None
3.1.IE.MarineInstitute.IrishNationalTideGaugeNetwork	N/A	ERDDAP data server	None
2.9.EE.TUT.OGIServiceCoproductionNeeds	5 years	Tallinn University's data server	None
3.1.EE.TUT.EstonianLandCadastre	N/A	N/A	None

1.7.EE.TUT.LandRegister	N/A	N/A	None
4.3.1.EE.TUT.EstonianBuildingRegister	N/A	N/A	None
4.3.1.EE.TUT.KVee	N/A	N/A	None
1.8.EE.TUT.CrimeStatisticalData	N/A	N/A	None
2.5.GR.MAREG.GreekPublicAgenciesPersonnel	5 years	MAREG's website	None
2.4.4.GR.MAREG.GovernmentVehiclesOperationMaintenace	5 years	MAREG's website	None
3.2.GR.MAREG.GreekMunicipalitiesPrefecturesRegions	5 years	MAREG's website	None
2.4.4.GR.MAREG.GovernmentVehiclesLifecycle	5 years	MAREG's website	None
2.4.4.GR.MAREG.GovernmentVehicles	5 years	MAREG's website	None
3.1.BE.VLO.AirWaterPollutionMetadata	N/A	Archival system	None
3.1.BE.VLO.AirWaterPollutionData	N/A	Archival system	None
2.5. BE.VLO.CBB	N/A	PostgreSQL	None
1.4.LT.EnterpriseLithuania.PermitsHygienePassports	Long lasting	Github	None
2.5.LT.EnterpiseLithuania.AdvertisementPermitsData	Long lasting	Website of Lithuania's national public health centre	None
1.2.UK.TRAF.NEET	5 years	Connexions management system	None
1.6.UK.TRAF.Benefits	5 years	IS of UK's Department of Work and Pensions (DWP)	None
1.8.UK.TRAF.Crime	5 years	N/A	None

1.1.UK.TRAF.Census	5 years	N/A	None
1.2.UK.TRAF.JCPlus	5 years	N/A	None

5   Conclusion

The purpose of this document was to provide the plan for managing the data generated and collected during the project; The Data Management Plan. Specifically, the DMP described the data management life cycle for all datasets to be collected, processed and/or generated by a research project. It covered:

- the handling of research data during and after the project
- what data will be collected, processed or generated.
- what methodology and standards will be applied.
- whether data will be shared/made open and how
- how data will be curated and preserved

Following the EU's guidelines regarding the DMP, this document may be updated - if appropriate - during the project lifetime (in the form of deliverables).

The DMP currently involves twenty five data sets from thirteen different CSA domains namely *Environment*, *Transport*, *Population and housing censuses*, *Government finance, fiscal and public sector statistics*, *Science, technology and innovation*, *Human settlements and housing*, *Regional and small area statistics*, *Justice and crime*, *Labour*, *Population and migration*, *Social protection*, and *Health*. All data sets except one will be openly provided to the public on Web servers, SPARQL endpoints or other, web-based solutions. Most of the data sets will be published as linked data using RDF and the RDF Data Cube vocabulary while the rest of them will be published as CSV, XLS and other formats. Moreover, for the linked data cubes the DCAT, Dublin rore and PROV vocabularies will be used to model their metadata. Finally, data sets will be preserved after the end of the project on the pilot's web sites, on web servers or other web-based solutions.

References

European Commission (2016), Guidelines on Data Management in Horizon 2020. Available at http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf (accessed 22 July 2016)

Appendix 1: DMP Template

Data set reference and name	*Identifier for the data set to be produced*
Data set description	*Description of the data that will be generated or collected, its origin (in case it is collected), nature and scale and to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse.* *e.g. Data will be collected through a survey with key actors ... The questionnaires will be disseminated to at least 20 recipient. Data will be useful for public authorities & policy makers.*
Standards and metadata	*Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created.*
Data Sharing	*Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.).* *In case a dataset cannot be shared, please mention the reasons for this (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).* *e.g. Data will be published in the website of the project. There will be no permission restriction placed on the data.*
Archiving and preservation (including storage and backup)	*Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered.* *e.g. To ensure data safety we will store and regularly back up them in existing file server. We do not need additional resources for the preservation of data. Data will be archived in servers for 3 years after the end of the project.*

