ISA Action 1.17: A Reusable INSPIRE Reference Platform (ARE3NA)

Study on RDF & PIDs for INSPIRE

D2: State of Play

Alice Vasilescu

Disclaimer

The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of this publication.

Bibliographic Information:

Vasilescu A. Study on RDF and PIDs for INSPIRE D2: State of Play. European Commission; 2015. JRC96756

TABLE OF CONTENTS

List of abbreviations

ISA	Interoperability Solutions for European Public Administrations
ADMS	Asset Description Metadata Schema
API	Application programming interface
ARK	Archival Resource Key
DCMI	Dublin Core Metadata Initiative
DOI	Digital Object Identifier
EC	European Commission
GML	Geography Markup Language
HTTP	HyperText Transfer Protocol
IETF	Internet Engineering Task Force
ISBN	International Standard Book Number
JRC	Joint Research Centre
LOD	Linked Open Data
OSS	Open source software
OSLO	Open Standards for Local Administrations
OWL	Web Ontology Language
PID	Persistent identifier
RDF	Resource Description Framework
SKOS	Simple Knowledge Organization System
SDI	Spatial Data Infrastructures
UML	Unified Modeling Language
URI	Uniform resource identifier
URL	Uniform resource locator
URN	Uniform resource name
VoiD	Vocabulary of Interlinked Datasets

List of terms and definitions

ARK	Archival Resource Key ARK is a scheme for the persistent identification of information objects, which can include finding aids and other metadata as well as digital archival objects; however, it can also be used to assign a persistent name to other resources, e.g. physical objects such as books and intangible objects (examples given include diseases, vocabulary terms and performances). ARK was developed as an alternative to schemes like PURLs, URNs and Handles, which address the problem of broken URLs by using a stable, indirect hostname scheme. Instead, the ARK scheme is founded on the principle that persistence is a matter of service, not syntax – it is reliant on the continued stability and support of the service behind the identifiers. http://www.paradigm.ac.uk/workbook/metadata/pids-ark.html https://www.google.be/?gws_rd=cr&ei=naCTUrGuLKWM0wWnyYC4Dw#q=ark+identifier
DOI	Digital Object Identifier A DOI name is an identifier (not a location) of an entity on digital networks. It provides a system for persistent and actionable identification and interoperable exchange of managed information on digital networks. A DOI name can be assigned to any entity — physical, digital or abstract — primarily for sharing with an interested user community or managing as intellectual property. The DOI system is designed for interoperability; that is to use, or work with, existing identifier and metadata schemes. DOI names may also be expressed as URLs (URIs). http://www.doi.org/index.html
HTTP URI	HTTP URIs, in the web architecture, have been used to denote documents -- "web pages" informally, or "information resources" more formally. However, with the growth of the Semantic Web, which uses URIs to denote anything at all, the urge to use and practice of using HTTP URIs for arbitrary things grew steadily. The W3C Technical Architecture group eventually decided to resolve the architectural problem that if an HTTP response code of 200 (a successful retrieval) was given, that indicated that the URI indeed was for an information resource, but with no such response, or with a different code, no such assumption could be made. This compromise resolved the issue, leaving a consistent architecture. http://www.w3.org/DesignIssues/HTTP-URI.html
Ontology	An ontology is a formal specification of a shared conceptualization. In the context of computer and information sciences, an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. Ontologies are considered one of the pillars of the Semantic Web, although there exist many definitions. http://semanticweb.org/wiki/Ontology http://www-ksl.stanford.edu/kst/what-is-an-ontology.html
OWL	Web Ontology Language The W3C Web Ontology Language (OWL) is a Semantic Web language designed to represent rich and complex knowledge about things, groups of things, and relations between things. OWL is a computational logic-based language such that knowledge expressed in OWL can be exploited by computer programs, e.g., to verify the consistency of that knowledge or to make implicit knowledge explicit. OWL documents, known as ontologies, can be published in the World Wide Web and may refer to or be referred from other OWL ontologies. OWL is part of the W3C's Semantic Web technology stack, which includes RDF, RDFS, SPARQL, etc. http://www.w3.org/2001/sw/wiki/OWL
PID	Persistence Identifier An identifier is a unique identification code that is applied to "something", so that the

	"something" can be unambiguously referenced. For example, a catalogue number is an identifier for a particular specimen, and an ISBN number is an identifier for a particular book. It is an overarching term and could take various forms such as persistent identifier systems include: Archival Resource Keys (ARKs), Digital Object Identifiers (DOIs), Persistent Uniform Resource Locators (PURLs), Uniform Resource Names (URNs).
PURL	Persistent Uniform Resource Locators PURLs are Web addresses that act as permanent identifiers in the face of a dynamic and changing Web infrastructure. Instead of resolving directly to Web resources, PURLs provide a level of indirection that allows the underlying Web addresses of resources to change over time without negatively affecting systems that depend on them. This capability provides continuity of references to network resources that may migrate from machine to machine for business, social or technical reasons. http://purl.oclc.org/docs/index.html
RDF	Resource Description Framework RDF is a standard model for data interchange on the Web. RDF has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed. RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a "triple"). Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications. http://www.w3.org/2001/sw/wiki/RDF
RDFS	RDF Vocabulary Description Language 1.0: RDF Schema RDFS is a general-purpose language for representing simple RDF vocabularies on the Web. Other vocabulary definition technologies, like OWL or SKOS, build on RDFS and provide language for defining structured, Web-based ontologies which enable richer integration and interoperability of data among descriptive communities. http://www.w3.org/2001/sw/wiki/RDFS
Vocabulary	On the Semantic Web, vocabularies define the concepts and relationships (also referred to as "terms") used to describe and represent an area of concern. Vocabularies are used to classify the terms that can be used in a particular application, characterize possible relationships, and define possible constraints on using those terms. In practice, vocabularies can be very complex (with several thousands of terms) or very simple (describing one or two concepts only). There is no clear division between what is referred to as "vocabularies" and "ontologies". The trend is to use the word "ontology" for more complex, and possibly quite formal collection of terms, whereas "vocabulary" is used when such strict formalism is not necessarily used or only in a very loose sense. Vocabularies are the basic building blocks for inference techniques on the Semantic Web. Core vocabularies refer to the vocabularies developed in the context of the ISA programme: person, business, location and public service http://www.w3.org/standards/semanticweb/ontology
URI	In computing, a uniform resource identifier (URI) is a string of characters used to identify a name of a web resource. Such identification enables interaction with representations of the web resource over a network, typically the World Wide Web, using specific protocols. Schemes specifying a concrete syntax and associated protocols define each URI.
URL	Uniform Resource Locator Is a specific character string that constitutes a reference to a resource on the web. It is the global address of documents and other resources on the World Wide Web. http://www.webopedia.com/TERM/U/URL.html
URN	Uniform Resource Name

Uniform Resource Names (URNs) are intended to serve as persistent, location-independent, resource identifiers.

http://datatracker.ietf.org/wg/urn/charter/

1. INTRODUCTION

1.1. Context

This deliverable has been prepared in the context of the INSPIRE Directive (2007/2/EC)[1] and the ARE3NA project (ISA Action 1.17) which aims to create a platform to support the reuse of location/geospatial data, metadata and services to support cross-border and cross-sector interoperability tasks in public administrations across the European Union (EU).

The INSPIRE Directive –Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community– entered into force in May 2007. The European Commission initiated the INSPIRE initiative to enhance the sharing of spatial data between public authorities in Europe and between the Member States and the European Institutions in particular. The EU Member States transposed the Directive into national legislation between May 2007 and May 2009. They have to implement this Directive and its Implementing Rules by following the Technical Guidelines which are based on international standards (ISO/TC 211, CEN/TC 287 and OGC). However the different approach of implementing the standards, the regular evolution of standards and challenges in coordinating changes between standards, alongside varying choices in the technologies being adopted are creating interoperability challenges.

In order to address these challenges, the European Commission's (EC) Joint Research Centre (JRC), as part of the Interoperability Solutions for European Public Administrations (ISA[2]) Programme, has established a *Reusable INSPIRE Reference Platform* (ARE3NA – ISA Action 1.17)[3].

1.2. Motivation

Whereas INSPIRE is focusing on addressing the interoperability of geospatial data sets and services through the creation of data models (using UML) and geospatial encodings mechanisms (using GML), for the exchange of data related to one of the 34 spatial data themes defined in the INSPIRE Directive, e-Government applications and tools usually use Linked Data based on the Resource Description Framework (RDF) of W3C. Several European projects and initiatives in Member States have created RDF vocabularies based on the conceptual INSPIRE UML data models. Several approaches have been applied and several issues remain unanswered:

- Lack of agreed rules or guidelines on how to create RDF vocabularies from the UML models. In the context of INSPIRE, data models are developed on a conceptual level using the Unified Modeling Language (UML), and the default encoding for most INSPIRE themes is based on the Geography Markup Language (GML)[4]. However the EU Member States and EU projects use the Resource Description Framework (RDF)[5] to make available the e-government applications and tools as Linked Data. Though some EU Member States and EU projects are creating RDF vocabularies based on the conceptual INSPIRE UML data models, there are no rules agreed or guidelines on how this creation should be performed.
- Lack of best practices and guidelines in the area of global identifiers PIDs. Though some EU Member States have created governance structures, processes, rules/guidelines and tools to create, manage, maintain and use persistent identifiers (PIDs) in their Spatial Data Infrastructures (SDIs) that INSPIRE is built upon, different approaches have been identified.

[1] Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE); http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32007L0002:EN:NOT

2 http://ec.europa.eu/isa

3 See: http://ec.europa.eu/isa/actions/01-trusted-information-exchange/1-17action_en.htm

4 See: http://www.opengeospatial.org/standards/gml

5 See: http://www.w3.org/RDF/

1.3. Scope

Before the elaboration of the common methodology and the guidelines, it is necessary to describe the current state-of-play for RDF and PIDs.

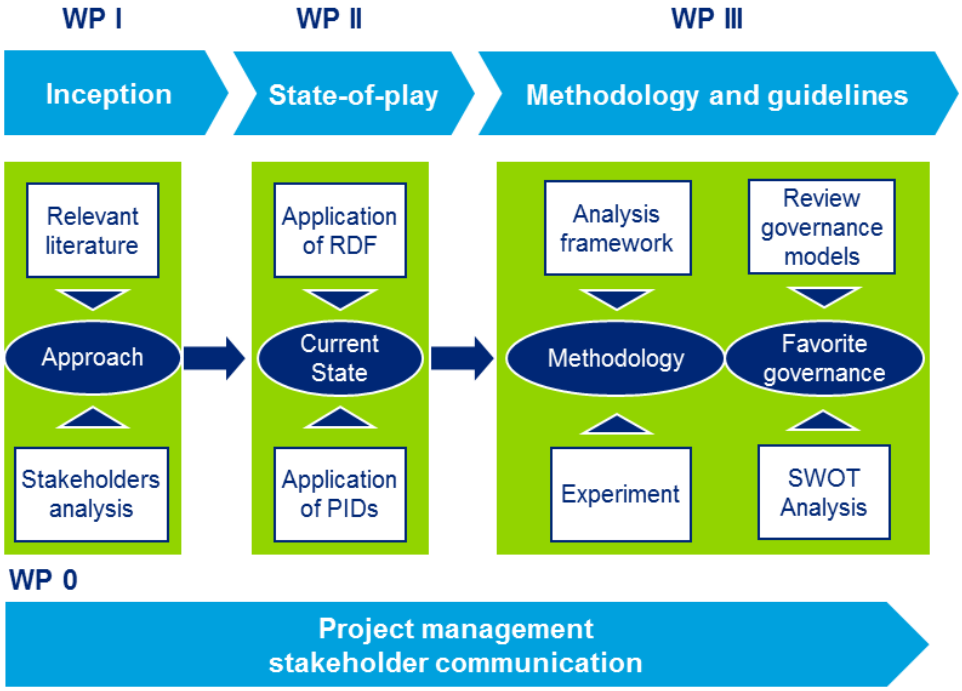


Figure 1: Scope JRC Study

The current document describes the context information about the Semantic Web and Linked Data, the methodology used for the research of State-of-Play together with the results and conclusions.

2. THE SEMANTIC WEB AND LINKED (OPEN) DATA

This chapter introduces the concepts and principles of the Semantic Web. It explains how Linked Data can be used to link data through the Web and by doing so creating new knowledge. It frames how Linked Data can support the efforts of the EU to improve interoperability between administrations, businesses and citizens, from the technical, semantic, as well as organisational and legal perspectives. Finally this chapter summarizes some of the challenges of publishing data as Linked Data.

2.1. Introducing the semantic web

The Semantic Web dates from 1998. It was coined as term for the first time by Tim Berners-Lee for a Web of data that can be processed by machines. The Semantic Web represents a different perspective on the World Wide Web to describe, query and reason on data and web content using HTTP, the Resource Description Framework (RDF) as a universal data structure, Uniform Resource Identifiers (URIs[6]) to denote things, SPARQL Protocol and RDF Query Language (SPARQL)[7] as a means to query the data, and different ontologies to represent, describe and annotate data. The major benefit of the Semantic Web lies in the ability of the Web to "know" and "understand" the data, to provide logic and meaning to the data. This should allow more automatic data and information processing and is expected to improve productivity and efficiency, e.g. because of more accurate Web search through the removal of ambiguity.

Linked Data, an integral and essential part of the Semantic Web, is a method for exposing, publishing and sharing structured data, with the idea of interlinking different datasets in order to use them more and more and to make them useful by making use of Semantic Web technologies. Indeed, by linking data with other data on the web the user can get new insights. The major benefit of Linked Data lies in the fact that it is sharable and extensible but, most importantly, re-usable. Furthermore, Linked Data gives the possibility to find new opportunities for data through new found relations between certain datasets. Especially in a world where there is an increasing pressure on publishing open data, the idea of linking different open data sets to each other seems very appealing.

[6] This is a more generic naming for anything on the web. Often it is also referred to as HTTP URIs which are defined as web information objects. However, those can take many different forms, generally they are abstractions which may have many different bit representations, as a function of, for example of time, content-type in which the bits are encoded, natural language in which a human-readable document is written, etc. It always must be used to refer to a unique conceptual object whose various representations have a very large an mount in common. See http://www.w3.org/DesignIssues/HTTP-URI.html.

[7] http://www.opengeospatial.org/standards/geosparql

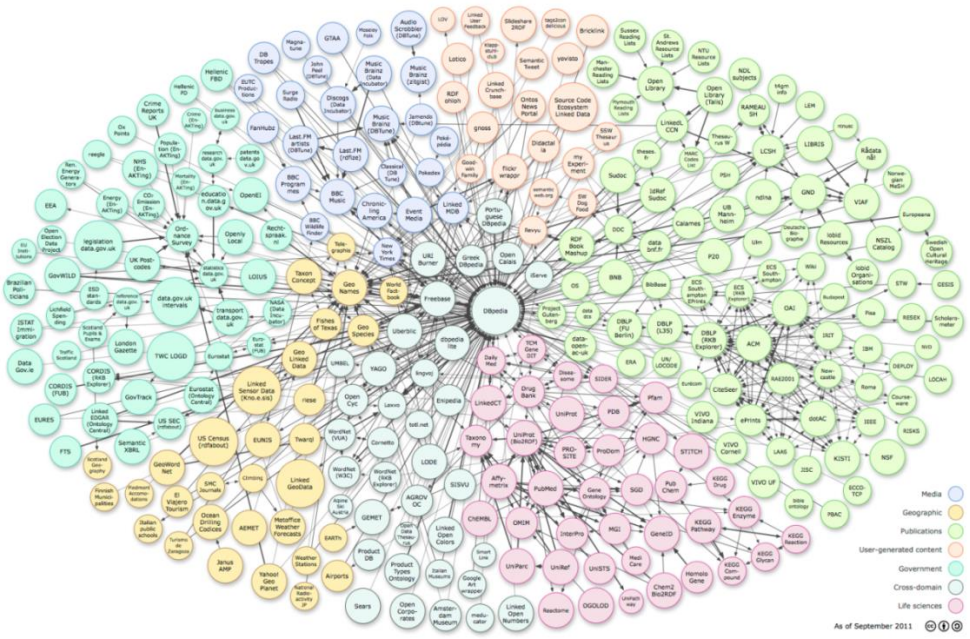


Figure 2: Different Linked Open Datasets and their links to other Linked Open Datasets in September 2011 (Jentzsch, 2011).

Linked Data is growing rapidly. In 2007, the published linked open datasets consisted of over two billion RDF triples, interlinked by over three million triples (W3, 2013). By September 2011 this had grown to 295 datasets with 31 billion RDF triples, interlinked by about 504 million RDF links (Heath, Hausenblas, Bizer, Cyganiak, & Hartig, 2008). Figure 2 gives a picture of the world of Linked Open Data by 2011.

The growing importance of Linked Data did not pass by unnoticed by the EU. The EU supports different projects involving Linked Data, for example the PlanetData project (http://planet-data.eu/), Linked Open Data 2 project (http://lod2.eu) and GeoKnow (http://Geoknow.eu). Moreover, the developments towards the Semantic Web and Linked Open Data are one of the cornerstones of the ISA programme of the EU (see section 2.4).

2.2. The 4 key principles and 5-star model of Linked Data

Linked Data builds upon standard Web technologies such as HTTP and URIs, but rather than using them to serve web pages for human readers, it extends them to share information in a way that can be read automatically by computers. This enables data from different sources to be connected and queried more easily. Tim Berners-Lee defined four important design principles for implementing Linked Data:

1. Use Uniform Resource Identifiers (URIs) to uniquely identify things;
2. Use HTTP URLs, corresponding to these URIs, so that information can be retrieved;
3. Provide metadata using open standards such as RDF;
4. Include links to related URIs, so that people can discover other, related things.