Appendix 2: Datasets

Marine Institute datasets

Data set reference and name	1. **3.1.IE.MarineInstitute.IrishWeatherBuoyNetwork.Wave** 2. **OF.PointTimeSeriesObservation.IE.MarineInstitute.IrishWeatherBuoyNetwork.Wave**
Data set description	Real time oceanographic data on the wave climate collected from the Irish moored Weather Buoy network of stations in the North Atlantic Ocean. The Irish Weather Buoy Network moored surface buoys collect data on wave behaviour including wave height (in metres), wave period (in seconds), mean wave direction (degreeTrue) and maximum wave height (known as Hmax) on an hourly basis and this is disseminated via the Marine Institute ERRDAP data server. Information is collected on a frequency of an hourly basis on a continuous data collection basis when sensors are working correctly and in deployment. This weather buoy point time-series observation data is very useful to Engineer Ed when considering developing a wave energy converter and his requirement for access to real-time data on waves in order to predict the performance of his device and to monitor its safety once it is deployed in the environment.
Standards and metadata	Point time-series observation data on waves is presently available in a number of common open formats including CSV (comma-separated ASCII text table) and JSON (Java Script Object Notation). ISO 19139 metadata describing the observation data is hosted on the Irish Spatial Data Exchange and Irelands Open Data Catalogue data.gov.ie. http://www.isde.ie/#/464864ce-319b-4645-b4c9-f408d656fd76
Data Sharing	The Marine Institute presently publish point time-series wave data via an ERDDAP data server under a Creative Common CC-BY 4.0 open licence. Users who require this data simply access the http link provided and choose the variables of interest (e.g. wave) to fit their requirements. http://erddap.marine.ie/erddap/tabledap/IWBNetwork.html
Archiving and preservation (including storage and backup)	The Marine Institute ERDDAP server is hosted on a backend SQL Server database with standard SQL Server backup devices used to preserve data in case of technical issues. Weather Buoy data is a core Marine Institute data asset of continuous data collection since 2001 when the first moored surface buoy was deployed to now including a network of 5 buoys deployed and observing the wave climate of the ocean.

Data set reference and name	1. **3.1.IE.MarineInstitute.WaveBuoyNetwork** 2. **OF.PointTimeSeriesObservation.IE.MarineInstitute.WaveBuoyNetwork**
Data set description	The Irish Wave Buoy Network of three moored surface buoys collect data on wave behaviour including peak period, peak direction, upcross period, significant wave height (cm), mean wave direction (degreeTrue), sea surface height of the highest wave (Hmax), period of the highest wave (THmax), mean direction of sea water velocity and mean sea water speed on an hourly basis. This is disseminated via the Marine Institute ERRDAP data server and these hourly observations are from the national ¼ scale wave energy test site in Galway Bay and the full scale wave energy test site near Belmullet Co. Mayo. The wave buoys have been in operation since 2008 with some periods of observation downtime due to technical issues. This wave buoy point time-series observation data is very useful to Engineer Ed when considering developing a wave energy converter and his requirement for access to real-time data on waves in order to predict the performance of his device and to monitor its safety once it is deployed in the environments provided by the test facilities in Galway Bay and the North Atlantic Ocean near Belmullet.
Standards and metadata	Point time-series observation data on waves is presently available in a number of common open formats including CSV (comma-separated ASCII text table) and JSON (Java Script Object Notation). ISO 19139 metadata describing the observation data is hosted on the Irish Spatial Data Exchange and Irelands Open Data Catalogue data.gov.ie. http://www.isde.ie/#/55eb27e0-2fc3-4dab-9963-a99d12402a9e
Data Sharing	The Marine Institute presently publish point time-series wave data via an ERDDAP data server under a Creative Common CC-BY 4.0 open licence. Users who require this data simply access the http link provided and choose the variables of interest (e.g. wave) to fit their requirements. http://erddap.marine.ie/erddap/tabledap/IWaveBNetwork.html
Archiving and preservation (including storage and backup)	The Marine Institute ERDDAP server is hosted on a backend SQL Server database with standard SQL Server backup devices used to preserve data in case of technical issues. Weather Buoy data is a core Marine Institute data asset of continuous data collection since 2008 when the