The four principles of Linked Data referred to above were developed by Tim Berners-Lee in 2006. Those principles did not yet mention openness at that time. The "Open" in Linked Open Data came from later projects, notably the W3C Linking Open Data project[8] and more recently, LOD2. Tim Berners-Lee proposed a 5 star rating system for data.

Table 1: The 5-star rating model (Berners-Lee)

[8] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

Rating	Requirement	Formats
★	Data available on the Web, whatever format, under an open license	HTML, PDF
★★	Data available as structured data in a format that it can be re-used	XML
★★★	Data available in an non-proprietary or open format	CSV
★★★★	Data available in RDF format and SPARQL to retrieve and use the data; use URI's to denote things so that you can point at it	RDF, URI
★★★★★	Link your data to other data to provide context	RDF, URI

2.3. How linked data can be used

The consumption of Linked Open Data (LOD) opens a new world of possibilities for any user on the Web. LOD allows users to link their data with third-party information and data collections. Benefits of using LOD emerge when one can generate mash-ups from distributed data sources, new applications are created with real time data and new knowledge is created from the linkage of data sources. Consuming Linked Data requires several steps to be taken.

First, concrete use cases should be defined. This can be a service or application that will help solving a particular problem. The own data and external data sources should be defined. Second, evaluate the relevant data sources that should be integrated. The evaluation of the quality needs is part of this step, including whether the data are up-to-date or not. Tools exist to explore what is available (e.g. Sindice[9]) and catalogues contain many data sources (e.g. The Data Hub[10]). In the third step the user needs to evaluate the applicable licences. Some licences are restrictive which might cause barriers for re-use. Particular attention should be paid to the restrictions regarding the mash-up of the data with other data. Fifth, it has to be defined which parts of the data sets will be (re-)used by defining consumption patterns. The sixth step is focussing on the alignment of vocabularies. Moreover, since the needed data sources are not always online, one might need cashing of the data. Finally users need to create user friendly GUI's to embed the linked data for end-users.

In the context of the Pilot 1 project on Linked Open Data in the Netherlands, and as part of the study *"principles for publishing, linking and accessing open government data as service on the Web"* (Archer et al., 2013), several use cases and projects describe the concrete examples of the application of Linked Data technology to support policy making or operational solutions. Some interesting examples are given in table

Use case	Area and Scope
City of Nijmegen	Linking spatial data (e.g. Building and Address Register) with cultural-historical data regarding monuments and information from DBPedia (e.g. linking to past events, linking to other buildings of the same architect, who lived there, etc.). Development of a viewer capable to recognise addresses based on the Core Vocabulary Location and a war monuments has been developed by 2CoolMonkeys for mobile devices.
'Huiskluis' (Housesafe) by Arcadis, Geonovum, Dutch Tax Service, Envolve …	Linking data related to our house from different sources: governmental data (e.g. permits), owner information, information about inhabitants, neighbours, businesses, etc., linked with information from the

[9] See http://sindice.com/ - the semantic web index: it is a platform with tools to help the user link their data with other web resources.

[10] See http://datahub.io/ - an open data management platform from the Open Knowledge Foundation

	owner/inhabitants (e.g. energy consumption). The project defined several potential applications based on the needs of the owners/inhabitants and selected three use cases: "fire in my house", "the house I love" and "I want to sell my house".
Supporting Fire Brigades by Geonovum and Netage	Based on the need to link their own (spatial) data to many other information sources (e.g. chemical products, accessibility buildings). The fire brigade of xxxx started to work with RDF to link own data with external data sources. One of the results is that the occurrence of fire incidents is projected on a wall in the fire station. Work started also to link information from the BAG (Central Register of Buildings and Addresses) to information on the complex legislation from government
Quality of Bathing Waters UK - DEFRA	Bathing Water Quality data was first published as linked data back in 2011. It exposes weekly sampling data for 500 locations around Britain's coast. The data is made accessible through a SPARQL endpoint using a Linked Data API. One application developed is the Bathing Water Explorer which allows users to link information on the quality, with images and factual information, local sewerage facilities, etc. DEFRA also provides a tool to generate widgets that can be embedded on any Web site and thus include

Table 2: Some examples of applications using Linked Data

2.4. Key challenges of publishing and using Linked Data

In *"principles for publishing, linking and accessing open government data as service on the Web"* Phil Archer (2012) focusses on the possible value, costs and benefits of Linked Open Data through different case studies. The main finding of the study is in line with what can be expected: Linked Open Government Data provides more flexible data integration, it increases data quality and reduces the costs and it allows the creation of new services whereby the key partners are the government itself and businesses and NGO's. These are, together with academic researchers, also the major customers of Linked Data. The major costs are in the development, maintenance and promotion of Linked Data. The study ends with an overview of challenges or "enablers and roadblocks" where the major benefits and problems of Linked Open Government Data, found in the case studies, are highlighted. The "enablers" mentioned are:

- *Efficiency gains in data integration – the network effect*: Linked Data supports the internal integration activities or facilitates data exchange within already existing collaboration structures. It is no longer necessary to download data of others within the organization.
- *Forward-looking strategies*: Some of the parties involved in Linked Open Government Data participate because of their mandate to distribute their information as widely as possible. In this case Linked Data is only one of the many data formats in which the data is published. But whereby Linked Data also offers more efficient mechanisms to create better integration across collections.
- *Increased linking and integrated services*: The value for re-users can often be found in the simplicity of linking external resources in their own user interface without the need of creating specific software or systems for specific providers.
- *Ease of model updates*: Traditional relational databases have a certain well-defined structure with tables, linked by primary keys which has its benefits but also makes it rather difficult to make changes in the structure. In the case of Linked Data, this is very easy and does not require changes in the existing architecture.

- *Ease of navigation*: Looking at URI's of certain data may reveal further information, which can help in navigation through complex data.
- *Open licensing and free access*: In many of the cases, open licenses are used, free of charge.
- *Enthusiasm from 'champions'*: It is often an enthusiastic individual that plays a leading role on the issue of Linked Data, especially in creating awareness of the possibilities and benefits.
- *Emerging best practice guidance*: Different players in the field of Linked Data are busy on creating best practice guides (e.g. W3C) for Linked Data and URI design which makes it easier for providers and (re)users to apply the same procedures.

The "roadblocks" mentioned in the study are:

- *Necessary investments*: The necessary investments differ among different projects: where Linked Data is only an additional format amongst other formats the investment is limited to negligible, when entirely new or replacement services are required, the costs are significantly higher but should always be compared to the costs of doing something similar for an alternative technology.
- *Lack of necessary competencies*: There is a lack of competencies, especially in other organizations like the data providers or the potential users and the first steps into Linked Data are often difficult for new users.
- *Perceived lack of tools*: Many publishers and users develop their own tools because they feel that the necessary tools are not available, especially for automated procedures. This is mainly because, although tools are available, they are far from performing as well as the current tools for relational databases.
- *Lack of service level guarantees*: The lack of service level guarantees hinders reuse by external parties. This lack is caused because the publisher mainly uses the Linked Data for its own benefit and not so much for third parties.
- *Missing, restrictive, or incompatible licenses*: The licence under which the data can be reused is not always available in which case the user should contact the data provider, which of course hinder reusability. Also very restrictive licences hinder the reusability.
- *Surfeit of standard vocabularies*: To many, it is evident to use Dublin Core for describing title, author, description and publication data of the published data, though this is not obligated. The development of many other standard vocabularies, with different other vocabularies based on these vocabularies may be confusing. When different data institutions use different vocabularies to describe the same things, generating Linked Data remains a challenge.
- *The inertia of the status quo*: Change in practice in the public sector had evolved (mostly) rather slow. Opening data creates fear of exposure and scrutiny for example, about the quality of the data. Making the data open exposes the errors in the data to people who may not take the full context into account.

2.5. Activities on Linked Data in the context of the ISA programme and in other DG's

One of the key actions under the ISA programme is *"Improving semantic interoperability in European eGovernment systems: Methodologies for the development of semantic assets"* (Action 1.1). The action starts from the observation that many barriers and challenges still exist to date to the exchange data during the execution of European Public Services. These barriers include divergent interpretations of the data, lack of commonly agreed and widely used metadata, absence of universal reference data (e.g. code lists, taxonomies), the multilingual challenge, etc. The Action tries to reduce these barriers and the (potential) negative impacts of semantic interoperability conflicts. The action is focussing on the collaborative development of Core Concepts and Vocabularies, the search and re-use of semantic

(interoperability) assets[11] with the use of enhanced indexing and browsing functionalities, and its promotional activities.

The main objectives include:

- Collecting, organizing, maintaining and providing access to a repository of highly reusable semantic interoperability assets as well as promoting best practices, experiences and lessons-learnt in the area of semantic interoperability.
- Providing the infrastructure via the Joinup platform for accessing, sharing and reusing semantic interoperability assets and open source software (OSS) that may be hosted in national repositories.
- Promoting the ADMS-based federation of semantic assets repositories.
- Supporting alignments and agreements on common definitions and specifications at the semantic layer.
- Promoting the use of the ISA Core Vocabularies at the European, national and local level to increase interoperability in the provision of European Public Services.
- Increasing the awareness on the importance of semantic interoperability and appropriate metadata management policies.
- Promoting linked data approaches and technologies for improving the interoperability of public administration systems.

Within this action, DG DIGIT developed some specific studies and activities in the field of Linked Data to better reach those objectives. The above mentioned study *"Improving semantic interoperability in European eGovernment systems: Methodologies for the development of semantic assets"* is one of the results of these efforts. There exists a close collaboration with the activities of DG CONNECT in this field. In this context, the Open Data Support project[12] aims to improve the visibility and facilitate the access to datasets published on local and national open data portals by supporting their publication and their visibility through the European Open Data Portal in order to increase their re-use within and across borders. The project provides to (potential) publishers of open datasets, three types of services: 1) data and metadata preparation, transformation and publication services that will enable data providers to share the metadata via the metadata infrastructure delivered by the project; 2) training services in the area of (linked) open data, aiming to build both theoretical and technical capacity to EU public administrations; and 3) IT advisory and consultancy services in the areas of linked open data technologies, data and metadata licensing, and business aspects and externalities of (linked) open data.

[11] A semantic asset is a collection of highly reusable metadata (e.g. xml schemata, generic data models) and reference data (e.g. code lists, taxonomies, dictionaries, vocabularies) which are used for e-Government system development. An interoperability asset describes resources that support the exchange of data in distributed information systems. They can be syntactic and semantic. See https://joinup.ec.europa.eu/contexthelp_faq/what-interoperability-asset.

[12] https://joinup.ec.europa.eu/community/ods/description

3. METHOD FOR ASSESSING THE STATE OF PLAY

This chapter describes the methodology used for the State-of-Play for RDF and PID Governance. It starts with an overview of how the study was conducted (desktop study, direct contacts and interviews), and then details the approach for assessing the results for the RDF and PID parts of the study.

3.1. Literature review and information from experts

The approach for assessing the different methods for generation of RDF from INSPIRE data models and the governance of PIDs was based on desktop research on the one hand and direct contacts with experts in the field on the other hand.

3.1.1. Literature review

Table 3 provides an overview of the literature that has been taken into account during the desktop research. The references indicated:

- In black are related to background aspects regarding the semantic web
- In green are related to general aspects of Linked (Open) Data
- In brown are related to the generation of RDF
- In blue are related to PID's

Table 3: Literature review

Author(s)	Date	Title	Type[13]	Description
Folmer, E., Reuvers, M., & Wilko, Q.	2013	Pilot Linked Open Data Nederland (NL)	B	Source: http://www.pilod.nl/doc/boek2.pdf
Hart, G., & Dolbear, C.	2013	Linked Data: a Geographic Perspective	B	Source: Boca Raton: CRC Press, Taylor & Francis Group
Heath, T., Hausenblas, M., Bizer, C., Cyganiak, R., & Hartig, O.	2008	How to Publish Linked Data on the Web	W	Source: http://events.linkeddata.org/iswc2008tutorial/
Jentzsch, A.	2011	LOD Cloud Diagram as of September 2011	W	Source: http://en.wikipedia.org/wiki/File:LOD_Cloud_Diagram_as_of_September_2011.png
W3C	2013	Linking Open Data - W3C SWEO Community Project.	W	Source: http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
Folmer, E.	2013	Introductie tweede Linked Open Data Pilot	PPT	Source: http://www.geonovum.nl/sites/default/files/Presentatie%200%20ErwinFolmerv4.pdf
Farazi, F. et al.	2012	Trentino government linked open geodata: first results	PPT	Source: http://inspire.jrc.ec.europa.eu/events/conferences/inspire_2012/presentations/14.pdf
Lopez-Pellicer, F.J., Florczyk, A.J., Nogueras-Iso, J., Muro-Medrano, P.R. & Zarazaga-Soria, F.J.	2011	Linked Open Data for INSPIRE: From 3 to 5 star geospatial data	PPT	Source: http://inspire.jrc.ec.europa.eu/events/conferences/inspire_2011/presentations/170.pdf
Schade, S. & Lutz, M.	2010	Opportunities and Challenges for using Linked Data in INSPIRE	P	Source: http://publications.jrc.ec.europa.eu/repository/handle/111111111/15247
Tschirner, S., Scherp, A. & Staab, S.	2011	Semantic access to INSPIRE: How to publish and query advanced GML data	P	Source: http://ceur-ws.org/Vol-798/paper7.pdf
Vanbockryck, J. & Robbrecht, J.	2012	Start to Link, a practical POI approach	PPT	Source: http://www.poweredbyinspire.eu/documents/0503-linkeddata-robbrecht.pdf
Beckers, V. & Tirry, D.	2013	Linked Open Data: Pilot Project of NGI-Belgium	R	Source: currently internal (report waiting for approval)
De Keyzer, M., Loutas, N., Colas, C. & Goedertier, S.	2013	TM1.2. Introduction to Linked Data (en)	PPT	Source: https://joinup.ec.europa.eu/community/ods/document/tm12-introduction-linked-data-en
De Keyzer, M., Loutas, N. & Goedertier, S.	2013	TM1.3. Introduction to RDF & SPARQL (en)	PPT	Source: https://joinup.ec.europa.eu/community/ods/document/tm13-introduction-rdf-sparql-en
Loutas, N., De Keyzer, M. & Goedertier, S.	2013	TM2.3. Design & Manage Persistent URIs (en)	PPT	Source: https://joinup.ec.europa.eu/community/ods/document/tm23-design-manage-persistent-uris-en

[13] Type: Report (R), Paper (P), Position Paper (PP), Book (B), Book chapter (BC), Presentation (PPT), Standard (S), Web resource (W)

Author(s)	Date	Title	Type[13]	Description
OGC and W3C	2014	Linking Geospatial Data Workshop - 5th - 6th March 2014, Campus London, Shoreditch	W	Source : http://www.w3.org/2014/03/lgd/
van den Brink, L., Janssen, P., Quak, W.	2013	From geodata to linked data: Automated Transformation from GML to RDF	BC	Source: http://www.pilod.nl/wiki/Boek/BrinkEtAl-GML2RDF
Archer, P., Loutas, N. & Goedertier, S.	2013	Cookbook for translating Data Models to RDF Schemas	R	Source: https://joinup.ec.europa.eu/community/semic/document/cookbook-translating-data-models-rdf-schemas
Colas, C., Goedertier, S., Kourtidis, S, Loutas, N. & Rubino, F.	2013	Core Location Pilot - Interconnecting Belgian Address Data	R	Source: https://joinup.ec.europa.eu/asset/core_location/document/core-location-pilot-interconnecting-belgian-address-data
Athanasiou, S. et al.	2013	Deliverable 2.2.1 Integration of External Geospatial Databases	R	Source: http://svn.aksw.org/projects/GeoKnow/Public/D2.2.1_Integration_of_Geospatial_Databases.pdf
Williams, H. et al.	2013	Deliverable 2.3.1 Prototype of built in Geospatial Capabilities	R	Source: http://svn.aksw.org/projects/GeoKnow/Public/D2.3.1_Prototype_of_Built-in_Geospatial_Capabilities.pdf
Ngonga, A., Sherif, M. & Hassan, M.	2013	Deliverable 3.1.1 Development of First Prototype for Spatially Interlinking Data Sets	R	Source: http://svn.aksw.org/projects/GeoKnow/Public/D3.1.1.Development_of_First_Prototype_for_Spatially.pdf
Wauer, M., Both, A., Stadtler, C. & Isele, R.	2013	Deliverable 6.1.2 Report on Customer Data Preparation and Transformation for Linked Data Usage	R	Source: http://svn.aksw.org/projects/GeoKnow/Public/D6.1.2_Customer_data_preparation.pdf
GeoKnow	2013	Task 2.7: Exposing INSPIRE data as Linked Data.	R	Source: http://geoknow.eu/t2-7.html
OGC	2012	GeoSPARQL - A Geographic Query Language for RDF Data	S	Source: http://www.opengeospatial.org/standards/geosparql
Abbas, S. & Ojo, A.	2013	Towards a Linked Geospatial Data Infrastructure.	P	Source: EGOVIS/EDEM 2013: 196-210. DOI: 10.1007/978-3-642-40160-2_16
Kerry Taylor, K., Lefort, L., Squire, G., Walker, G., Woolf, A., Shu, Y., Ratcliffe, D., Cox, S., Haller, A.	2014	Developing Ontologies for Linked Geospatial Data	P	http://www.w3.org/2014/03/lgd/papers/lgd14_submission_41.pdf
Tsinaraki, C., Stavrakantonakis, I. & Christodoulakis, S.	2007	XS2OWL: Representation of XML Schemas in OWL syntax	W	http://www.music.tuc.gr/projects/sw/xs2owl/
Tsinaraki, C. & Christodoulakis, S.	2007	XS2OWL: A Formal Model and a System for Enabling XML Schema Applications to Interoperate with OWL-DL Domain Knowledge and Semantic Web Tools	B	http://link.springer.com/chapter/10.1007%2F978-3-540-77088-6_12
Hyland, B., Atemezing, G. & Villazón-Terrazas, B.	2014	Best Practices for Publishing Linked Data	W	http://www.w3.org/TR/ld-bp/

Author(s)	Date	Title	Type[13]	Description
Berrueta, D. & Phipps, J.	2008	Best Practice Recipes for Publishing RDF Vocabularies. W3C Working Group Note	W	http://www.w3.org/TR/swbp-vocab-pub/
InGeoClouds	2013	D2.2: Interface of Web Services and Models of Data (& Annex)	R	http://www.ingeoclouds.eu
Portele, C. et al.	2013	INSPIRE Generic Conceptual Model, Version 3.4rc3. INSPIRE Drafting Team "Data Specifications".	R	http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/D2.5_v3.4rc3.pdf
ISA Programme – Action 1.1	2011	Deliverable 3.1 - Process and Methodology for Developing Core Vocabularies	R	https://joinup.ec.europa.eu/node/43160
W3C	2013	A JSON-based Serialization for Linked Data. W3C Candidate Recommendation	S	http://www.w3.org/TR/json-ld-syntax/
W3C	2013	Linked Data Platform 1.0. W3C Last Call Working Draft	W	http://www.w3.org/TR/ldp/
Janowicz, K., Pehle, T., Hart, G. & Maué P.	2010	Proceedings of the Workshop on Linked Spatiotemporal Data 2010. In conjunction with the 6th International Conference on Geographic Information Science (GIScience 2010)	B	http://ceur-ws.org/Vol-691/
Heath, T. & Bize, C.	2011	Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology	B	Morgan & Claypool - http://linkeddatabook.com/book
Ibáñez de Elejalde, P.	2013	Comparative Assessment of Linked Data and SDI Concepts in Terms of Data Access and Querying	R	Erasmus Dissertation, KU Leuven
OGC (Ed. Arctur, D.)	2011	Summary of the OGC Web Services, Phase 8 (OWS-8) Interoperability Testbed	R	OGC Document 11-139r2
van den Brink, L., Janssen, P. & Quak, W.	2014	Linking spatial data: automated conversion of geo-information models and GML data to RDF	P	Article under Review for the International Journal of Spatial Data Infrastructures Research (IJSDIR), submitted 2014-02-11
Archer, P., Dekkers, M., Goedertier, S., & Loutas, N.	2013	Study on business models for Linked Open Data (BM4LOGD)	R	Source: https://joinup.ec.europa.eu/sites/default/files/Study_on_business_models_for_Linked_Open_Government_Data_BM4LOGD_v1.00_2.pdf
Overbeek, H. & van den Brink, L.	2013	Towards a national URI Strategy for Linked Data of the Dutch public sector	R	Source: http://www.pilod.nl/images/a/aa/D1-2013-09-19_Towards_a_NL_URI_Strategy.pdf
Overbeek, H. & Brentjes, T.	2013	Concept URI Strategy for the NL Public Sector	PPT	Source: http://www.geonovum.nl/sites/default/files/2013-03-12_uri-strat.pdf
Portele, C.	2013	URI strategy of INSPIRE	PPT	Source: http://www.geonovum.nl/sites/default/files/4clemensportele.pdf
Vanbockryck, J. & Robbrecht, J.	2012	Concepts of Meta-SDI	PPT	Source: http://www.poweredbyinspire.eu/documents/0403-linkeddata-robbrechtvanbockryck.pdf

Author(s)	Date	Title	Type[13]	Description
Berendt, B.	2014	USEWOD 2014: Building a Web Observatory for research on LOD usage	W	Source: http://people.cs.kuleuven.be/~bettina.berendt/USEWOD2014/
DDGI	2013	Open Government Data – Verwaltungsdaten frei für Wirtschaft und Gesellschaft	PP	Source: http://www.google.be/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCsQFjAA&url=http%3A%2F%2Fwww.ddgi.de%2Fdownloads-topmenu-8%2Fdoc_download%2F1066-positionspapier-open-government-data-verwaltungsdaten-frei-fuer-wirtschaft-und-gesellschaft&ei=Rl3vUvnAOMijhgeLx4D4Dg&usg=AFQjCNHDyPud3AJxKZTzBI7ovAJ3z3lzWg
Geonovum	2013	Projectplan Linked Open Data	R	Source: http://www.geonovum.nl/sites/default/files/projectplan_linked_open_data_2012-2013.pdf
Schade, S. & Smits, P.	2012	Why Linked Data should not lead to next generation SDI	P	Source: IEEE International on Geoscience and Remote Sensing Symposium (IGARSS), pp. 2894-2897
Archer, P., Goedertier, S. & Loutas, N.	2012	10 Rules for Persistent URIs	R	Source: https://joinup.ec.europa.eu/community/semic/document/10-rules-persistent-uris
W3C	2012	223 Best Practices URI Construction. W3C GLD WG wiki.	W	Source: http://www.w3.org/2011/gld/wiki/223_Best_Practices_URI_Construction
Davidson, P., Murray, K. & Williams, S.	2011	Designing URI Sets for Location. Version 1.0.	R	Source: http://data.gov.uk/sites/default/files/Designing_URI_Sets_for_Location-V1.0_10.pdf
Davidson, P.	2010	Designing URI sets for the UK public sector. Version 1.0. Interim paper.	P	Source: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/60975/designing-URI-sets-uk-public-sector.pdf
UK CTOC	2014	URI patterns for identifiers proposal	W	http://standards.data.gov.uk/proposal/uri-patterns-identifiers-proposal
Abbas, S. & Ojo, A.	2014	Applying Design Patterns in URI Strategies - Naming in Linked Geospatial Data Infrastructure	P	Source: https://www.deri.ie/content/applying-design-patterns-uri-strategies-naming-linked-geospatial-data-infrastructure
Sauermann, L. & Cyganiak, R.	2008	Cool URIs for the Semantic Web. W3C Interest Group Note	W	Source : http://www.w3.org/TR/cooluris/
	2014	Digital Object Identifier System (DOI)	W	Source: http://www.doi.org/
	2014	The Handle System	W	Source : http://www.handle.net/
	2014	EZID (Easy ID) – Long Term Identifiers Made Easy	W	http://n2t.net/ezid/
EC, DG JRC		Implementation of Identifiers using URIs in INSPIRE – Frequently Asked Questions.	W	http://inspire.ec.europa.eu/index.cfm/pageid/5120
PURL	2013	Persistent Uniform Resource Locators	W	http://purl.org/
KE – Knowledge Exchange	2013	Persistent Identifiers Studies	W	http://www.knowledge-exchange.info/Default.aspx?ID=332
EC-JRC	2013	INSPIRE Registry workshop	W	https://ies-svn.jrc.ec.europa.eu/projects/inspire-registry/documents

Author(s)	Date	Title	Type[13]	Description
Koistinen, K. & Tiainen, E.	2014	URI design for semantics with ontology approach	PPT	

3.1.2. Interviews

Complementary to the desktop research, information was collect through contacts with Linked Data experts. Several interviews took place in the context of one of the pilot projects in Belgium (between November and December 2013), while other contacts took place during events such as the Joint W3C/OGC Workshop on Linking Geospatial Data (London, 5-6 March 2014)[14], providing useful input for the State of Play.

People from whom information gathered were, among others:

- Paul Hermans – Independent consultant for the government who was involved in different Linked Data projects in Belgium and the Netherlands.
- Jef Vanbockryck – Senior consultant at inno.com and kZen Labs, who worked on the project of Meta-SDI for AGIV.
- Marc Portier – Project manager at ICT Westtour, where they did an internal Linked Data project.
- Noël Van Herreweghe – Advisor-Informaticus at CORVE (Flemish eGovernment Coordination Unit) of the Flemish Government.
- Raf Buyle - V-ICT-OR (Vlaamse ICT organisatie). V-ICT-OR was part of the consortium that made the OSLO standard.
- Glen Hart & John Goodwin – Head of Research and Principal Scientist at Ordinance Survey (UK). Glen Hart was co-author of the book Linked Data: A Geographic Perspective.
- Erwin Folmer – GeoNovum (NL), project leader of the Linked Open Data Pilot in the Netherlands.
- Stijn Goedertier – Consultant at PWC, working on European Projects such as the ISA program and specialized in Core Vocabulary.
- Linda van den Brink – Geonovum - Involved in the first and second pilot on Linked Data in the Netherlands
- Clemens Portele – Active Instrucments – Chair INSPIRE DT Data Specifications, GML expert and involved in several Linked Data pilots

3.2. Method for RDF State-of-play

Several aspects should be considered when dealing with transforming INSPIRE data and data models to Linked Data (Schade and Lutz, 2010).

1. The INSPIRE data specifications and the method to describe them.
2. The different INSPIRE registers such as the INSPIRE Feature Concept Dictionary (FCD) which includes natural language definitions of the concepts underlying the feature types defined in the INSPIRE data specifications.
3. INSPIRE-conformant metadata for the INSPIRE data sets.
4. The use of multilingual thesauri, such as GEMET.
5. The INSPIRE network services, which enable the access to data and metadata of INSPIRE.

Several national and European projects have or are currently studying the transformation of INSPIRE data models and data sets into RDF. This is done in different ways, with often different starting points (e.g. from existing data sets to RDF, from GML to RDF) and each of those projects have a different focus (e.g. leaving out the metadata).

[14] http://www.w3.org/2014/03/lgd/

Currently, GML is the default encoding for INSPIRE data specifications. The decision of using a GML or RDF representation of geospatial data depends on the intended use. Both are optimized for different purposes. While RDF allows for sophisticated querying and reasoning, numerous GIS clients process GML (Schade and Lutz, 2010). However, transforming INSPIRE data models into RDF is not straightforward (Tshirner et al. 2011). INSPIRE data models are relatively complex. The models include many features with many dependent elements which, in the GML implementation, result in a heavily nested, verbose XML-tree structure. That is why transforming GML to RDF vocabularies is not straight-forward. The target OWL-models in which to transform INSPIRE data must be well-thought-out (Tshirner et al. 2011). Also, with Linked Data, the distinction between data and metadata becomes less pronounced (Schade and Lutz, 2010).

Tschirner et al. (2011) propose a general approach for deriving INSPIRE ontologies from the INSPIRE UML/GML data models in order to define the target models which then can be queried using SPARQL. They specified a way to translate SPARQL to the WFS-query language[15] "OGC Filter Encoding (FE)" and tackled the prerequisite of references between the INSPIRE ontology concepts and the INSPIRE GML data structure. Also in The Netherlands, different approaches have been followed in several Pilot Projects (Folmer, 2013). Van den Brink et al. (2013) describe a method for transforming structured GML to RDFS/OWL automatically, using XSLT. They argue that the GML's object-property structure translates very well to triples, and that therefore the transformation to RDF is straightforward. Well-known GML content elements such as names and descriptions are mapped to their RDF equivalent. However, any semantics specific to the input GML data are ignored in this translation. They also studied and tested how more meaningful RDF can be created from GML, given the underlying information model, by transforming it from UML to RDF/OWL. They argue that this mapping is also straightforward but that there are still important differences between both worlds: the re-use of existing concepts in vocabularies takes a central role in RDF/OWL while in UML the re-use of vocabularies is not directly supported. They describe how annotating the UML model could improve this translation. The figure below represents the workflow for creating and publishing data as Linked Data.

Besides the work in The Netherlands, important work has been done in Germany and the UK, as well as in Finland, Belgium (National Mapping Agency), Spain (University of Zaragoza) and Italy (Trentino Region). Important work has been done in the context of the OGC Interoperability Experiment OWS-8 and is ongoing in the context of the GeoKnow project as well[16]. The State of Play will analyse these experiences in more detail and define some guidelines for testing and comparing the different methods applied.

Based on the discussion above we need to consider different aspects with regard to the transformation of spatial data and spatial data models to RDF. Key elements to be analysed in the state of play are: the transformation process and transformation rules, the use of ontologies, the elements of conversion (spatial data, metadata, service …), the geometries involved and the tools used. Besides an abstract summarizing the projects, we therefore we analyse the cases in the next chapter according the following scheme:

Transformation: the transformation flow (UML to RDF, GML to RDF, data to RDF) and transformation rules

Elements considered: which elements were considered in the transformation (data and data models, metadata, services, codelists …)

Data sets and geometry: which type of data were tested and what type of geometry

Vocabularies and ontologies used

[15] Filter Encoding Standard (FE) defines an XML encoding for expressing filters for spatial queries in order to select a subset of features based upon specific attributes

[16] Task 2.7: Exposing INSPIRE data as Linked Data - http://geoknow.eu/t2-7.html

Tools used for transforming

We also will describe the (open) issues that those projects revealed and that should be taken into account when carrying out the experiments.

3.3. Method for PIDs State-of-Play

Why identifiers?

The web contains many diverse information, information that often changes, which makes it difficult to publish and share all the sources of information in an accurate meaningful way. This challenge required the creation of special mechanism that correctly identifies the data sets and part thereof. This is the basis for the concept of Linked Data.

In the context of INSPIRE, Linked Data can be another way of making the spatial data sets (and other components, see below) more visible and connectable.

One of the key challenges when publishing (spatial) data on the web is a mechanism for creating identifiers, not only for the spatial data sets, but also for the spatial objects they contain. Also other components of INSPIRE, such as metadata, services and code lists need a mechanism for Persistent Identifiers.

What is a persistent identifier?

When publishing (spatial) data on the web, independently of the time when the publishing takes place, it is important that the identifiers refer unambiguously to the *same resources* over time. A persistent identifier is s a long-lasting globally unique reference to a digital resource.

Is there only one kind of Persistent identifier?

There are several technologies available to realize persistent identifiers and each has its own set of advantages and disadvantages.

A. HTTP URIs as persistent identifiers

The more commonly used URI (Universal Resource Identifiers) is the Uniform Resource Locator (URL), which refers to the subset of URIs that, in addition to identifying a web resource, specifies the means of acting upon or obtaining the representation, specifying both its primary access mechanism and network location. This type of URI is promoted by the World Wide Web Consortium that they rely on the HTTP protocol for retrieving the associated information. In order to ensure persistence, HTTP servers can alias URLs to redirect to other URLs. "Moreover, work on the Semantic Web (see for example http://en.wikipedia.org/wiki/Semantic_Web) has dispensed with the notion that URLs can only be used to address online resources: the Semantic Web uses URLs to refer to off-line abstractions as well. For that reason, URLs are now generalized to URIs: Universal Resource Identifiers, rather than mere Locators." (Barnes)

B. HTTP URIs containing persistent identifiers

HTTP URI as persistent preserves the advantage of using the HTTP protocol while making the responsibility of persistence clearer to users. In this category we can have Persistent Uniform Resource Locator (PURL) and ARK (Archival Resource Key[17]) which has been used at the California Digital Library.

The Persistent Uniform Resource Locator (PURL) is a kind of URI that act as permanent identifiers in the face of a dynamic and changing Web infrastructure. "Instead of resolving directly to Web resources, PURLs provide a level of indirection that allows the underlying Web addresses of resources to change over time without negatively affecting systems that depend on them. This capability provides continuity

[17] http://tools.ietf.org/html/draft-kunze-ark-15

of references to network resources that may migrate from machine to machine for business, social or technical reasons."[18]

The ARK (Archival Resource Key) is a HTTP URI (e.g. 'http://example.org/ark:/12025/654xz321) that embeds a persistent identifier (e.g. ark:/12025/654xz321).

C. Non HTTP URIs

Non HTTP URIs, are the URIs are the identifiers schemes separated from the HTTP protocol.

URN (Uniform Resource Name) identifies a resource by name in a particular namespace. The International Standard Book Number (ISBN) system for uniquely identifying books provides a typical example of the use of URNs.

Another type of persistent identifier is the Digital Object Identifier (DOI), managed and maintained by the DOI Foundation, and are often used to refer to published articles, but they cost money.

The current PID State-of-play is not focused on any of the above technology, however it should be noted that most work done and mentioned by the stakeholders in the interviews is around URI, and in particular HTTP URIs that meet the persistence requirement.

Why PID Governance?

A clear-cut PID strategy formulated in consultation with the stakeholders, independent of the technology available to realize persistent identifiers, must ensure that the parties that wish to set to work with Linked Data can make the sound choices that are needed to generate Linked Data-solutions (Overbeek & van den Brink, 2013).

Ideally, the PID-strategy, oriented towards the technical implementation of the identification of data, should be embedded in a broader Linked Data Strategy, in which organizational aspects are considered as well. The PID-strategy is mainly intended for data that is used to define objects or concepts, to which other applications can refer (Overbeek & van den Brink, 2013).

The UK Chief Technology Officer Council defined URI sets for location with guidance on how to define URI's for Spatial Things, INSPIRE Spatial Objects, Reference Data and for derived Classes and Property Definitions (Davidson et al., 2011).

The Dutch URI-strategy, developed by Geonovum, supports the reuse of concepts and reference objects by other data collections with particular attention for the terms in the models and the reference data. Geonovum stresses the importance of registers in this context: i.e. a specification of terms/concepts in a standard or an authentic registration of reference objects. They state: "no register, no identifier". The ARE3NA-funded work on the Re3gistry software [19]and INSPIRE Registry instance as a source of INSPIRE vocabularies and will be taken into account in the state of play activities. Another paradigm which the Dutch strategy is relying upon is that it is not mandatory to have one single shared internet domain, which is different from the UK strategy (although that strategy is currently changing). The British URI-strategy assumes that all of the Linked Data URIs of the government resides under a single main domain: 'data.gov.uk'. The Dutch strategy also defines different URI patterns and ends with open issues that need to be resolved: Uniform Resource Names (URNs) vs. URIs; recognizable Internet domain; and the degree to which the strategy must be formalized.