Data set reference and name	1. **3.1.IE.MarineInstitute.EastAtlanticSWANWaveModel** 2. **OF.GridObservation.IE.MarineInstitute.EastAtlanticSWANWaveModel**
Data set description	The SWAN wave model runs for a domain covering Irish waters at a resolution of 0.025 degrees. The model runs every day to produce a 6 day forecast using NCEP GFS wind forcing and FNMOC Wave Watch 3 data for the wave boundaries. The forecast is generated for research purposes and for comparison with measured wave parameters. Modelled parameters include, Significant Wave Height (m), Mean Wave Direction (degreesTrue) and Mean Wave Period (seconds). Predictions should not be used for safety critical applications. Please note that weather data used for hindcast models is more reliable than weather data used for future model forecasts. At any one time, model data is available for the previous 30 days and 6 days into the future. SWAN is a third-generation wave model, developed at Delft University of Technology, that computes random, short-crested wind-generated waves in coastal regions and inland waters. This wave model data is critical to Engineer Ed in providing a real-time snapshot of the generic wave climate in the Atlantic Ocean around Ireland. The regional marine picture of wave climate allows Ed to potential decide on the constraints and opportunities for deployment of his device in the marine environment.
Standards and metadata	Grid time-series observation data on wave climate is presently available in a number of common open formats including CSV, esriAscii, JSON, GeoTiff, PNG, PDF and the ocean data format NetCDF. ISO 19139 metadata describing the model data is hosted on the Irish Spatial Data Exchange and Irelands Open Data Catalogue data.gov.ie. http://www.isde.ie/#/4b0e8380-9b9b-42f3-bc56-91b6bcda7d9a
Data Sharing	The Marine Institute presently publish gridded wave data via an ERDDAP data server under a Creative Common CC-BY 4.0 open licence. Users who require this data simply access the http link provided and choose the variables of interest (e.g. mean wave direction) to fit their requirements.

	first moored surface buoy was deployed to a network of 3 buoys now deployed and observing the wave climate of the ocean at the designated ¼ scale and full scale test facilities.

	http://erddap.marine.ie/erddap/griddap/IMI_EATL_WAVE.html
Archiving and preservation (including storage and backup)	The Marine Institute ERDDAP server is linked to the Thredds server hosting model data with standard backup devices used to preserve data in case of technical issues with Thredds. The SWAN wave model grid observation data is a core Marine Institute data asset of continuous data collection since 2010 when the first successful models on the East Atlantic were released to production.

Data set reference and name	1. **3.1.IE.MarineInstitute.IrishNationalTideGaugeNetwork** 2. **OF.PointTimeSeriesObservation.IE.MarineInstitute.IrishNationalTideGaugeNetwork**
Data set description	Network of 14 tide gauges located around the coast of Ireland collecting water level observation data as part of the development of a permanent tidal monitoring infrastructure. Parameters collected include; Station (ID), dateTime (yyyy-mm-ddThh:mm:ss), Water Level (Sea Level above Lowest Astronomical Tide [LAT], m) and Water Level (Water Level Ordnance Datum Malin Head, m). Observation data is realised on a hourly frequency daily since deployments started in 2007. The tide movement observation data is useful to Engineer Ed in understanding the marine environment in which his device may be deployed.
Standards and metadata	Point time-series observation data on water level observations for tides is presently available in a number of common open formats including CSV and JSON. ISO 19139 metadata describing the observation data is hosted on the Irish Spatial Data Exchange and Irelands Open Data Catalogue data.gov.ie. http://www.isde.ie/#/0da11e42-2d03-4c64-8b0d-0e78c04efc34
Data Sharing	The Marine Institute presently publish point time-series tide data via an ERDDAP data server under a Creative Common CC-BY 4.0 open licence. Users who require this data simply access the http link provided and choose the variables of interest (e.g. water level) to fit their requirements. http://erddap.marine.ie/erddap/tabledap/IrishNationalTideGaugeNetwork.html
Archiving and preservation	The Marine Institute ERDDAP server is hosted on a backend SQL Server database with standard SQL Server backup devices used to preserve