In order to find workable solutions for PIDs for the geospatial domain, it is important to analyse the initiatives and activities in the broader ICT context. For example, the Digital Object Identifier [20](DOI)

[18] http://purl.oclc.org/docs/index.html

[19] https://joinup.ec.europa.eu/asset/re3gistry/asset_release/re3gistry-01

[20] http://www.doi.org/

system provides a technical and social infrastructure for the registration and use of persistent interoperable identifiers for use throughout digital networks. The DOI system implements the Handle System [21]and the indecs Framework. The DOI Foundation is the governance and management body for the federation of Registration Agencies providing DOI services and registration, and is the registration authority for the ISO standard (ISO 26324) for the DOI system.

From the discussion above, it emerges that a possible use of PIDs in INSPIRE involves considerations beyond the governance mechanism allowing for decision-making. Day-to-day operations and the overall PID architecture are two additional important elements in the creation and maintenance of PIDs. Of course, the investment that is required to put this in place and to ensure its sustainability should be thoroughly analysed. All these elements can be grouped together in a single holistic conceptual framework composed of four areas:

- Governance: Elements required for controlling and steering the decisions on PIDs;
- Operations: Processes and tools needed to run PIDs;
- Financing: Resources needed for the operations and the architectural updates;
- Architecture: Formal specifications around PIDs.

This framework will be used for identifying the current state-of-play with regards to PIDs.

Figure 3 Method State-of-play PIDs

[21] http://www.handle.net/

4. REPRESENTING SPATIAL DATA AS LINKED DATA USING RDF

This chapter will explores the current landscape with regard to the transformation of INSPIRE data and data models to RDF.

4.1. INSPIRE and Linked data

INSPIRE is addressing the interoperability of geospatial data sets and services through harmonised data models and encodings for the exchange of data related to one of the 34 spatial data themes listed in the Annexes to the INSPIRE Directive. These data models have been developed on a conceptual level using UML. The default encoding recommended to be used in the INSPIRE Technical Guidelines is automatically generated from these UML data models based on explicitly defined encoding rules. The current default encoding for most INSPIRE themes is based on the Geography Markup Language (GML).

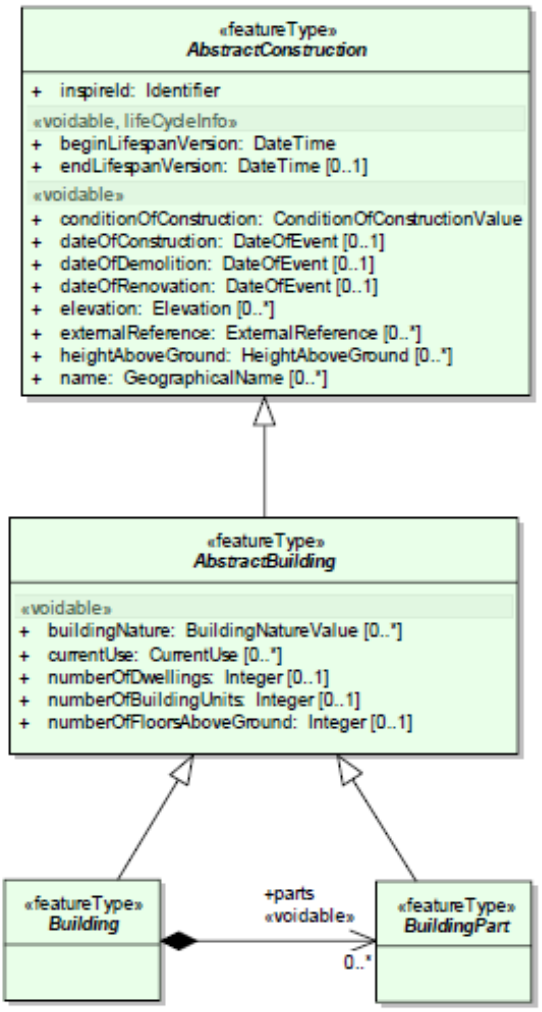


Figure 4: Feature type of Buildings Base Application Schema

Several European projects and national initiatives are publishing geospatial data as Linked Data using the RDF. However, no agreed rules or guidelines exist on how to create such RDF vocabularies from the UML models.

Table 4 compares the world of Linked Data and INSPIRE from different perspectives.

Table 4: Comparison of technical characteristics of spatial data in Linked Data and INSPIRE (based on Clemens Portele, 2013)

Aspect	Linked Data	INSPIRE
Schema description	✓ RDF-S / OWL are preferred; other languages are ok, too	✓ UML as specified by ISO 19109
Data encoding	✓ RDF (RDF/XML or Turtle) is the preferred encoding; other encodings are ok, too; any encoding should use an open specification or at least provide the data in a structured form GeoSPARQL specifies two RDF geometry encoding options using WKT and GML ✓ Another RDF serialisation to be mentioned is JSON-LD, which recently became a W3C Recommendation	✓ GML – derived from the UML model using the standard GML encoding rule – as default encoding; other encodings are ok, too. Over the next years a broad variety of encodings will be used, see http://inspire.ec.europa.eu/media-types
Terms and vocabularies	✓ Managed as resources, typically encoded in SKOS	✓ Managed as resources, currently encoded in GML and in the future likely using SKOS, too
Identifiers	✓ HTTP URIs for all resources The identifiers should be stable and not depend on implementation	✓ Identifiers not required for all data ✓ HTTP URIs may be used, but this is only a recommendation, not a requirement
Links	✓ Links to other data is qualified by a link type HTTP URIs are used to reference the linked resource	✓ Links to other features are qualified by a link type (property) ✓ URIs are used to reference linked resource in GML encoding ✓ Links are restricted to associations identified in application schemas ✓ Most datasets do not have links to external resources
Access to resources	✓ Using HTTP	✓ Pre-defined Dataset Download Service (Atom feed option): no access to individual resources, only datasets ✓ Pre-defined Dataset and Direct Access Download Service (WFS option): GetFeatureById query supports access to each feature using a HTTP URI
Queries on datasets	✓ Optional, but typically provided using a SPARQL endpoint, if RDF is supported as an encoding ✓ GeoSPARQL provides extensions for spatial query predicates	✓ Only supported in Direct Access Download Services (WFS)
Extensions (e.g., additional attributes)	✓ Linked data follows the open world assumption: ✓ Additional information may be attached to any resource by anyone Extensions may be part of another dataset	✓ INSPIRE follows a closed world assumption: ✓ Extensions are supported and specified in extensions to the UML schema ✓ The complete information about a feature is always part of one dataset

4.2. Methods for publishing spatial data as linked data

4.2.1. Use of ontologies

Ontology refers to the study of the nature of the world itself. The information technology community use this term for the explicit specification of a concept. Information technology and artificial intelligence consider that reality may be abstracted differently depending on the context from which "things" are perceived and, as such, recognize that multiple ontologies about the same part of reality may exist. In geographic information, ontology refers to a formal representation of phenomena of a universe of

discourse with an underlying vocabulary including definitions that make the intended meaning explicit and describe phenomena and their interrelationships. An ontology can be formalized using taxonomies, thesauri, conceptual models, logical theory.

Ontology is a fundamental notion for semantic interoperability on the Semantic Web since it defines the meaning of data and describes it in a format that machines and applications can read. As such, an application using data also has access to their inherent semantics through the ontology associated with it. Therefore, ontologies can support integration of heterogeneous data captured by different communities by relating them based on their semantic similarity. The W3C has proposed the Web Ontology Language (OWL) family of knowledge representation languages for authoring ontologies characterised by formal semantics on the Web.

Semantics is an important topic in the field of geographic information. The meaning of geographic information is essential for their discovery, sharing, integration, and use. Geographic information standards have recognized this fact in the ISO 19100 series of standards, and more specifically the standards on rules for application schemas (ISO 19109) and the methodology for feature cataloguing (ISO 19110). Basically, semantics relates phenomena and signs used to represent them (i.e. data) by the way of concepts. Typically, concepts are maintained in repositories called ontologies.

The projects studied and the experts that were interviewed used different ontologies for their data. The content of the data was of course often the major reason to choose a specific ontology. But, because of the large amount of existing ontologies, it is often difficult to choose one. In some cases it might be better to create a new ontology instead of using an existing one. When choosing an ontology, one has to investigate how the ontology is defined, whether there is any overlap and/or conflict with other ontologies, and what should be done with a suitable definitions of which you only use some of the fields?

Examples of ontologies used by the experts are described in table:

Ontology	Full name	Reference	Comments
Schema.org	Schema.org	https://schema.org/	This site provides a collection of schemas that webmasters can use to markup HTML pages in ways recognized by major search providers, and that can also be used for structured data interoperability (e.g. in JSON).
SKOS	Simple Knowledge Organization System	http://www.w3.org/2004/02/skos/	Support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading lists and taxonomies within the framework of the Semantic Web
Dublin Core	Dublin Core Metadata Initiative	http://dublincore.org/	The Dublin Core Schema is a small set of vocabulary terms that can be used to describe web resources (video, images, web pages, etc.), as well as physical resources such as books or CDs, and objects like artworks.
OSLO	Open Standards for Local Administrations	http://www.v-ict-or.be/kenniscentrum/OSLO	The standards of the Flemish's OSLO project are local extensions of the core Person, Busines, Location, and Public Service vocabularies created at European level in the context of the ISA Programme

PROV	PROV	http://www.w3.org/TR/prov-o/	The PROV Ontology (PROV-O) expresses the PROV Data Model [PROV-DM] using the OWL2 Web Ontology Language. It provides a set of classes, properties, and restrictions that can be used to represent and interchange provenance information generated in different systems and under different contexts.
VoiD	Vocabulary of Interlinked Datasets	http://semanticweb.org/wiki/VoID	VoiD is an RDF based schema to describe linked datasets.

Some of the experts made their own ontologies, often just for internal use. But most experts agree that, if possible, it is best to re-use existing ontologies. Ontologies play an important role in the transformation of spatial data models to RDF.

4.2.2. From GML to RDF, from UML to RDF and from data sets to RDF

Publishing Linked data can be done in different ways and different components of an SDI can be part of this transformation. Figure 6 provides an overview of the steps and parts of this transformation process. The XML or GML to RDF transformation of data and metadata can be done using so called RDF-izers. The linked data wrapper allows transforming data coming from OGC type of Web services or data sources exposed through an API, while the transformation from the data models is another mechanism.

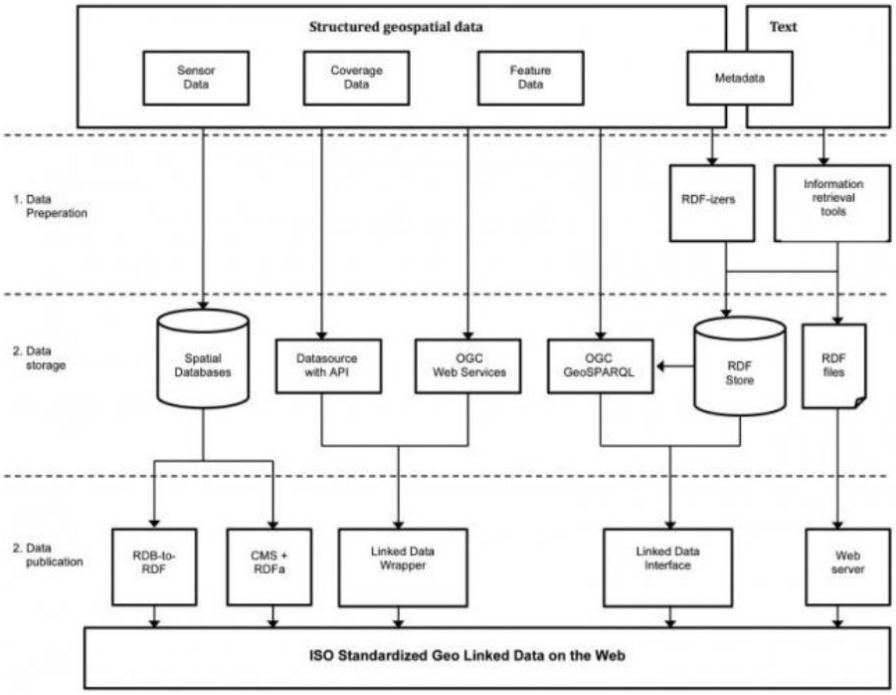


Figure 5: Scheme for publishing spatial data as Linked Data (based on Francisco et al. 2012, and Heath & Bizer, 2011)

One of the key discussions when transforming spatial data (models) is to define and apply transformation rules. This has been done in different projects, one of them being the OGC OWS-8 Interoperability Experiment (see further). Standardization efforts are under way with work of ISO/TC 211 that has set-up a

project *"ISO/DIS 19150-2 - Rules for developing ontologies in the Web Ontology Language (OWL)"*. Currently this is still a Draft International Standard. While part 1 of the standard provides an overall framework, part 2 will provide the formal representation of abstraction of the reality in OWL to support the Semantic Web. Accordingly, this part of ISO 19150 defines rules to convert UML static views of geographic information and application schemas into OWL ontologies in order to benefit and support interoperability of geographic information using the Semantic Web. These rules are required for:

- ontology description completeness;
- consistency in the set of OWL ontologies for geographic information;
- consistency in conversion of UML diagrams to OWL ontologies; and
- cohesion and unity between UML models and ontologies.

These rules are based on but also extend the OMG's Ontology Definition Metamodel. OWL ontologies are complementary to UML static views and serve different purposes.

4.3. Tools for publishing Linked Data

During the desktop study and the interviews it was analysed which software and syntax was used for working with Linked Data. The software tools for storing triples and modelling ontologies used were:

- OpenLink Virtuoso: Virtuoso is a commercial software package[22] that is often used for working on Linked Data. The problem with Virtuoso is limited integration of GeoSPARQL;
- Protégé: Protégé was often mentioned as a good choice for editing ontologies. An important pro concerning Protégé is that it is free and open source. The major problem with Protégé is that your own work is hardly recognizable after serialization;
- TopBraid Composer: TopBraid Composer is a commercial software package for developing, managing and testing configurations of ontologies and linked data was also mentioned a few times. In contrast to Protégé, in TopBradeComposer you can edit your work in text format and it is still recognizable after serialization;
- Open Refine (a.k.a. Google Refine): Sometimes Open Refine (another free and open source software package) is being used, especially when the data to start from are spreadsheets, big XML files or JSON. The pros of Open Refine that were mentioned were: - the capability of easily checking the quality of your data, - the presence of a reconciliation client, - the ability to convert to RDF. A mentioned con of Open Refine was the limited integration of the geographical aspect.

Other ones mentioned were: Strabon (which now supports GeoSPARQL), Sesame, Apache Jena, BigOWLIM, R&Wbase Parliament and Oracle. An overview of tools for Semantic Web Development can be found via the W3C webpage: http://www.w3.org/2001/sw/wiki/Tools.

None of this software tool is perfect, or the only way to go, all lack some maturity compared to their relational counterparts. The development of new software is ongoing. Until recent, only a few of the tools support the GEOSPARQL standard. In addition, regularly new tools are being developed and presented, for example, to easily convert a relational database into RDF.

Concerning syntax, a few different formats were mentioned:

- Turtle (ttl), wherefore no specific problems were mentioned;
- RDF/XML, which was not recommended by one of the experts partly because it is difficult to split up, which can be a problem for large datasets;
- Notation3 (or N3): One of the experts found it to be the easiest to make and load Linked Data and the easiest for massive files because you can split it up quite easily;

[22] A free version is available as well

- JSON-LD, a recent syntax seen as the future for Linked Data syntax, because of the already wide use of JSON. But it isn't yet strongly implemented in the market and (for now) doesn't have a geographic component. There is however a GEOJSON standard, so chances are that JSON-LD will also get a geographic component sooner or later.

4.4. Overview of different pilots and projects in Europe

4.4.1. The Netherlands

Geonovum - Pilot 1 on Linked Data

Summary:

Geonovum considers Linked Data as a complementary route for disseminating spatial information, not contradictory, but rather complementary to 'traditional' SOA-based SDI approach. This approach has created a wealth of standardized and structured GML encoded spatial data. Linked data could provide an open mechanism for sharing and combining these data with other data and information once the data is also available as linked data. It would make the existing data more visible and available for new innovative services and applications. That was the motivation for investing in an extensive pilot with many parties involved. The pilot on Linked Data in The Netherlands ran from 2012-2013. It developed two ways of transforming spatial data sets to Linked Data: one method focused on deriving RDF from GML, while the other focused on the transformation from the data models to RDFS/OWL.