(including storage and backup)	data in case of technical issues. Tide gauge network is a core Marine Institute data asset of continuous data collection since 2007 when the first tide gauge was deployed to a network of 14 tide gauges now deployed and observing the tidal movements around the coastline of Ireland at key locations.

TUT datasets

Data set reference and name	**2.9.EE.TUT.OGIServiceCoproductionNeeds**
Data set description	The dataset was generated by TUT in the framework of Tasks 1.2 and 1.4 of WP1. The data was collected through e-mail interviews and online survey involving key actors in the field of open data and public sector innovation in the six pilot countries: Belgium, Greece, Estonia, Ireland, Lithuania and the UK. The respondents were chosen to represent public administrations as well as relevant business and civil society actors, based on the suggestions of OGI pilot partners in each country.
	The questionnaire consisted of 11 questions of varying length and depth. The questionnaire was disseminated to 120 respondents and returned 63 responses.
	The collected data is qualitative and involves the respondents' experience with open data-driven co-creation, their opinions on issues such as drivers and barriers to open data-driven co-creation, the capacities and needs of the organisations involved, good examples of policies that have been used to encourage open data innovation, etc.
	The data will be used in an anonymized and generalized form in a summary report of the survey results in deliverable D1.1. The results may potentially also be used in this form in related scientific publications. In addition to the OGI project, the generalized results could be potentially interesting for researchers, policy-makers, different types of users and providers of open data. There is also potential value in comparing the data with previous studies on open data innovation.
Standards and metadata	Data will be stored in xls and csv table format.
Data Sharing	The original, identifiable data will be stored at the Tallinn University of Technology and will not be shared with any other parties outside the research team. The data will be published only in a generalized and anonymized form through project reports and scientific publications, applying open access rules whenever possible.
Archiving and preservation (including storage and backup)	The original data will be stored by the Tallinn University of Technology in a secure server in Estonia, which only the research team has access to. The key allowing re-identification will be stored in an encrypted form only until necessary for the analysis, and will be deleted after the project's end.
	The anonymized data will be stored on a secure server of the Tallinn

	University of Technology up to 5 years after the project's end date. No additional resources are needed for the preservation of the data.

Data set reference and name	3.2.EE.TUT.EstonianLandCadastre
Data set description	The Estonian Land Cadastre is a database owned by the Estonian Land Board, an executive agency under the jurisdiction of the Estonian Ministry of the Environment. The cadastre contains data about land such as the county where it is located, local government, settlement unit, address, attribute, registration date, changing date, intended uses, area, evaluation zone, fertility zone, etc.
	Existing data from the Estonian Land Cadastre will be re-used for the pilot to produce an aggregated dataset.
	The data is currently accessible through the Land Board's web application (http://xgis.maaamet.ee/). Queries to the database are mostly free but some are available for a small fee. The software is developed on .NET platform in C# language and MS SQL database servers are being used. Information between the Land Cadastre and other registries is exchanged via the secure data exchange system X-Road.
	The data will be used for the Estonian pilot case within OGI and will as such be useful for citizens (property or land buyers, sellers and renters), real estate brokers, real estate developers, investors, notaries, public administration officials responsible for urban planning and strategic development (in particular at the local level).
	Pre populated CSV lists are available here: http://xgis.maaamet.ee/adsavalik/ads?xBTN.newvalja
Standards and metadata	CSV format
Data Sharing	Public online access will be provided through the web solution to be created in the framework of the pilot.
Archiving and preservation (including storage and backup)	To be determined.