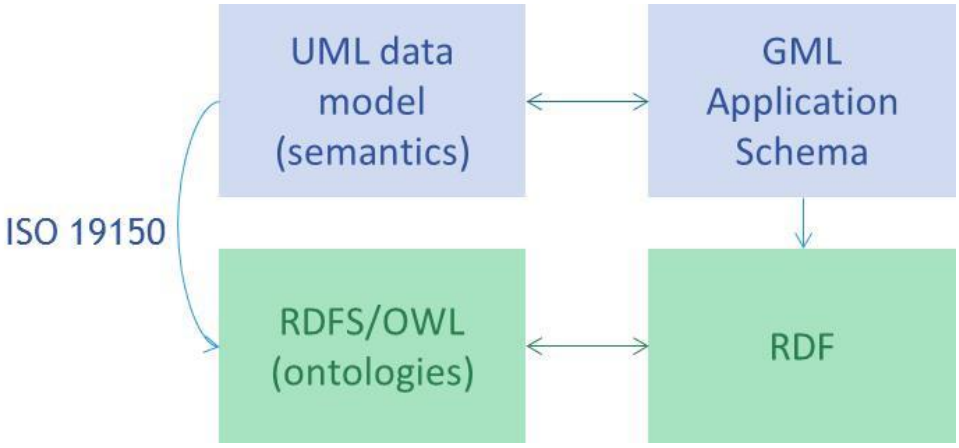


Figure 6: Schematic overview of mapping UML and GML to RDF

Transformation:

- GML to RDFS/OWL
- UML to RDFS/OWL

Data sets and geometry:

- Data set IMRO (land use plans)
- Point and Surface geometry (which can be extended) transformed to a WKT serialization conform to GeoSPARQL

Ontologies and vocabularies:

- No ontology exists for IMRO data. In the GML to RDF transformation, no ontology was used but names, descriptions, objects (including nested features) were mapped towards their RDF equivalent.
- An UML model and application schema exists for IMRO and the generation of an OWL vocabulary from this model has been done in the context of the pilot. To improve the mapping from UML to OWL, missing information is added by using tagged values in the UML model using mapping annotations

Tools used:

- GeoSPARQL with WKT serialization
- XSLT 2.0, a formal language for the transformation of XML documents (GML included). The supported output formats are HTML, XML (RDF/XML included) or plain text (TTL, N3 and JSON-LD fall in this category).
- ShapeChange – modified to allow annotation of UML models for the UML to RDF transformation

Elements:

- Data models, codelists

Issues:

- The applied approach generated an OWL representation of the frame-based, closed-world UML model (because that makes the transformation straightforward). To be able to generate an open-world oriented ontology representation of the same UML model, an open-world mapping approach is needed
- Issues regarding UML modelling conventions

4.4.2. Germany

Federal Institute of Hydrology and University of Koblenz

Summary:

A research project of the Federal Institute of Hydrology and the University of Koblenz, both in Germany, led to experiments regarding the publication and querying of GML data. The motivation for the work is the following. The Semantic Web has technologically fostered the Linked Data initiative which builds up huge repositories of freely collected data for public access. Querying both data categories within distributed searches looks promising. To tackle the associated prerequisites, firstly a general approach to translate sophisticated INSPIRE GML data models into Semantic Web OWL ontologies should be proposed. This is done according to Linked Data principles while preserving selective INSPIRE structural information (in the UML models) as annotations. Secondly, a conversion of the Semantic Web query language SPARQL to its GeoWeb counterpart "OGC Filter Encoding" has been developed.

Transformation:

- GML to RDFS/OWL
- UML to RDFS/OWL
- SPARQL conversion to OGC Filter Encoding

Data sets and geometry:

- Data set protected sites (Slovakia) and administrative units (The Netherlands)
- Point, Linear and Area geometry

Ontologies and vocabularies:

- The thematic domains are transferred to their respective domain ontology while preserving the INSPIRE element names. Basic rules for the conversion have been established.
    - Every INSPIRE UML-class except stereotype union is converted to an owl:Class. Subtypes of stereotype union are modeled each as one owl:Class;
    - Every value of codelist or enumeration is converted to an owl:Individual typed with the corresponding enumeration/codelist owl:Class
    - Every UML-attribute corresponding to a GML-property is converted to an owl:ObjectProperty or owl:DatatypeProperty. If multiple, equally-named GML-properties lead to both OWL property types we name the owl:DatatypeProperty with suffix _dataValue to be conform to OWL-DL

    - every UML-association is converted to an owl:ObjectProperty
- Ontological refinements have been introduced: classifications with codelist- and enumeration types; temporal and measure values; registries for reference data (e.g. CRS) and identifiers

Tools and standards used:

- SPARQL and GeoSPARQL
- Sesame framework: NetworkedGraphs, DistributedSail

Elements:

- Data models, codelists, CRS, WFS

Issues:

- Linking the INSPIRE RDF vocabularies with other Linked Data is not so obvious and needs to be explored
- Challenge is the development of a coordinated Semantic Web infrastructure for INSPIRE
- Work on metadata is a matter of future work

4.4.3. Belgium

Experiment transforming data from the NGI-BE

Summary:

The national mapping agency of Belgium, NGI-BE started a project in October 2013 on Linked Data. The aim of this pilot project is to get a clear view on the state of the art of Linked Data, and more specifically better understand the benefits of implementing Linked Data compared to the cost of publishing data as Linked Data. The methodology consisted of two steps. First, experts in Flanders and abroad were interviewed in order to assess current experiences regarding the implementation of linked data. A brief assessment was made of the different technologies are in use for producing Linked Data. Next, an experiment has been developed in which an existing dataset from the NGI was transformed into RDF. The objective of the experiment was to demonstrate how an operational dataset of NGI can be transformed, offering more insight in the technical steps and the software that can be applied.

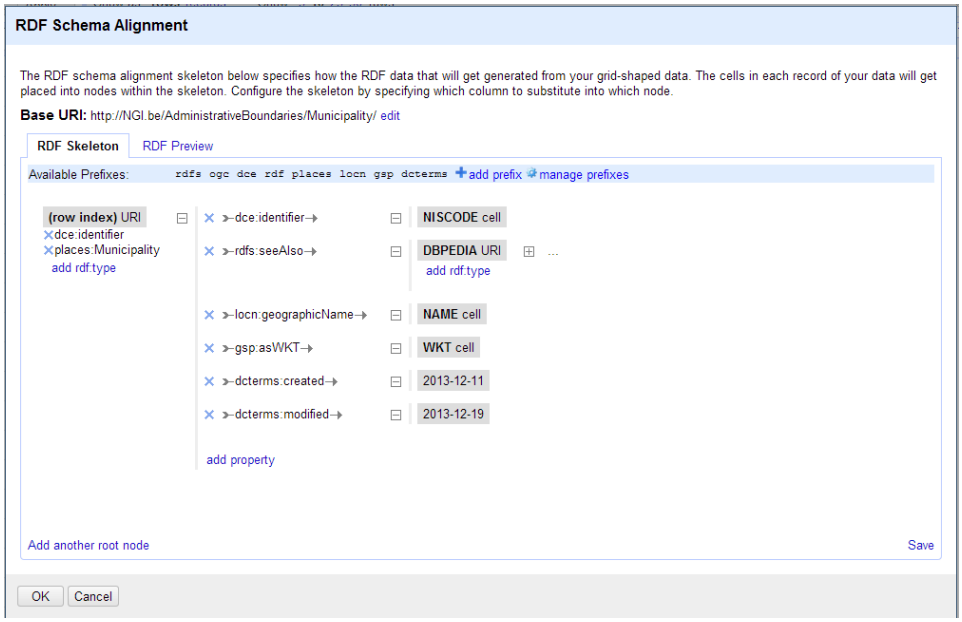


Figure 7: Configuring the base URIs (NGI administrative boundaries case)

Transformation:

- From raw data (SHAPE file) to CVS to Turtle

Data sets and geometry:

- Administrative Boundary data set
- Area geometry

Ontologies and vocabularies:

- DBPedia was used to add a reconciliation service based on a SPARQL endpoint. Other options such as Freebase and Geonames were not tested (the because at the time of the experiment Geonames was not able to find most of the municipalities of the province of Limburg)
- Dublin Core Elements Set, the Qualified Dublin Core were added as ontologies, as well as the Location Core Vocabulary and the Places Ontology from Smethurst et al.

Tools and standards used:

- GeoSPARQL used for geometry of polygons according to WKT serialization
- Open Refine with RDF Refine extension sued for conversion to CVS

Elements:

- Data geometry and attributes, no metadata

Issues:

- The representation of geometry is very extensive
- Restrictions on the use of Google Refine

4.4.4. Italy

Trentino Government – Government Linked Open geoData project

Summary:

The Government of the Trentino Region in northern Italy has initiated a project to publish certain geospatial data sets as linked open data. The project started from the observation that data in the public administration domain comes from different entities, can be produced, stored and delivered in different formats and can have different levels of quality. Hence, this heterogeneity needs to be addressed, while performing various data integration tasks. 161 core geographic datasets were released ad linked data by leveraging the geo-catalogue application within the existing geo-portal.

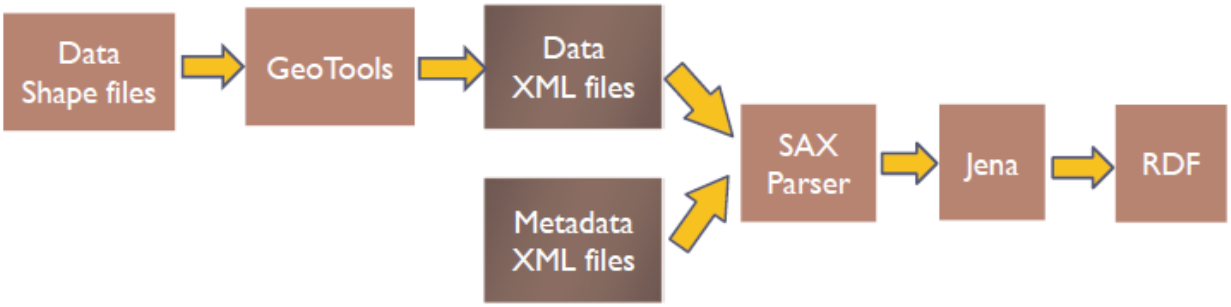


Figure 8: The transformation scheme from raw data to RDF (Trentino Region)

Transformation:

- From raw data (SHP-files) to XML to RDF

Data sets and geometry:

- 161 data sets from the regional SDI
- Point, polylines and polygon geometry

Ontologies and vocabularies:

- Transformation was performed based on a rule file providing the mapping between the XML and the RDF Objects
- Tests were performed to link the data and metadata to external vocabularies (DBPedia and Freebase)
- For metadata Dublin Core and DCMI-BOX standard vocabularies were used

Tools and standards used:

- Geotools was use to pre-process the data to produce XML
- Relevant metadata were extracted with a standard SAX Parser
- The Jena tool was used to convert to RDF

Elements:

- Data (Classes, entities and attributes) and metadata have been transformed

Issues:

- There remained an open question regarding the URIs, more specifically the patterns to be adopted. An approach for a good URI design is the next step in the process of implementing Linked Data

4.4.5. OGC

OWS-8 – Interoperability experiment: Semantic Mediation ER

Summary:

The 8th OGC OWS Interoperability Experiment addressed, among others the set-up of a Semantic Mediation approach to integrate different data and data models, and making them available through web services. Within OWS-8 both application schema and feature catalogues play a very important role with the definition of the feature type holding information for interpreting the semantics of the data. The mediation component, as a CSW client, accesses a symbology registry to generate maps based on the feature type, the symbols, and rules registered for those feature types. The mediation component also translates between instances of domain models in GML and RDF, queries a knowledge base and integrates the results in a map with proper styles. The knowledge base contains the common model (Rosetta Mediation Model, RMM), ontologies representing each data model, mappings from each data model to RMM, and rules.

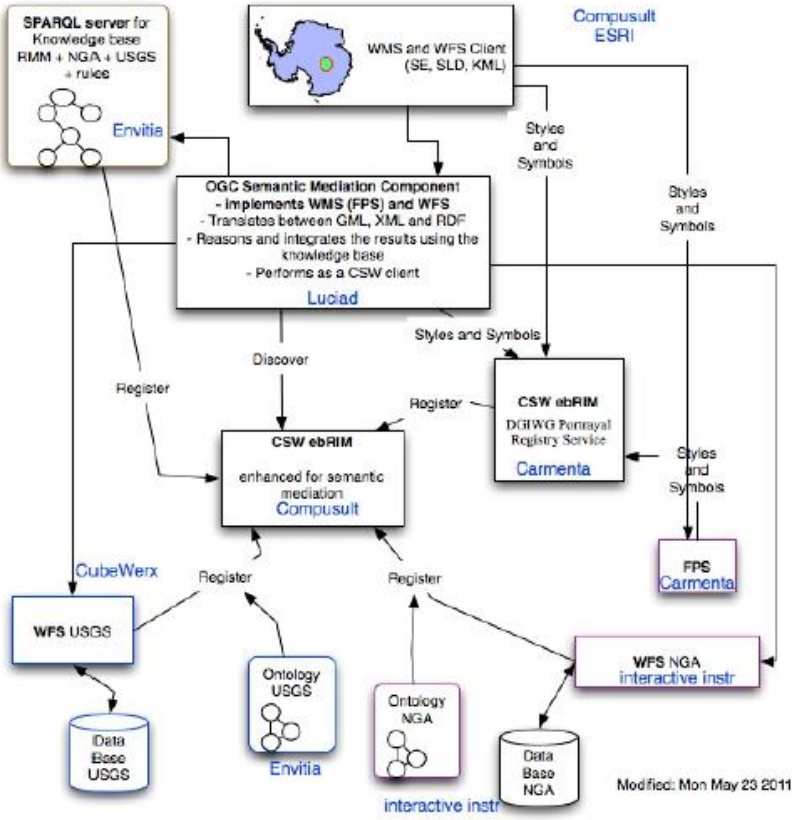


Figure 9: schema of the architecture of OWS-8

Transformation:

- From GML to RDF, From UML to RDF
- Both ontology mapping and rules based transformation applied

Data sets and geometry:

- Data sets were from the NGA Topographic Data Store and the USGS National Map
- Different formats and different geometries

Ontologies and vocabularies:

- Specific ontologies created from the respective data (UCGS and NGA)

Tools and standards used:

- GeoSPARQL
- ShapeChange

Elements:

- Data sets and models, CRS, services

Issues:

- Complex mapping might be needed. How much of the data transformation mapping can be standardized and how much should be left to implementations

4.4.6. GeoKnow

[Report not yet available, to be added in later stage]

4.5. Open issues and challenges, potential questions for the experiments

Throughout the discussions at the Joint W3C/OGC Linking Geospatail Data workshop and from the experiences in the different projects described, several issues and challenges can be defined.

From UML to OWL

There seems that there are too many semantics in UML stereotypes that are not amenable to a generic UML to OWL toolset. Some experts like Kerry Taylor recommend not performing straight translations from UML to OWL, as you will lose much of the intrinsic OWL capabilities. It also hampers the interoperability with other RDF datasets. Different opinions exist on this matter. It was empirical observed during one of the projects where a team of Kerry Taylor tried to convert existing ontologies on water and climate data in XML to OWL. For the transformation of domain models, the idea is then to start over again from scratch for the development of an OWL model. An important issue to tackle is the problem of modelling conventions and restrictions in UML which lead to awkward modelling construct, which could better be differently modelled in OWL (van den Brink).

Rules for the generation of OWL

The work in the context of ISO/TC 211 is important to this regard. Although this is ongoing work, and new NWIPs will be defined, including a project on OWL versions of several ISO standards, there is still much discussion ongoing. Simon Cox prepared a subset of the ISO Harmonized Model as OWL ontologies, available as RDF/XML[23]. This work focusses on a rule-based conversion, and there is no attempt to harmonize with any existing ontologies, so it is merely to represent the ISO UML models into OWL. The thinking is that this provides a formal conversion. Relating it to other OWL ontologies can then be done using RDFS and OWL mechanisms – subClassOf, subPropertyOf, sameAs, etc. These relationships should probably be a distinct resource. One of the questions is: *"Is it desirable to have a flat rule-based conversion, without any attempt to harmonize with existing ontologies"?* In several experiments specific rules *were* defined and integrated in toolsets (e.g. in ShapeChange.

Which standard for geometry

During the workshop in London there was a lot of discussion on existing standards and specifications to capture geometry. Although GeoSPARQL is an OGC standard, some experts have serious objections to use it and prefer other vocabularies such as the Core Location or NeoGeo for their purposes. One of the conclusions was that W3C and OGC will meet and discuss in order to come up with one standard. Most probably there will be a revision of the GeoSPARQL standard.

Persistent Identifiers

Several questions need an answer:

- What (spatial dataset, spatial object, datatypes, codelists, CRSs, ,…) needs to be identified within/outside INSPIRE (e.g. CRS outside INSPIRE, which registers….),
- For what purpose (object identifiers e.g. INSPIRE-Id versus real-world identifiers aka thematic identifiers), how to deal with multiple spatial objects that represent the same real-word phenomenon
- and how will it be managed.

[23] http://def.seegrid.csiro.au/isotc211/

5. PIDs State-of-play

In the context of this research it was considered that a possible use of PIDs in INSPIRE involves considerations beyond the governance mechanism allowing for decision-making, as mentioned in section "3.3 Method for PIDs State-of-Play". Day-to-day operations and the overall PID architecture are two additional important elements in the creation and maintenance of PIDs. Of course, the investment that is required to put this in place should be thoroughly analysed. All these elements can be grouped together in a single holistic conceptual framework composed of four areas:

- Governance: Elements required for controlling and steering the decisions on PIDs;
- Operations: Processes and tools needed to run PIDs;
- Financing: Resources needed for the operations and the architectural updates;
- Architecture: Formal specifications around PIDs.

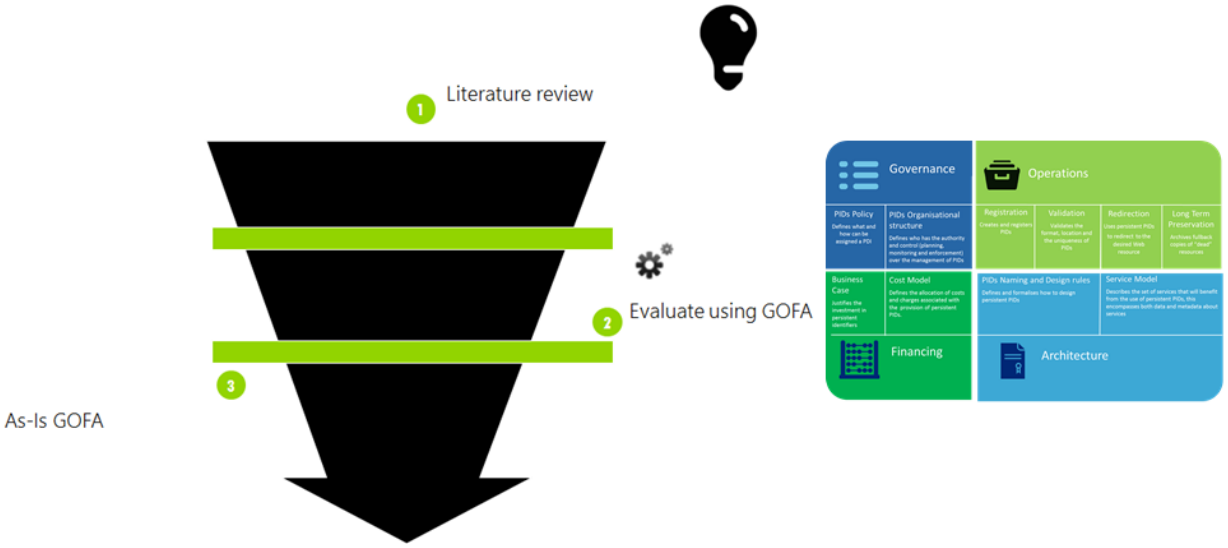


Figure 10 PID State-of-play framework

5.1. Governance

When discussing PID Governance, this study looked for the set of quality control disciplines for managing, using, improving, maintaining, monitoring, and protecting PIDs with the associated resources across the organisation.