Data set reference and name	1.7.EE.TUT.LandRegister
Data set description	The Land Register is a register of real rights in immovable property (immovable property = a plot of land, an apartment ownership, a right of superficies or a right of superficies in apartments). The owner of the database is the Estonian Ministry of Justice. Existing data from the Estonian Land Register will be re-used for the pilot to produce an aggregated dataset.
	The register includes data on the owners of immovables and the

	restricted real rights encumbering the immovable: mortgages, servitudes, real encumbrances, rights of superficies, rights of pre-emption and notations (preliminary notations, notations concerning the prohibition, objections and notes). The potentially relevant data for the pilot includes the registered immovable number, validity mark, digital file, land registry division, type of immovable, name of the registered immovable, cadastral code, specific purpose, address, area, apartment no, address, owner, restrictions, etc.
	The data is currently accessible from the Land Register. Basic queries are for free but detailed queries cost 1-2 euros. The existing service allows a verification of general data, size, owners, restrictions and encumbering mortgages of immovable properties. Users can make queries in the Land Register by entering the address of a land or building. More detailed queries can also be made by location (address or cadastral code), by immovable and by person.
	The data will be used for the Estonian pilot case within OGI and will as such be useful for citizens (property or land buyers, sellers and renters), real estate brokers, real estate developers, investors, notaries, public administration officials responsible for urban planning and strategic development (in particular at the local level).
Standards and metadata	To be determined
Data Sharing	Public online access will be provided through the web solution to be created in the framework of the pilot.
Archiving and preservation (including storage and backup)	To be determined

Data set reference and name	**4.3.1.EE.TUT.EstonianBuildingRegister**
Data set description	The Estonian Building Register is a database owned by the Estonian Ministry of Economic Affairs and Communications (MKM). Some data from the register will be re-used for the pilot, e.g. the type of property, general data about the building, type of building, year of first use, situation of the building, purpose/function of the building, surface area, number of floors, constructions and materials, technical data, data about electricity, sewerage, water supply and heating, etc.
	The data will be used for the Estonian pilot case within OGI and will as such be useful for citizens (property or land buyers, sellers and renters), real estate brokers, real estate developers, investors, notaries, public administration officials responsible for urban planning and strategic development (in particular at the local level).
Standards and metadata	To be determined

Data Sharing	Public online access will be provided through the web solution to be created in the framework of the pilot.
Archiving and preservation (including storage and backup)	To be determined

Data set reference and name	4.3.1.EE.TUT.KVee
Data set description	This information was scraped from KV.ee (an online housing portal in Estonia) from 2009 to 2015. This data provides address, type of property, value, condition, etc. Using this information, it will be possible to build a general idea of the value of different types of properties based on area.
Standards and metadata	CSV file, some characters will need to be fixed as the Estonian have not come through correctly.
Data Sharing	Public online access will be provided through the web solution to be created in the framework of the pilot.
Archiving and preservation (including storage and backup)	To be determined

Data set reference and name	1.8.EE.TUT.CrimeStatisticalData
Data set description	The criminal data is recorded and monitored by the Estonian Ministry of Justice. Currently, this data is only accessible here: http://www.kriminaalpoliitika.ee/et/statistika-ja-uuringud/baromeeter. Soon we will be receiving a data set, which will provide the following information: crime total, crime by type, location the activity took place, time of the crime. This data set will be useful in helping individuals gauge the safety of different neighbourhoods or locations within Estonia. As this data will go back a number of years, it will also be possible to track trends in crime so users can check whether an area is getting worse or better in regards to crime (also sortable by crime type).
Standards and metadata	Likely to be received in CSV format.
Data Sharing	Public online access will be provided through the web solution to be created in the framework of the pilot.
Archiving and preservation (including storage and backup)	To be determined

MAREG datasets

Data set reference and name	**2.5.GR.MAREG.GreekPublicAgenciesPersonnel**
Data set description	Data on the personnel of Greek Public Agencies will be generated from the Civil Servants' Registry. The data set will be used in the production of data analytics in combination with datasets of Government Vehicles. Data is already being published in the apografi.gov.gr platform.
Standards and metadata	Data will be published as linked data cubes using the RDF data cube vocabulary. Metadata will be also created using the DCAT and/or PROV ontology standards to accompany the data cubes.
Data Sharing	The generated data set will be available through the website of the project.
Archiving and preservation (including storage and backup)	Data will be stored in MAREG's website for 5 years after the end of the project. No additional costs required.