Data governance design needs to comprise the three governance building blocks and integrate these into one suitable governance model.

- Governance principles: The principles define basic and overarching rules or directions that data steering and control rely on (e.g. in Risk Management). They form the general framework for a company's information governance
- Governance bodies: The intention of data governance bodies is to achieve effective decision making processes including relevant roles & responsibilities
- Governance processes: These processes are the means to implement and execute the data governance and tie tightly in with the information governance bodies

From the literature, with regards to the PID governance the following two aspects were identified:

- PID Policy – defines what and how can be assigned a PID;
- PIDs Organizational structure – Defines who has the authority, control and liability.

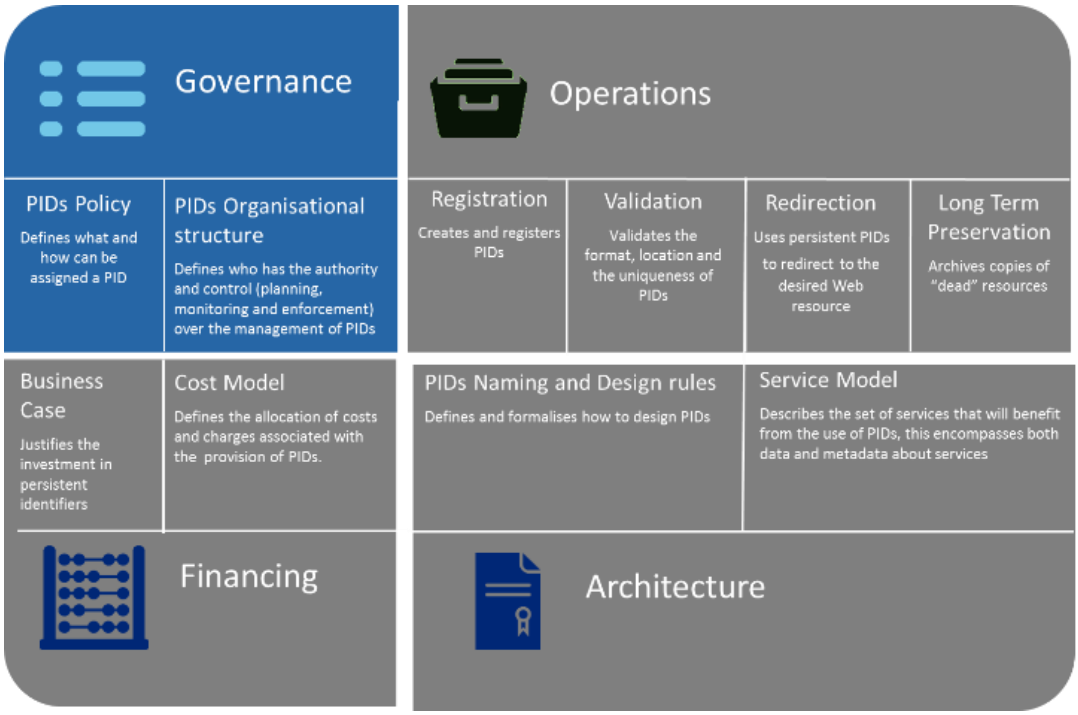


Figure 11 PID Governance - relevant literature

5.1.1. Policy on what and who is allowed to assign identifiers

The policy on what and who is allowed to assign identifiers is generally associated with PID design patterns and the management of domains.

For example UK (Public sector) mandate that Government departments and agencies should keep a list for each type of 'Thing' that they are responsible for. Additionally they should engage with the stakeholders to ensure the quality.

Policy is generally present in organizations however not always documented due to the small scale. (e.g. Austria – REEEP(Renewable Energy and Energy Efficiency Partnership). Additionally, the policy is similar however the workflow is different as is custom to the organisation.

Several guidelines have been put forward for the URI persistence (IETF - RFC2616., UK - Designing URI sets for UK Public Sector), but currently these are just recommendations. There is no EU agreed policy for stable http URIs for centrally managed, shared resources.

With regards to the resources that can be assigned an identifier, it is specific to the domain (books, spatial objects, spatial data sets etc.) and the assignment is managed by the domain owner. The resources are assigned an identifier within a domain or a sector or register. In the Strategy for Linked Data of the Dutch public sector the provision for the authentic definition and identification of concepts or reference objects as a register – "No register, no identifier".

5.1.1.1. Standardization bodies –W3C

W3C has created a collaboration page for creating URIs for use in government linked data.

When discussing persistence, it is mentioned that "*URI persistence is a matter of policy and commitment on the part of the URI owner. The choice of a particular URI scheme provides no guarantee that those URIs will be persistent or that they will not be persistent.* "

With regards to the management of URI persistence, IETF put forward RFC2616. (e.g. HTTP redirection (using the 3xx response codes) which permits servers to tell an agent that further action needs to be taken by the agent in order to fulfill the request. Content negotiation also ensures consistency as for new format there is no need to create a new URI (except for FTP).

5.1.1.2. Austria – REEEP (Renewable Energy and Energy Efficiency Partnership) Reegle.info

In the paper put forward by the European Commission "Study on business models for Linked Open Data", Florian Bauer, REEEP Operations and IT Director was interviewed. With regards to the policy he mentioned that there is no need to have a policy at this scale.

5.1.1.3. Belgium - Meta-SDI

AGIV, Jef Vanbockryck and Joeri Robbrecht discuss about Meta-SDI. Meta-SDI consists of the following aspects:

- A set of rules to manage meta-information
- A well-defined syntax for stable and long-living URI's
- Registries and repositories

Meta-SDI is presented as important to support and implement the information policy and governance

  - Availability of up-to-date, harmonised and qualitative public data inside and outside the public sector, to stimulate the use and reuse of geographical information.
  - Centralized publication of the offering
  - Cost-saving: produce once, use many times.
  - Process integration, efficiency and quality: A guaranteed offering of authentic data (addresses, parcels, buildings, roads, …) as a reference framework for data usage and data processing.
  - Harmonized data models, metadata models, code lists, identifiers, service interfaces, exchange formats, to facilitate the use of the SDI for a large and diverse user base (geo and non-geo) and to meet INSPIRE requirements.

Meta-SDI project in Flanders (Concepts of Meta-SDI) reflected that the URI scheme is an essential part of the Meta-SDI. The scheme represents the "who-what-where":

• Who (provider): All information providers are registered and have their own "orgns" (organisation namespace)

• What (resource): All kind of information is registered, categorized and associated

• Where (online access): All registered information is accessible through a standardized URL

5.1.1.4. Denmark (KE) and Netherlands (SURF)

KE (Knowledge Exchange) and SURF (Studies on Persistent Identifier Infrastructure and development of a URN-NBN based Global Resolution Service) have published several studies on Persistent Identifier Infrastructure and development of a URN-NBN based Global Resolution Service. These studies are related to other broader PIDs initiatives, such as DEFF, SURF, DANS – the national libraries of Germany, Finland and Sweden and also CNR and FDR in Italy.

For the Current state and state of the Art report, they have developed a framework for evaluation. This framework consists of the following:

- Functions – all functions from lifecycle of a persisted ID
- Policy – defines how the functions should behave
- Workflow – describes how different functions are dealt with
- Implementation – the procedures to implement the workflow (e.g. who does what).

With regards to the policy on who is allowed to assign identifiers it was concluded that organization have a policy however this is not always formalized or it is outdated, while the demand for PIDs is increasing.

The policy in all organizations is similar, the workflow are different as they are custom to the organizations they are implemented in. The same goes for the implementation which is different from country to country, mainly the format of the identifiers, metadata schemas used when registering.

Function	Comments
Becoming a registrar	Upon request (email or phone) an external can request to be a registrar. The difference relies in the type of organizations are allowed, and in general people can't be registrars.
Creating an identifier	Uniqueness normally assured by using decentralized repositories with a unique namespace so they can create unique identifiers. Use opaque identifiers minimize the use of semantics. Sweden and Finland – provide a service for the creation of unique valid and durable IDs Italy – free to create in NBN space and it is checked at registration Germany adds checksums
Assigning identifiers	All have a policy but defers in content and type Many don't know how to deal with representation Unclarity with regards to the level of identification
Registering identifiers	The goal of this function is: • To make the resolver aware of the resource • Validate the identifiers. The resolvers are based on either pull or push mechanism.
Updating locations	Keeping the location up to date is in the responsibility of the repository. The

Function	Comments
	location is firstly validated at the registration.
Updating identified objects	A new identifier to a new version of the object. Germany provides features for version management. However the rest have a more strict policy.
Transferring objects to new owners	The PIDs should not be changed even if the owner changes. Creation of a new identifier it is also not desirable. The transfer between repositories is based on: • Single records • Collections • Complete repositories. Italy – the object can be registered in the new repository and removed from the previous one. Germany – the new location is registered as e new version. Holland – the ownership is determined on an individual basis.
Removing identified objects	In most cases is a 404 page displayed. In other cases of removal, a fallback to a copy in a Long Term Preservation (LTP).
Fetching identified object	GET or POST web request that fetches the object followed by a 302 redirection. The way to submit identifiers differs : • Path to the webserver (http://example.com/urn:example) OR • Query parameter (http://example.com?identifier=urn:example)
Fetching metadata of identified objects	Most use lookup of the metadata.
LTP facilities	Usually the National Library have a legal obligation to a LTP (Long Term Preservation).

5.1.1.5. Germany – German National Library

In the paper put forward by the European Commission "Study on business models for Linked Open Data", Lars G. Svensson (Advisor for Knowledge Networking) was interviewed. With regards to the policy he mentioned that all linked data is supplied through the domain d-nb.info. URI patterns are:

- http://d-nb.info/(internal reference number) for bibliographic data
- http://d-nb.info/gnd/(authority record identifier) for authority data
- http://d-nb.info/standards/elementset/(term) for ontologies
- http://d-nb.info/standards/vocab/(term) for value vocabularies

5.1.1.6. Italy - Agenzia per l'Italia Digitale

In the paper put forward by the European Commission "Study on business models for Linked Open Data", Giorgia Lodi & Antonio Maccioni were interviewed. When discussing about policy it was mentioned that AgID is responsible for defining technical rules for the interoperability of the Italian base registries, as defined by the national master law of innovation named CAD – Codice dell'Amministrazione Digitale. The technical guidelines include a section on persistent URIs and they have contributed to and refer to the ISA Programme work.

5.1.1.7. Netherland

In the paper "URI Strategy of INSPIRE" by Hans Overbeek and Thijs Brentjens, a set of recommendations for the URI Strategy of INSPIRE were put forward;

- Need to map all identifiers in INSPIRE to http URIs;
- URIs must be independent of implementation details and should be short and mnemonic;
- Member States, the Commission and other organizations assigning identifiers need to develop URI schemes to manage assignment of http URIs;
- Typically this should be done with a wider scope than just spatial data;
- Infrastructure needs to be set up and maintained to resolve http URIs and return information resources.

Conclusions:

- Stable http URIs as identifiers for spatial objects and spatial data sets are currently recommendations;
- Additional discussions between Member States and the Commission required to develop an agreed URI strategy;
- Stable http URIs for centrally managed, shared resources are being defined.

Aims to establish a EU-wide spatial data infrastructure (SDI):

- Obligations on public sector organizations;
- To share spatial information relevant for the environment to assist policy-making across boundaries.

5.1.1.8. UK

The document **"Designing URI sets for UK Public Sector (2009)"** recommends the allocation of lead departments/agencies for sectors. Further, Government departments and agencies should keep a list for each type of 'Thing' that they are responsible for.

In addition, the paper recommends engaging the stakeholders of each sector. In this perspective, lead departments/agencies should engage with stakeholders to ensure that the set is of sufficient quality to meet a wide range of purposes. The department or agency responsible for a real-world 'Thing' should also be responsible for defining it and naming instances of it, on behalf of the appropriate sector.

At the same time governance is needed to avoid naming collisions, determining and prioritizing URI sets within each sector, accreditation of URI sets to be rooted at data.gov.uk, to ensure the longevity and infrastructure.

UK – BBC

In the paper put forward by the European Commission "Study on business models for Linked Open Data" when discussing with Dave Rogers and Oli Bartlett about policy, it was mentioned that they have an internal one but at the whim of a bigger organization and other departments might think otherwise. "*BBC has a mixed set of policies that may or may not be followed. There is no one canonical URI policy. We do recognise need for canonical URIs in LDP but basically it is bbc.co.uk/things/GUID – a flat policy that avoids ownership. Of course there's a URI for everything in the platform but we don't want to stop product teams exposing their data in their own URI system - i.e. we're very distributed. Trying to impose URI designs across the BBC wouldn't work.*"

UK – Companies House

In the paper put forward by the European Commission "Study on business models for Linked Open Data" when discussing with Mark Fairhurst, Chris Smith, Stacey Smith, it was mentioned that the policy is given by URI Structure, Domain. The recommendation is to use "business" to represent the business sector, in-

line with the examples of education and transport, and within the data.gov.uk collection of UK public sector URIs. NB – business.data.gov.uk exists.

UK – Department of Environment, Food and Rural Affairs (DEFRA)

In the paper put forward by the European Commission "Study on business models for Linked Open Data", when discussing with Alex Coley, DEFRA/Environment Agency, Chair UK Government Linked Data WG it was mentioned that "*We have a framework policy. Not an environment.data.gov subdomain policy – but is in development. Things like what to do in our namespace. We're working on it – conventions coming from what we've done.* "

UK – National Archives

In the paper put forward by the European Commission "Study on business models for Linked Open Government Data" (Commission, 2013) John Sheridan, Head of Legislation Services mentioned about their policy: "*We specify UK Gov URIs in each contract in the areas that they cover. For gazette we also require there's a URI template. IETF work at the heart of ELI and this. We're trying to eliminate the reliance on paper and persistence is at the heart of what the National Archives does of course so you can be confident that the URIs will persist as long as they're useful.*"

UK – Ordnance Survey

In the paper put forward by the European Commission "Study on business models for Linked Open Data", when discussing with John Goodwin, Principal Scientist, Ordnance Survey Research mentioned that they do have an identifier policy of which the policy on URIs forms a fragment.

UK- Public Sector Information Domain of the CTO Council's cross-Government Enterprise Architecture

A report from Public Sector Information Domain of the CTO Council's cross-Government Enterprise Architecture (The Designing URI sets for UK Public Sector, 2009), gives emerging best practices regarding implementation decisions are made to meet the specific needs of the UK public sector:

- Use of data.gov.uk as the domain to root those URI sets that are promoted for re-use;
- Organization of URI sets into 'sectors' (e.g. education, transport, health) with a lead department or agency;
- Consistent use of metadata to describe the quality characteristics of each URI set;
- Organization of URI sets into 'sectors' (e.g. education, transport, health) with a lead department or agency.

Public Sector Information Domain of the CTO Council's cross-Government Enterprise Architecture suggests that URIs should include the following components (The Designing URI sets for UK Public Sector, 2009):

- A concept: a word or string to capture the essence of the real-world 'Thing' that the set names
- A reference: a string that is used by the set publisher to identify an individual instance of concept. The reference should match the way that it is used in normal use.

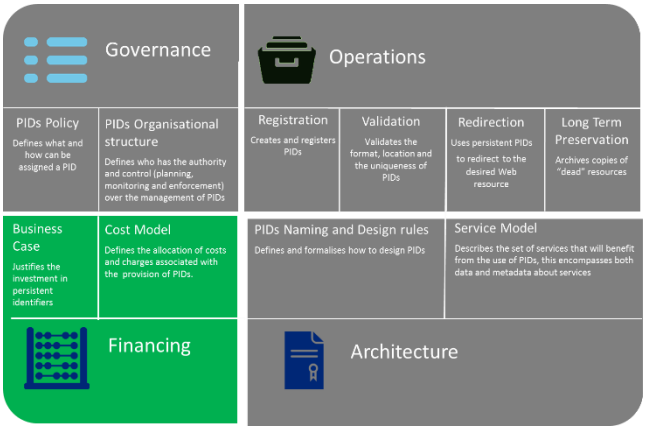
In addition, the UK INSPIRE Compliance Board ensures that implementation of INSPIRE is aligned with wider Government information and data policy. The Architecture and Interoperability Board is required to oversee and drive the implementation of the UK location interoperability standards and practice guidelines, in conjunction with the INSPIRE Regulations (United Kingdom INSPIRE Member State Report 2013).

5.1.2. PID Organisational structure

Based on the collected data, there is no clear organizational structure for the management of PIDs. It can only be implied that the domain owner manages the PIDs, however the organizational structure behind has not been described.

5.1.2.1. Netherlands

In the paper "Towards a national URI-Strategy for Linked Data of the Dutch public sector[24]" (Brink), when discussing about persistence, it was mentioned that persistence should be independent of the organizational changes. A company or authority must feel confident to develop critical operational systems on its basis.