Data set reference and name	**2.4.4.GR.MAREG.GovernmentVehiclesOperationMaintenace**
Data set description	Data on the operation and maintenance of Government Vehicles will be collected by an Information System, which will monitor and keep track of all aspects of Government Vehicles. The technical annex for the respective EU funded project has recently been submitted. Data on the life cycle of Government Vehicles will be useful for policy makers of the Ministry of Interior and Public Administration, as well as for all Public Agencies using Government Vehicles and other interested third parties, such as insurance and toll operation companies.
Standards and metadata	Data will be published as linked data cubes using the RDF data cube vocabulary. Metadata will be also created using the DCAT and/or PROV ontology standards to accompany the data cubes.
Data Sharing	Data will be published in the website of the project after receiving approval from the Minister. Certain data columns may be excluded from sharing.
Archiving and preservation (including storage and backup)	Data will be stored in MAREG's website for 5 years after the end of the project. No additional costs required.

Data set reference and name	3.2.GR.MAREG.GreekMunicipalitiesPrefecturesRegions
Data set description	Data on the personnel of Greek Public Agencies will be requested from the Hellenic Statistical Authority. The data set will be used in the production of data analytics in combination with datasets of Government Vehicles. Data is already being published in the www.statistics.gr portal
Standards and metadata	Data will be published as linked data cubes using the RDF data cube vocabulary. Metadata will be also created using the DCAT and/or PROV ontology standards to accompany the data cubes.
Data Sharing	The generated data set will be available through the website of the project.
Archiving and preservation (including storage and backup)	Data will be stored in MAREG's website for 5 years after the end of the project. No additional costs required.

Data set reference and name	2.4.4.GR.MAREG.GovernmentVehiclesLifecycle
Data set description	Data on the life cycle of Government Vehicles will be collected by an Information System, which will monitor and keep track of all aspects of Government Vehicles. The technical annex for the respective EU funded project has recently been submitted. Data on the life cycle of Government Vehicles will be useful for policy makers of the Ministry of Interior and Public Administration, as well as for all Public Agencies using Government Vehicles and other interested third parties, such as automobile vendors.
Standards and metadata	Data will be published as linked data cubes using the RDF data cube vocabulary. Metadata will be also created using the DCAT and/or PROV ontology standards to accompany the data cubes.
Data Sharing	Data will be published in the website of the project after receiving approval from the Minister. Certain data columns may be excluded from sharing.
Archiving and preservation (including storage and backup)	Data will be stored in MAREG's website for 5 years after the end of the project. No additional costs required.

Data set reference and name	2.4.4.GR.MAREG.GovernmentVehicles
Data set description	Data describing Government Vehicles have been collected by the Ministry of Interior from Public Agencies using Government Vehicles. An .xls template was sent to all Public Agencies, who used it to produce their own .xls files, which were later merged into one unified .xls file. The dataset is constantly being undated to include all changes regarding Government Vehicles. The dataset includes all data that are necessary to describe the vehicle itself, as well as its use by the respective Public Agency. A similar dataset exists and has been provided by the Ministry of Infrastructure, Transport and Networks, but comparison of the two datasets showed significant inconsistencies between them. Government Vehicles Data will be useful to policy makers of the Ministry of Interior, as well as to Public Agencies using Government Vehicles and to other interested third parties for their exploitation in other apps and the generation of statistics.
Standards and metadata	Data will be published as linked data cubes using the RDF data cube vocabulary. Metadata will be also created using the DCAT and/or PROV ontology standards to accompany the data cubes.
Data Sharing	Data will be published in the website of the project after receiving approval from the Minister. Certain data columns (e.g. Vehicle Registration Numbers) may be excluded from sharing.
Archiving and preservation (including storage and backup)	Data will be stored in MAREG's website for 5 years after the end of the project. No additional costs required.

Flemish Government datasets

Data set reference and name	3.1.BE.VLO.AirWaterPollutionMetadata
Data set description	Flemish companies need to report yearly on their pollutions (air and water emissions). The metadata of these reports are kept in an archival system (DSPACE). These metadata will be converted into RDF during the project and loaded in a triple store.

	These data are interesting to be combined with other Flemish Government datasets (IMJV data, CBB) and already open datasets such as KBO (National Registry of Companies), NIS (Administrative geography nomenclature), NACE (Economical Activities Nomenclature), CRAB (Flemish Addresses Basic Registry).
Standards and metadata	These metadata on the documents are using the Dublin Core metadata element set. For metadata on the metadata the PROV vocabulary will be used.
Data Sharing	These data will be available via a SPARQL endpoint to be used in applications available to the general public.
Archiving and preservation (including storage and backup)	The data are already in an archival system, which has all the necessary procedures for long lasting preservation in place.

Data set reference and name	**3.1.BE.VLO.AirWaterPollutionData**
Data set description	Flemish companies need to report yearly on their pollutions (air and water emissions). The emission data are available in XML files and archived in an archival system (DSPACE). These data will be converted into RDF during the project and loaded in a triple store.
Standards and metadata	These emission data will be converted into the RDF Data Cube Vocabulary. These data are interesting to be combined with other LNE datasets (IMJV metadata, CBB) and already open datasets such as KBO (National Registry of Companies), NIS (Administrative geography nomenclature), NACE (Economical Activities Nomenclature), CRAB (Flemish Addresses Basic Registry). For metadata on the data the PROV vocabulary will be used.
Data Sharing	These data will be available via a SPARQL endpoint to be used in applications available to the general public. All privacy related data will be filtered out before publishing.
Archiving and preservation (including storage and backup)	The data are already in an archival system, which has all the necessary procedures for long lasting preservation in place.

Data set reference and	**2.5. BE.VLO.CBB**

name	
Data set description	The Environment, Nature and Energy Department (LNE) of the Flemish government keeps company (and their sites) information in a PostgreSQL DB. These data will be virtualised as a SPARQL endpoint during the project.
Standards and metadata	These company data will be converted into the org, regorg and locn vocabularies. For metadata on the data the PROV vocabulary will be used.
Data Sharing	These data will be available via a SPARQL endpoint to be used in applications available to the general public. No privacy related data are involved.
Archiving and preservation (including storage and backup)	The backup procedures are already in place for the PostgreSQL system.

EL datasets

Data set reference and name	**2.3.LT.EnterpiseLithuania.AdvertisementPermitsData**
Data set description	Vilnius city municipality issues permits for installation of outdoor advertising in municipal territory for businesses. The data is available in CSV format file and published in Github (https://github.com/vilnius/reklamosleidimai/blob/master/data/REKLAMOS_REGISTRAS.csv).
Standards and metadata	This data will be converted into the RDF Data Cube Vocabulary. For metadata on the data the PROV vocabulary will be used.
Data Sharing	This data will be available via a SPARQL endpoint to be used in applications available to the general public. There is no privacy related data.
Archiving and preservation (including storage and backup)	The data is already in Github.com system, which has all the necessary procedures for long lasting preservation in place.

Data set reference and name	**1.4.LT.EnterpriseLithuania.PermitsHygienePassports**
Data set description	National public health center issues Permits-Hygiene passports for businesses. The data is available in HTML format file and published in

	website of the authority (http://vssis.vvspt.lt/lhp/index.php?veikla10=&veikla09=&veikla07=&veikla05=&objpavad=&objkodas=&vk=&objapskritis=Vilniaus&lhp_data_nuo=&veikvieta=&lhp_data_iki=&veikpavad=&lhpnr=).
Standards and metadata	This data will be converted into the RDF Data Cube Vocabulary. For metadata on the data the PROV vocabulary will be used.
Data Sharing	This data will be available via a SPARQL endpoint to be used in applications available to the general public. There is no privacy related data.
Archiving and preservation (including storage and backup)	The data is already available in website of the authority, which has all the necessary procedures for long lasting preservation in place.