5.2. Financing

Financing the PID Governance implies both the cost model and the business case. These two aspects are important factors that support the decision on whether an organisation should have PID Governance or not.

The cost model is about how much the PID Governance would cost, while the Business case is about the benefits of having PID Governance.

5.2.1. Business Case

The business case is generally not documents as there is a high interest in Linked Data in general. As John Sheridan, Head of Legislation Services would say "A business case for using linked data would like making a business case for using electricity."

5.2.1.1. UK

In the paper put forward by the European Commission "Study on business models for Linked Open Data", when discussing with UK – Department of Environment, Food and Rural Affairs (DEFRA) it was mentioned about the business case is about making decisions based on evidence "*We're about environmental outcomes that we want to achieve – decisions based on evidence – we want other people to base their decisions on evidence too so we make the data available for that.*" In the same paper UK National Archive representative mentioned that "*For legislation, we don't have a business case doc, but the essence of approach to managing legislation as data draws deeply on LD principles so there wasn't a business case to write. A business case for using linked data would like making a business case for using electricity.*"

5.2.2. Cost Model

The cost model is occasionally document and it is different from one case to another. Generally speaking, the cost can always be categorized as:

- Development costs;

[24] http://www.pilod.nl/w/images/a/aa/D1-2013-09-19_Towards_a_NL_URI_Strategy.pdf

- Maintenance costs;
- Promotion costs;
- Other costs.

5.2.2.1. EU – European Commission

In the paper put forward by the European Commission "Study on business models for Linked Open Data" when evaluation the cost, the following three categories were taken into account:

- Development costs
- Maintenance costs
- Promotion costs

5.2.2.2. Austria – REEEP

In the paper put forward by the European Commission "Study on business models for Linked Open Data", Florian Bauer, REEEP Operations and IT Director was interviewed. When discussing about costs he mentioned: "*The problem is separating out the relevant costs. Reegle costs a few hundred K per year and over the past 10 years has probably cost several million all told. The linked data aspects have come to less than €1M.*"

5.2.2.3. Belgium

For Flanders, the IT Infrastructure (Hardware and core software components) costs are dived in set-up costs and maintenance (yearly costs) (Member State Report: Belgium, 2010-2012). The report considers the individual cost of the following aspects:

- Set-up costs (one-off costs)
    - Software (adapting software, creating new software, setting catalogues)
    - Production
        - Creation of metadata for discovery
        - Creation of metadata for evaluation and use (new metadata elements required by Data
        - Specifications Implementing Rules)
        - Testing for compliance
        - Participation of national experts into INSPIRE development process
- Maintenance (recurrent yearly costs)
    - Software (adapting software, creating new software, setting catalogues)
    - Production
        - Maintenance of metadata for discovery MD
        - Maintenance of MD for evaluation and use
        - Testing for compliance

5.2.2.4. Cyprus

In Cyprus INSPIRE Member State Report 2013, the costs required for the development of the National Infrastructure for Geospatial Information are divided into three categories:

- Development cost: The cost relates to the development of infrastructure and web services and applications as well as to the preparation of specifications and rules on common use and sharing of the system and the data.
- Cost of production of basic geospatial reference data: The cost relates to the homogenization of existing geospatial data in order to produce the background to be

used for the presentation and dissemination of geospatial data through the information system.

- Compliance costs for data producers: The cost relates to the transformation of geospatial data and the development and maintenance of appropriate services, from PAs which produce geospatial data and are obliged to participate.

5.2.2.5. Germany

In the paper put forward by the European Commission "Study on business models for Linked Open Data", Lars G. Svensson (Advisor for Knowledge Networking) was interviewed. When discussing about costs he mentioned: "*Development cost is estimated at approximately 221 person days.*

Up to June 2012, the service was a project, from July 2012 and onwards, it is a product. Costs are considered to be part of the general bibliographic services and general product maintenance."

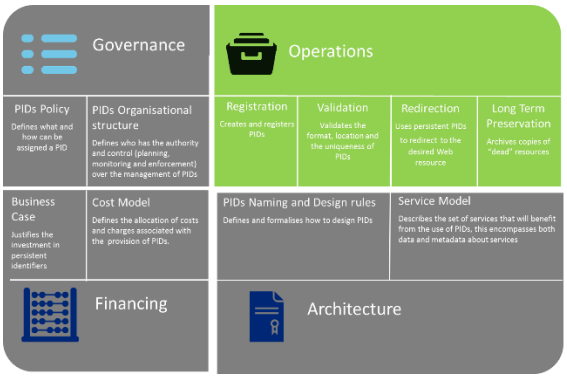
The Federal authorities in Germany split the costs related to as follow (Germany INSPIRE Member State Report 2013):

- Operating costs of the IT infrastructure
- Production of interoperability
- Business model for the implementation of INSPIRE
- IT structure
- Processing of spatial data
- Personnel

5.2.2.6. UK

The report from the UK to the European Commission in accordance with Article 21 of the INSPIRE Directive 2007/2/EC (United Kingdom INSPIRE Member State Report 2013) considers the costs of the UK Location Programme for the past three financial years. This costs includes organizational cost (per department) as well as the area of expenditures (Defra funded staff costs, Consultancy, IT development and other costs).

5.3. Operations



Operational aspects of PID include registration, validation, redirection and long term preservation.

Registration is the process of creation and registration of resources and their Persistent Identifiers. Validation is the process that checks that the format and location is correct and that the Persistent Identifier allocated to the resources is unique. Redirection is the process that uses the Persistent Identifier to redirect to the desired web resource. The long term preservation is the operation of archiving "dead" copies.

5.3.1. Registration

Creation and registration of valid PIDs are normally managed at the domain level. With regards to the domain discussion, the opinions are split. On one side UK goes for one rooted domain data.gov.uk with sub-domains which would give more trust to the consumer. On the other side Netherlands stresses the importance of the registers and the issues related to the sharing of a domain (need to find an owner for each sub-domain and not clear where to fit information.

5.3.1.1. Netherlands

In the paper "Towards a national URI-Strategy for Linked Data of the Dutch public", with regards to the registers it is summarized as following "No register, no identifier". It is suggested that one should mint URIs for concepts or objects that are recorded in a register. If such register doesn't exist, than one should be created.

With regards to the Domain discussion, it was recommended not to share internet domain, specially referring to the UK Public sector. Having a shared domain, though it gives more confidence to the consumer about the quality of the data, it also introduces several issues such as:

- Need to find an owner for each sub-domain
- Not clear where to fit information : train station is location or transport

5.3.1.2. UK

Public Sector Information Domain of the CTO Council's cross-Government Enterprise Architecture provides 2 recommendations for avoiding naming collisions (Designing URI sets for UK Public Sector (2009)

- URIs from a set that is promoted for re-use should not contain the name of the department or agency currently responsible for it. Accreditation of URI sets to be rooted at data.gov.uk;
- It is expect that the governance regime for "location.data.gov.uk" URI will be delegated from the COI/Cabinet Office to the UK Location Council.

Domains should:

- Expect to be maintained in perpetuity;
- Not contain the name of the department or agency currently defining and naming a concept, as that may be re-assigned;
- Support a direct response, or redirect to department/agency servers;
- Ensure that concepts do not collide;
- Require the minimum of central administration and infrastructure costs;
- Be scalable for throughput, performance, resilience;

"*The choice of domain should provide the confidence to the consumer, that the URI set has met minimum quality criteria, including implementing these design considerations. In other words, the domain itself should convey an assurance of quality and longevity.* "

5.3.2. Validation

When discussing about validation, most papers are about the management of the uniqueness of the resource. Uniqueness normally assured by using decentralized repositories with a unique namespace so they can generate unique identifiers and use opaque identifiers minimize the use of semantics.

5.3.2.1. Germany

PersID III.a – Current State and State of the Art & III.b – User Requirements states that uniqueness normally assured by using decentralized repositories with a unique namespace so they can create unique identifiers. Use opaque identifiers minimize the use of semantics. Germany adds a checksum of the identifier at the end of the identifier to enable detection errors when transferring identifiers.

Regarding the updated of identified objects, PersID III.a – Current State and State of the Art & III.b – User Requirements reports that Germany has features for version management to provide new identifier to a new version of the object.

For transferring objects to new owners, the PIDs should not be changed even if the owner changes. Creation of a new identifier it is also not desirable. The transfer between repositories is based on:

- Single records
- Collections
- Complete repositories.

In particular, the new location is registered as a new version

5.3.2.2. UK

Public Sector Information Domain of the CTO Council's cross-Government Enterprise Architecture explores design considerations for URIs to persist (Designing URI sets for UK Public Sector (2009). In order to maintain the design, The URI set publisher may provide a URI alias to the current version and publishers may wish to offer a facility to subscribe to their set so that they can alert consumers to changes, improvements, etc.

The URI template should not include the name of the organisation or project that minted the URI. This makes it much less susceptible to change should the project end or the organisation be merged or renamed.

The document also gives a recommended pattern for a URI designed for persistence:

- http://{domain}/{type}/{concept}/{reference}

Other general rules are provided in the study regarding the URI template in order to ensure valid persistent identifiers:

- Avoid Stating ownership
- Avoid version numbers
- Re-use existing identifiers
- Avoid using auto-increment
- Avoid query strings
- Avoid file extensions

5.3.3. Redirection

Redirection is about using Persistent URLs that redirect to another Web resource. For the redirection, most of the work points to the HTTP response code 303 to a document that describes the object.

5.3.3.1. UK

When de-referenced, URIs that identify real world objects that cannot be transmitted as a series of bytes (such as buildings, places and people) should redirect using HTTP response code 303 to a document that

describes the object. This should be done in a consistent manner that can be written as a URI re-write rule, typically replacing the URI {type} of 'id' with 'doc.' (**10 Rules for Persistent URIs).**

5.3.4. Long Term Preservation

With regards to the long term preservation it is only mentioned that the URI should be designed in such a way that independently of the life cycle of the resource, the URI should persist unaltered.

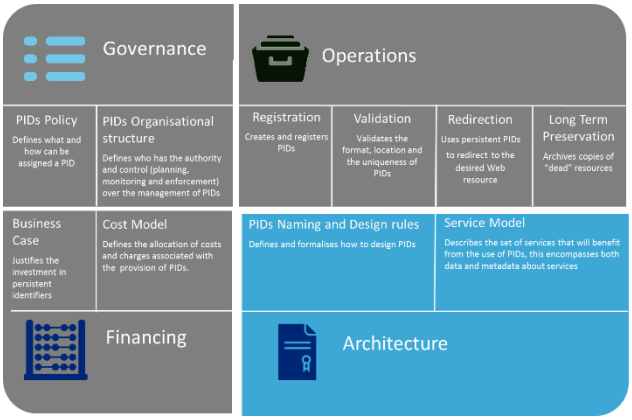
5.3.4.1. Germany

PersID III.a – Current State and State of the Art & III.b – User Requirements reported that National Library have a legal obligation to a LTP (Long Term Preservation).

5.3.4.2. UK

In the "Designing URI sets of UK public sets" it is recommended that once created, a URI should persist unaltered. To reflect the different versions or stages, these should be into the structure of the URI path, thus creating a separate URI for each stage.

5.4. Architecture



The architecture comprises both the PIDs naming/design rules and the service model.

The service model describes the set of services that will benefit from the use of PIDs, including both the data and the metadata about services.

5.4.1. Service Model

The typical services described are the View and Download services. These services are mostly described in the member state INSPIRE reports.

Country	Website (geoportal)	Download URI	View URI
Austria	http://www.inspire.gv.at/Geoportale/National.html	https://geometadaten.lfrz.at/at.lfrz.discoveryservices/srv/de/csw202?service=CSW&request=GetCapabilities&version=2.0.2	https://geometadaten.lfrz.at/at.lfrz.discoveryservices/srv/de/csw202?service=CSW&request=GetCapabilities&version=2.0.2
Belgium	http://www.geopunt.be/geowijzer/inspire/monitoring-en-rapportering		https://metadata.agiv.be/zoekdienst/apps/tabsearch/index.html?hl=dut&uuid=93795cd6-66d3-4310-83b2-5443adfee403
Bulgaria	http://www.esmis.government.bg/en/?t=inspire		http://www.esmis.government.bg/upload/docs/2013-04/MAP_100_broadband_EESM_ENGL_c.pdf
Czech Republic	http://geoportal.gov.cz.	http://geoportal.gov.cz/arcgis/rest/services/CENIA/cenia_rt_III_vojenske_mapovani/MapSe	http://geoportal.gov.cz/web/guest/map

		rver/WMTS/tile/1.0.0/CENIA_cenia_rt_III_vojenske_mapovani/default/default%20028mm/0/0/0	
Germany	http://www.geoportal.de/EN/GDI-DE/INSPIRE/Directive/Monitoring-and-Reporting/monitoring-and-reporting.html?lang=en		http://www.geoportal.de/gds/xml.php?uuid=03c2141a-8104-4a98-855f-bcc16ee9dce3
UK	http://data.gov.uk/		http://data.gov.uk/data/map-based-search

5.4.2. Design patterns

Most of the work is focused on the design patterns. Most papers refer to the 10 Rules of Persistent Identifiers and UK Design pattern rules.

5.4.2.1. EU – European Commission – 1O Rules of Persistent Identifiers

Paper often referred to in other papers. The main principles are summarized below:

Recommended URI Format:

- http://{domain}/{type}/{concept}/{reference}

Design Principles:

- Avoid Stating ownership
- Avoid version numbers
- Re-use existing identifiers
- Avoid using auto-increment
- Avoid query strings
- Avoid file extensions

Design and build for multiple formats:

- Multiple representations of the same resource should all link to each other using a suitable method

Implement 303 redirects for real-world objects

- When de-referenced, URIs that identify real world objects that cannot be transmitted as a series of bytes (such as buildings, places and people) should redirect using HTTP response code 303 to a document that describes the object. This should be done in a consistent manner that can be written as a URI re-write rule, typically replacing the URI {type} of 'id' with 'doc.' Use a dedicated service

5.4.2.2. Ireland - Digital Enterprise Research Institute

In the paper "Applying Design Patterns in URI Strategies - Naming in Linked Geospatial Data Infrastructure" (Sonya Abbas)the following URI design related issues:

1. Consolidating existing URI design rules,
2. Distilling core URI design aspects or facets from these rules and
3. Abstracting the rules into a set of consistent URI Design Patterns specifications.

The paper summarized as weaknesses of the existing guidelines the following:

- Limited pragmatic value;
- Abstract nature of the rules;
- Weak elaboration on nature of problems addressed.

The paper summarises the following conclusions:

1. Many cases are too abstract and underspecified making their implementation difficult, e.g. “Ensure that URIs do not have to change with every re-design”,
2. The guidelines do not elaborate on nature of problems addressed and consequences of prescribed design decisions, e.g. “if no auto-increment is to be allowed in generating URIs, how will URI’s for a large dataset be generated?
3. There are similar rules across documents without any explicit references to related rules in other sources, e.g. “one rule specifying no file extension in URI and another rule from a second source indicating no mutable element in URI”;
4. The guidelines are inconsistent when consolidated across different sources, e.g. “one rule indicating having the right domain in URI and another rule specifying not having domain information in URI”.

5.4.2.3. UK – Chief Technology Officer Council

The UK Public Sector design patterns, often referred as starting point in other work, describes how to design URI sets and the path structure for URIs.

Principle	
Use HTTP so that URIs can be resolved	MUST
Use a consistent path structure to explicitly indicate the type of URI	RECOMMEND
The publisher will make it clear whether the set is promoted for re-use by other parts of government and/or the public	MUST
Public sector URI sets should publish their expected longevity, and potential for re-use	MUST
Those public sector URI sets that are promoted for re-use should be designed to last for at least 10 years RECOMMEND	RECOMMEND
Where more than one Representation URI is available, provide a Document URI where Content Negotiation can be used to provide the most appropriate representation	RECOMMEND
Avoid exposing the technical implementation of a URI in its structure	RECOMMEND
As a minimum, provide a machine-readable Representation URI	MUST
If appropriate, provide a human-readable Representation URI in HTML	RECOMMEND
Provide a means of discovering each of the available Representation URIs for a single Document URI	RECOMMEND
A URI set will publish its authorisation, authentication, and data quality characteristics using a common vocabulary	MUST
A URI structure will not contain anything that could change, such as session IDs	MUST
A URI path structure will be readable so that a human has a reasonable understanding of its contents	RECOMMEND

More design patterns specific to Location in the paper "Designing URI Sets for Location".

Additionally, the Cabinet Office on the Standard Hub puts forward a URI pattern template: http://www.rfc-editor.org/rfc/rfc6570.txt and having the functional needs:

- The design of URI patterns must be clearly understandable for both new users and experts
- The approach needs to be flexible enough to meet the needs of different public sector organizations, some of which will have many different sets of data
- The patterns approach should as far as possible be compatible with existing significant government linked data publishing efforts, so that current best practices do not become wrong
- It should balance clarity of structure with flexibility and length of URIs
- It should cover the majority of cases but leave room for extensibility to deal with unusual cases
- Table below describes path structure for Location URIs prescribed by the Chief Technology Officer Council (**Designing URI Sets for Location**).

URI Template Field	Description
{sector}	The name of the data.gov.uk **sector** under which a **concept** and its related reference data is governed. e.g. transport, education, environment.
{concept}	The **sector** specific **concept name** for the type of entities associated with a given **reference** designator. e.g. road, school, river
{reference}	A **reference** value used to discriminate between individual instances of a **concept**. Reference values are typically derived from a common codeset.
{version}	An optional field used to distinguish between distinct versions of either spatial-thing or their reference documents/objects. Note that versioning of things and their corresponding reference documents/objects are independent of each other. References made without a **{version}** refer to the most recent version at the time the reference is de-referenced (followed).
{theme}	A two letter code for the corresponding INSPIRE theme – see Annex II.
{class}	The INSPIRE conceptual model **class name** corresponding to the most specific (leaf-level) feature-type used in expressing a spatial-object or in abstracting a spatial-thing. By convention class names begin with an uppercase letter to distinguish them from property names.
{codeset}	An optional **codeset** name that indicates the codeset from which {reference} values are taken.
{rendition}	Optionally provides a way of identifying different possible document renderings of a spatial-object, e.g. alternative renderings may be available in other formats such, html, rdf, json or plain-text amongst others. When this field is omitted an appropriate rendering may be selected through content-negotiation or a default rendering may be supplied.