Data set reference and name	**1.2.UK.TRAF.NEET**
Data set description	Number of 16 - 18 year olds not in education, employment or training. Updated monthly, case-level data is added to the Core Plus database by Connexions workers. Data will be provided to Trafford Innovation and Intelligence Lab in anonymised form, aggregated to Ward level.
Standards and metadata	Data is recorded according to requirements of the case management system. Statistical data will be marked up with metadata in line with Dublin Core standard.
Data Sharing	Data to be shared is anonymised and aggregated by the data collectors (Connexions) prior to release to the Lab. Data will be uploaded to Publish My Data platform, with appropriate geographical and temporal identifiers to allow linking data. Numbers will be suppressed if there are fewer than five cases in an area. Data will be released under Open Government Licence.
Archiving and preservation (including storage and backup)	Statistical data will be retained as long as required, to support operational and strategic research and decision-making. The Connexions case management system has its own procedures for storing, preserving and archiving case management data.

Data set reference and name	**1.6.UK.TRAF.Benefits**
Data set description	The number and percentage of people claiming different types of benefits across Greater Manchester. Data is available to Lower Super Output Area

	geographical level, and can be split into age groups, and gender. Data is available monthly / quarterly, and is provided by the Office For National Statistics, through their NOMIS data portal.
Standards and metadata	Data is recorded according to requirements of the case management system held by the Department of Work and Pensions. Statistical data will be marked up with metadata in line with Dublin Core standard.
Data Sharing	Data to be shared is anonymised and aggregated by the data collectors (Office for National Statistics) prior to publishing on NOMIS. Data will be dynamically converted to RDF, using an API, and uploaded to Publish My Data platform, with appropriate geographical and temporal identifiers to allow linking data. Numbers will be suppressed if there are fewer than five cases in an area. Data is available under Open Government Licence.
Archiving and preservation (including storage and backup)	Statistical data will be retained as long as required, to support operational and strategic research and decision-making. The DWP system which holds the source data should have its own procedures for storing, preserving and archiving case management data.

Data set reference and name	**1.8.UK.TRAF.Crime**
Data set description	The number of crimes recorded each month across Greater Manchester. Data is broken down by crime type and location. Data is available from data.police.uk.
Standards and metadata	Data is recorded according to requirements of the Home Office. Statistical data will be marked up with metadata in line with Dublin Core standard.
Data Sharing	Data to be shared is anonymised and aggregated by the data collectors (the Home Office). Geolocated crime data is snapped to nearby location to prevent identification of individuals. Data will be dynamically converted to RDF, using an API, and uploaded to Publish My Data platform, with appropriate geographical and temporal identifiers to allow linking data. Data is available under Open Government Licence.
Archiving and preservation (including storage and backup)	Statistical data will be retained as long as required, to support operational and strategic research and decision-making. Police forces across the country are responsible for the source data.

Data set reference and name	**1.1.UK.TRAF.Census**
Data set description	Anonymised and aggregated responses to various questions from the

	2011 Census in the UK. Travel to work, demographics, health and other measures will be used in the projects.
Standards and metadata	Census data is subject to internal standards. Statistical data arising from the dataset will be referenced according to the Dublin Core.
Data Sharing	Data to be shared is anonymised and aggregated by the data collectors. Data will be dynamically converted to RDF, using an API, and uploaded to Publish My Data platform, with appropriate geographical and temporal identifiers to allow linking data. Data is available under Open Government Licence.
Archiving and preservation (including storage and backup)	Statistical data will be retained as long as required, to support operational and strategic research and decision-making.

Data set reference and name	**1.2.UK.TRAF.JCPlus**
Data set description	Locations of all JobCentre Pluses in Greater Manchester, along with details of services provided.
Standards and metadata	Co-ordinates referenced according to WGS84 coordinate reference system. Metadata fits Dublin Core.
Data Sharing	Data uploaded to Publish My Data and made available through SPARQL endpoint.
Archiving and preservation (including storage and backup)	Data will be continually updated as Job Centre pluses change as a result of the project. Master data to be retained and maintained by Department for Work and Pensions.