URI Template Field	Description
{namespace}	The namespace component of an INSPIRE Unique Object Identifier.
{localId}	The localId component of an INSPIRE Unique Object Identifier.
{versionId}	The versionId component of an INSPIRE Unique Object Identifier.
{authority}	A fully qualified domain name that serves as the authority field of an RFC3986 URI. Typically governance of the URI namespace based on the value of this field falls to or is delegated from the organisational entity responsible for the domain name assignment. {sector}.data.gov.uk represents a pattern of {authority} with common governance requirements delegated from data.gov.uk. It is expected that in the first instance governance of location.data.gov.uk will fall to the UK Location Programme.
{property}	A **property name** derived from the INSPIRE conceptual model and its application schema. By convention property names begin with a lowercase letter to distinguish them from class names.
{term}	A defined term in an ontology, concept schema or codelist, typically corresponds with {class} or {property} values.
{package}	A package name from an INSPIRE conceptual schema, used it situations where it is necessary to disambiguate URI for otherwise similarly named but distinct vocabulary terms.

5.5. PID initiatives

CrossRef - "CrossRef is an independent membership association, founded and directed by publishers. CrossRef's mandate is to connect users to primary research content, by enabling publishers to work collectively."

DataCite - "The objectives of this initiative are to establish easier access to scientific research data on the Internet, to increase acceptance of research data as legitimate, citable contributions to the scientific record, and to support data archiving that will permit results to be verified and re-purposed for future study. DataCite will promote data sharing, increased access, and better protection of research investment."

EPIC - EPIC (European Persistent Identifier Consortium) is an initiative from Max Planck and CLARIN that provides allocation and resolution of persistent identifiers for the European research community. They acknowledge that: ...one needs a commonly agreed process and due to the importance of the resolution of the references to actual URLs for a lot of transactions, the needed resolution service has to have a high degree of robustness and reliability in the long-term. http://www.pidconsortium.eu

PersID12 is a joint project of the current implementers of the URN:NBN namespace, facilitated and funded by Knowledge Exchange13. Their main goal to harmonize the use of the systems of the different URN:NBN implementers on the level of policy, technology and communication. Such transparency is

required to support the different stakeholders of the URN:NBN system, but it also gives insight in how to deal with the use of 'different' persistent identifiers across an infrastructure.

The Australian National Data Service (ANDS) provides identifier services for staff at universities, government agencies, publicly funded research organizations, museums, galleries, archives, libraries and any custodian of data relevant to research. The service has an Australian focus but aims to assist the emergence of a global data commons.

Stelselcatalogus 2.0 http://www.e-overheid.nl/onderwerpen/stelselinformatiepunt/stelsel-van-basisregistraties/stelselvoorzieningen/stelselcatalogus.

OWMS, the metadata standard for Dutch public. OWMS is a Dublin Core Application Profile. It is specific for information objects within the domain of the Dutch public sector and allows in its turn to create more specific standard for information types within that domain by creating content models. The ranges of the metadata properties are specified in terms of the Dublin Core Abstract Model. Property values can be specified as pointers to concepts in the OWMS Ontology. The OWMS Ontology contains URI's for organizations, information types, themes, geographic enitities and others.

PiLOD – Platform Implementation Linked Open Data , http://www.pilod.nl/wiki/PiLOD.

PURL.org

OCLC Research Europe (located in the NL) + US (Zepheira) – not yet live. PURLS are Web addresses that act as permanent identifiers in the face of a dynamic and changing Web infrastructure. Instead of resolving directly to Web resources, PURLs provide a level of indirection that allows the underlying Web addresses of resources to change over time without negatively affecting systems that depend on them. This capability provides continuity of references to network resources that may migrate from machine to machine for business, social or technical reasons.

http://EZID.cdlib.org/

From US (University of California), EZID (easy-eye-dee) makes it easy to create & manage unique, long-term identifiers. The scope of this initiative is:

- create identifiers for anything: texts, data, bones, terms, etc.
- store citation metadata for identifiers in a variety of formats
- update current URL locations so citation links are never broken
- use EZID's programming interface for automated operation at scale
- choose from a variety of persistent identifiers, including ARKs and DataCite DOIs.

www.DOI.org

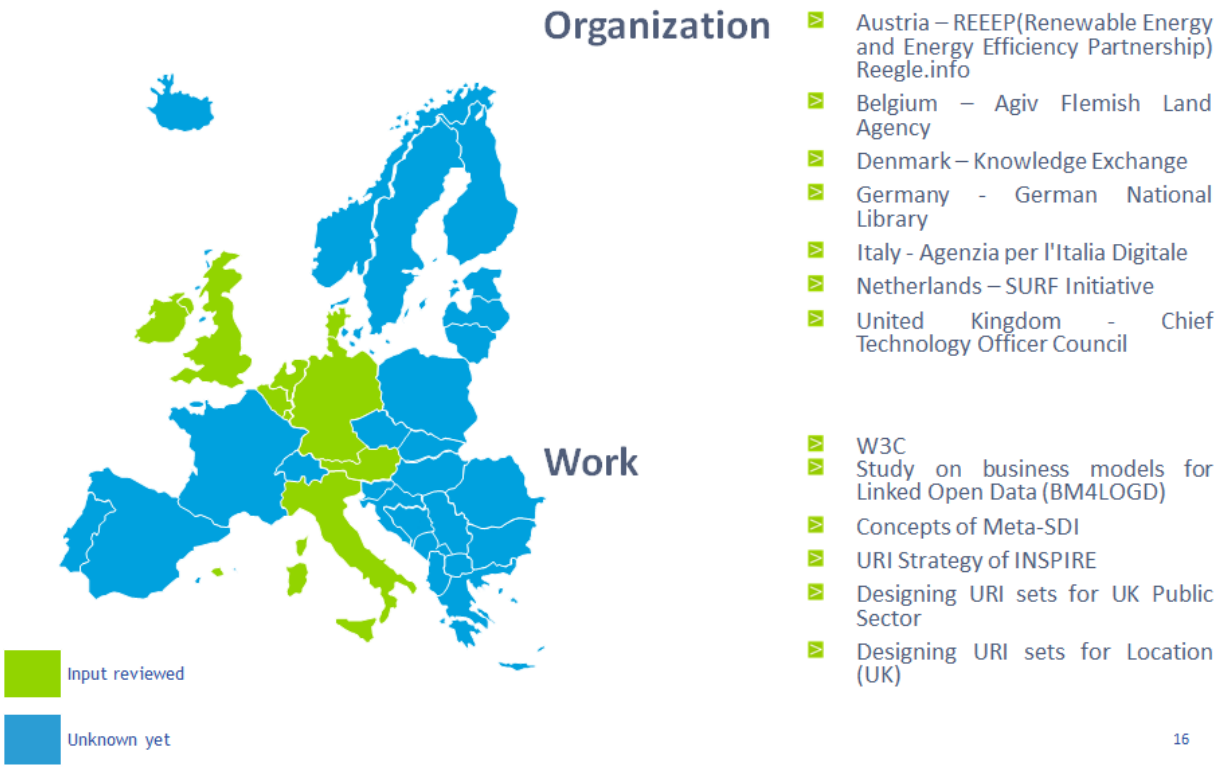
International DOI Foundation (IDF) – Delaware USA. Digital Object Identifier is an alternative to URI

- The DOI system concept
- DOI system components
- DOI name syntax
- DOI name resolution
- DOI® data model
- DOI system implementation
- Benefits of the DOI system

5.6. Conclusions

5.6.1. Governance

The picture below depicts the available data analyzed until now.



The conclusions that could be drawn based on the analyzed data are the following:

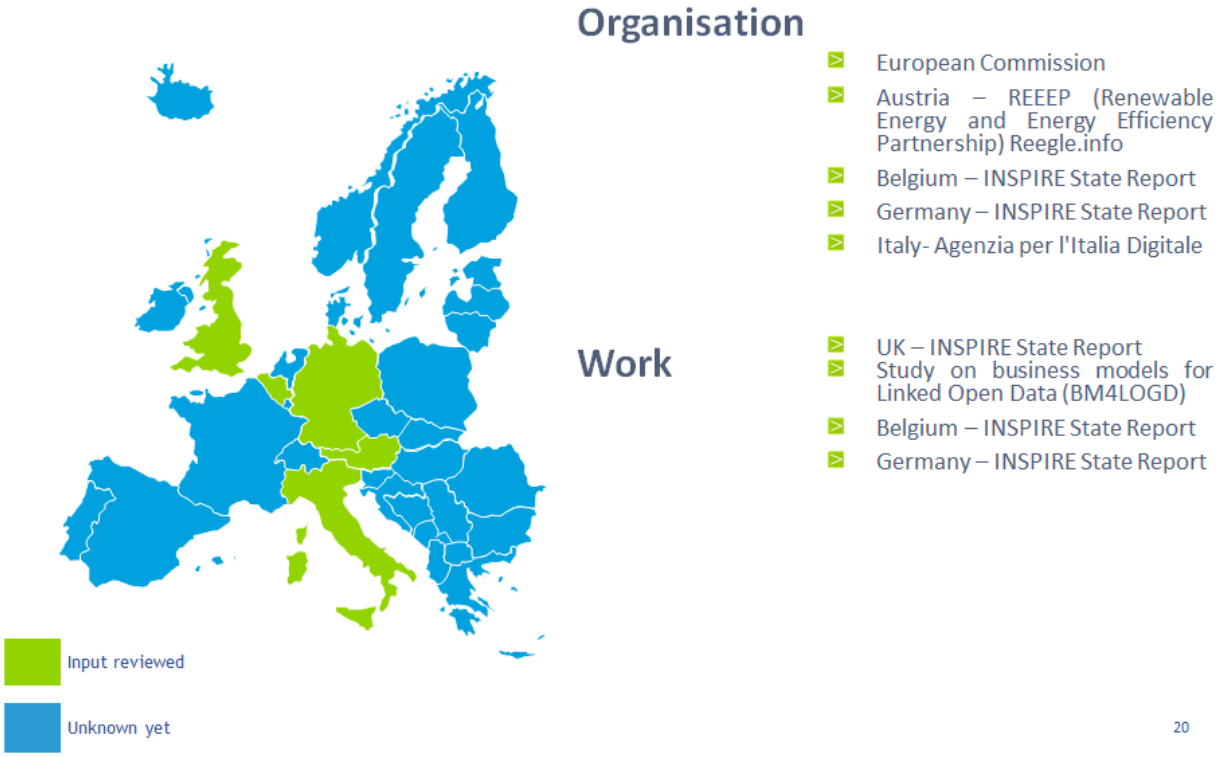
1) Policy is often **associated with URI design patterns** and the **management of domains**. However, there is definitely more than these two elements.
2) When a policy exists it **isn't always well formalized** depending on the size and organizational structure.
3) There **seem to be no examples of PID organizational structures** in the context of INSPIRE and in e-Government.
4) There is **a lot of theoretical work** about the "technical side" of persistent identifiers, often associated with linked data.
5) **Lack of a holistic picture** that integrates the several GOFA dimensions.

Having the above conclusions, the following barriers can be identified:

1) PIDs governance is an inherently complex topic given the way that "ICT governance and management" is done in the different Member States.
2) There is "no silver bullet" for PID governance

5.6.2. Financing

The picture below depicts the available data analyzed until now.



The conclusions that could be drawn based on the analyzed data are the following:

1) There are no formalized business cases for PIDs – we believe that the interest in linked data is sometimes enough:

e.g. John Sheridan, Head of Legislation Services "A business case for using linked data would like making a business case for using electricity."

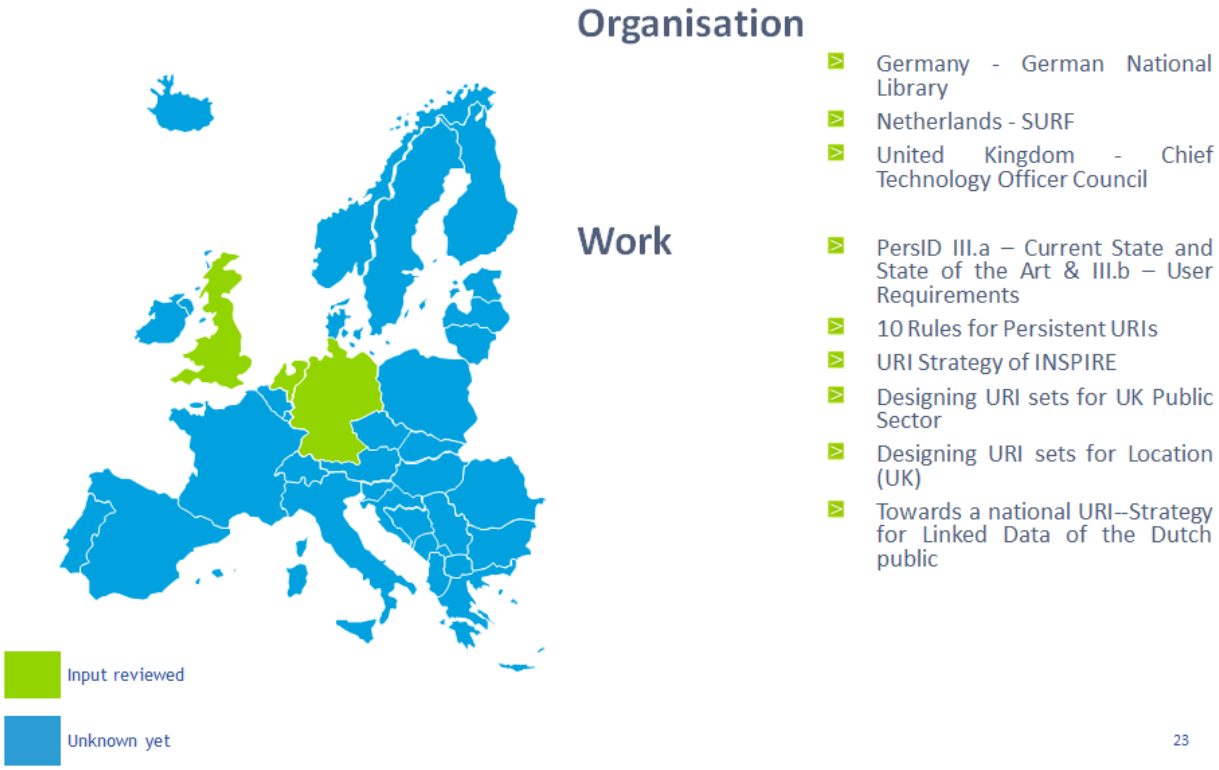
2) There seems to be no cost model for PIDs. For Linked Data or the implementation of INSPIRE, when they can be found, they are very different from one case to the other. Below are a few examples:

   - Development costs, Maintenance costs, Promotion costs – Linked Data
   - Operating costs of the IT infrastructure, Production of interoperability Business model for the implementation of INSPIRE, IT structure Processing of spatial data, Personnel – INSPIRE Germany

3) No clear understanding on the investment needed to set up PIDs.

5.6.3. Operations

The picture below depicts the available data analyzed until now.



The conclusions that could be drawn based on the analyzed data are the following:

1) The **domain owner is implicitly** considered to be responsible for registration and validation.
2) There is **little discussion about long term preservation** and most papers mention 303 Redirection.
3) There **is little discussion about validation** of PIDs (it is much more about standardization and very little about control).

5.6.4. Architecture

The picture below depicts the available data analyzed until now.

The conclusions that could be drawn based on the analyzed data are the following:

1) Most referenced work is the **10 Rules for Persistent URI** and **Designing URI sets for UK Public Sector.**
2) There is **no EU agreed policy** for PIDs for centrally managed, shared resources.
3) **Several guidelines have been put forward** for the URI persistence (W3C - RFC2616., UK - Designing URI sets for UK Public Sector, ISA), but currently these are just recommendations and not always aligned.
4) Control and validation are as important as standardization but often overlooked.

6. REFERENCES

Barnes, I. (n.d.). *Persistent Identifiers - Working level.* Retrieved from http://ands.org.au/guides/persistent-identifiers-working.pdf

Brink, L. v. (n.d.). *Towards a national URI-Strategy for Linked Data of the Dutch public sector.* Retrieved from http://www.pilod.nl/w/images/a/aa/D1-2013-09-19_Towards_a_NL_URI_Strategy.pdf

Commission, E. (2013). *Study on business models for Linked Open Data*. Retrieved from http://ec.europa.eu/isa/documents/study-on-business-models-open-government_en.pdf

Heath, T., Hausenblas, M., Bizer, C., Cyganiak, R., & Hartig, O. (2008, 10 27). *How to Publish Linked Data on the Web*. Retrieved 12 21, 2013, from http://events.linkeddata.org/iswc2008tutorial/

Jentzsch, A. (2011, 09 19). *File: LOD Cloud Diagram as of September 2011*. Retrieved 12 21, 2013, from Wikipedia: http://en.wikipedia.org/wiki/File:LOD_Cloud_Diagram_as_of_September_2011.png

Sonya Abbas, A. O. (n.d.). *Applying Design Patterns in URI Strategies - Naming in Linked Geospatial Data Infrastructure.* Retrieved from https://www.deri.ie/content/applying-design-patterns-uri-strategies-naming-linked-geospatial-data-infrastructure

W3. (2013, 11 26). *Linking Open Data - W3C SWEO Community Project*. Retrieved 12 21, 2013, from W3C: http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData