Requirements of the SmartOpenData Infrastructure

Deliverable D2.1 :: Public

Keywords: Open Data, Requirements, infrastructure

Linked Open Data for environment protection in Smart Regions
Table of Contents

Executive Summary ........................................................................................................................................ 6

1  Introduction ................................................................................................................................................ 7
   1.1  Open Geospatial Data ...................................................................................................................... 8
   1.2  SmartOpenData Requirements and Architecture ............................................................................. 11
   1.3  SmartOpenData vision and use ......................................................................................................... 13
       1.3.1  Open Data Support .................................................................................................................... 17

2  The SmartOpenData Infrastructure ............................................................................................................. 20
   2.1  Open Source Licence Model ............................................................................................................ 24

3  Relevant SmartOpenData components in the scenarios proposed ......................................................... 25
   3.1  Characteristics of the SmartOpenData components ........................................................................... 25
   3.2  Research issues that SmartOpenData needs to address ................................................................... 30
       3.2.1  Current Shortcomings to be addressed by SmartOpenData ................................................... 31
       3.2.2  Machine translation in Spatial Data and Metadata ................................................................... 32
       3.2.3  Large Volumes of Real Time Data ........................................................................................... 33
       3.2.4  Multilingualism ....................................................................................................................... 33
   3.3  Required impacts of SmartOpenData ................................................................................................. 34
   3.4  Specific requirements for each of the SmartOpenData Pilots ............................................................ 37

4  How the SmartOpenData components can be classified or generalised ............................................... 40
   4.1  Generic Components Interfaces ...................................................................................................... 41
   4.2  Multilingualism ............................................................................................................................... 42
   4.3  Machine Translation .......................................................................................................................... 43
   4.4  Large volumes of real-time data .................................................................................................... 43

5  Legacy Architectures and existing Systems .............................................................................................. 44
   5.1  CKAN data portal ............................................................................................................................. 45
   5.2  Junar platform .................................................................................................................................... 47
   5.3  Socrata suite of products .................................................................................................................. 47
   5.4  SemaGrow project ............................................................................................................................ 50
   5.5  GeoKnow project ............................................................................................................................... 52
   5.6  HABITATS Reference Laboratory .................................................................................................... 54
   5.7  Plan4Business Open Data Repository ............................................................................................... 56
   5.8  InGeoCLOUDS project ................................................................................................................. 58
   5.9  LOD2 project ................................................................................................................................... 61
5.10 GetLOD and the Emilia Romagna Region Geoportal ........................................ 64
5.11 iMarine e-Infrastructure .................................................................................. 67
5.12 NatureAlliance and TESS Project ..................................................................... 69
5.13 DataLift Platform ............................................................................................... 72
5.14 Other related projects and initiatives ................................................................. 75
6 Relevant legal and political fundamentals .............................................................. 76
6.1 Legal barriers to linked open data ....................................................................... 77
6.2 Creative Commons ............................................................................................... 78
6.3 Open GI Data and the PSI Directive .................................................................... 80
7 Quality Constraints .................................................................................................. 81
7.1 Metadata Quality in a Linked Data Context ......................................................... 84
7.2 Service Level Agreements (SLA) ......................................................................... 90
7.2.1 Common metrics ............................................................................................. 91
7.2.2 Web and Cloud SLAs ...................................................................................... 91
7.2.3 SLA Management ............................................................................................ 92
7.2.4 Five Star Open Data ....................................................................................... 94
7.2.5 SmartOpenData Service Level Agreement .................................................... 97
8 Conclusions and Requirements ............................................................................... 98
8.1 Generic Requirements ......................................................................................... 98
8.2 Service Requirements ......................................................................................... 99
8.3 Specific Requirements of the SmartOpenData infrastructure ............................. 101
Annex A: Linked Open Data ....................................................................................... 107
Global Open Data Initiative Declaration .................................................................. 112
Annex B: The SmartOpenData Pilots ........................................................................ 116
1. Spain & Portugal - Agroforestry Management ...................................................... 116
2. Ireland - Environmental Research and Biodiversity ............................................. 117
3. Italy - Water Monitoring ...................................................................................... 119
4. Czech Republic - Forest Sustainability ................................................................. 120
5. Slovakia-Environmental Data reuse ................................................................... 121
Annex C: OPQUAST OpenData best practices ......................................................... 123
Annex D: SMEspire Best Practice Catalogue ............................................................ 125
Annex E: List of Abbreviations .................................................................................. 130
List of Figures

Figure 1 Open Data Support services ................................................................. 18
Figure 2 SmartOpenData Infrastructure ............................................................... 21
Figure 3 CKAN High Level Architecture ............................................................ 46
Figure 4 SemaGrow Architecture ..................................................................... 51
Figure 5 GeoKnow High-Level Architectural Overview .................................... 54
Figure 6 HABITATS Reference Laboratory Architecture ................................... 55
Figure 7 Plan4business Open Data Repository Architecture ............................ 57
Figure 8 Plan4Business and INSPIRE data models .......................................... 57
Figure 9 InGeoCLOUDS architecture ............................................................... 59
Figure 10 LOD2 Stack of Tools ......................................................................... 62
Figure 11 gCube Infrastructure Enabling Services .......................................... 69
Figure 12 VISION Cloud SLA Management system ......................................... 94
Figure 13 5 Star Open Data ................................................................................ 94

List of Tables

Table 1 SmartOpenData Pilot Environmental Scenarios .................................... 37
Table 2 Functions required in each of the Pilot Scenarios ................................. 39
Table 4 Plan4Business Characteristics ............................................................... 56
Table 5 InGeoCLOUDS services ................................................................... 58
Table 6 Metadata-quality indicators ................................................................. 85
Table 7 Views on LOD Quality Criteria .......................................................... 87
Table 8 Rules for the 5 Stars of Open Data ...................................................... 95
Contractual Date of Delivery to the EC: January 2014

Actual Date of Delivery to the EC: 31st January 2014

Editor(s): John J O’Flaherty, MAC

Contributor(s): TRAGSA, UPM, MAC, Sindice, MWRA, ARPA, SpazioDati, HSRS, UHUL FMI, CCSS, IMCS, DGDOTER, SAZP

DOCUMENT HISTORY

<table>
<thead>
<tr>
<th>Version</th>
<th>Version date</th>
<th>Responsible</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>31/01/2014</td>
<td>MAC</td>
<td>Version for submission. Based on feedback and inputs from Partners.</td>
</tr>
<tr>
<td>0.3</td>
<td>MAC</td>
<td>Update based on Partners feedback and inputs.</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>MAC</td>
<td>Updates based on the Report from and discussion at the Kickoff Meeting on 4-5 Dec 2013.</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>15/11/2013</td>
<td>MAC</td>
<td>Initial draft based on feedback from Partners, desk research and DoW.</td>
</tr>
</tbody>
</table>

The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Communities. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

Copyright © 2014, SmartOpenData Consortium.
Executive Summary

The SmartOpenData project will create a Linked Open Data infrastructure (including software tools and data) fed by public and freely available data resources, existing sources for biodiversity and environmental protection, and research in rural and European protected areas and its National Parks. This deliverable identifies the basic environmental requirements of the SmartOpenData infrastructure. It describes (a) the relevant components in the scenarios proposed, (b) the characteristics of those components, (c) how the components can be classified or generalised, (d) what legacy architectures exist, (e) what the relevant legal and political fundamentals are, and (f) what quality constraints exist and how they can be defined in the context of service level agreements (SLA).
1 Introduction

The Linked Open Data strategy is becoming a source of unprecedented visibility, enabling the generation of new businesses and significant advances for research in the environmental area. Nevertheless, in order for this strategy to become a reality, it is necessary to advance the publication of existing environmental data, usually owned by public bodies.

The SmartOpenData project will create a Linked Open Data (LOD) infrastructure (including software tools and data) fed by public and freely available data resources, existing sources for biodiversity and environmental protection, and research in rural and European protected areas and its National Parks. This will provide opportunities for organizations to generate new innovative products and services that can lead to new businesses in the environmental, regional decision-making and policy areas among others. The value of the data will be greatly enhanced by making it available through a common query language that gives access to related datasets available in the linked open data cloud. Organizations such as Environmental Agencies and National Parks will benefit by improving their knowledge of their biodiversity, maintenance and protection. Public bodies, researchers, companies and European citizens will take a central role in user-driven pilots developed to enhance the potential of protected areas. Innovation by third party companies and SMEs will be encouraged by the promotion of royalty-free open standards and best practices generated, initiated or simply highlighted by SmartOpenData.

SmartOpenData will provide Linked Open Data for environment protection in Smart Regions, by focusing on how the LOD Initiative can be linked with INSPIRE, GEOSS Data-CORE, GMES and external third parties and how it can impact on the economic and sustainability progress in European Environmental research and protection.

There exists many different information sources for protecting the environment in Europe, mainly focused on the Natura 2000 network, and areas where environmental protection and activities like tourism need to be balanced. SmartOpenData aims to define mechanisms for acquiring, adapting and using Open Data provided by existing sources for environment protection in European protected areas. Through target pilots in these areas, the project will harmonise metadata, improve spatial data fusion and visualisation and publish the resulting information according to user requirements and Linked Open Data principles to provide new opportunities for SMEs.

This deliverable describes the basic environmental requirements for the SmartOpenData infrastructure, including:

(a) the relevant components in the scenarios proposed
(b) the characteristics of these components
(c) how these components can be classified or generalised
(d) what legacy architectures exist
(e) what the relevant legal and political fundamentals are, and
(f) what quality constraints exist and how they can be defined in the context of service level agreements (SLA).
This project is focused on how Linked Open Data can be applied generally to spatial data resources and specifically to public open data portals, GEOSS Data-CORE, GMES, INSPIRE and voluntary data (OpenStreetMap, GEPWIKI, etc.), and how it can impact on the economic and sustainability progress in European Environmental research and Biodiversity Protection.

There are many different information sources for protecting biodiversity and environmental research in Europe - in coastal zones, agricultural areas, forestry, etc. -, mainly focused on the Natura 2000 network, and areas where environmental protection and activities like agriculture, forestry or tourism need to be balanced with the Habitats Directive and the European Charter for Sustainable Tourism in Protected Areas. Nevertheless, the economic value of these areas is still largely unknown.

SmartOpenData will define mechanisms for acquiring, adapting and using Open Data provided by existing sources directly involved in the project for biodiversity and environment protection in rural and European protected areas and its National Parks. Through target pilots in these areas, the project will

(i) harmonise geospatial metadata (ISO19115/19119 based) with principles of the Semantic Web,
(ii) provide spatial data fusion introducing principles of Linked Open Data,
(iii) improve spatial data visualisation of Geospatial Linked Open Data
(iv) publish the resulting information according to user requirements and Linked Open Data principles to provide new opportunities for SMEs.

The project will reuse existing European Spatial Data Infrastructures (SDI), based on INSPIRE, GMES and GEOSS (Free Pan European Data Sets such as CLC, Natura 2000, Habitats, Plan4all, Plan4business, EnviroGRIDS, Briseide1, GEOSS registries, national INSPIRE portals, thematic portals like National Forestry portals together with local and regional data) and will extend it using Linked Open Data. Research and Development Partners will provide extensions of current INSPIRE/GMES/GEOS based Spatial Data Infrastructure. The SMEs involved will develop new services based on this data and research on biodiversity. Environmental Agencies and National Parks will benefit by improving their knowledge of their biodiversity, maintenance and protection. Public bodies, researchers, companies and European citizens will take a central role in user-driven pilots developed to enhance the potential of protected areas. Innovation by third party SMEs will be encouraged by the promotion of royalty-free open standards and best practices generated, initiated or simply highlighted by SmartOpenData.

1.1 Open Geospatial Data

Open public data resources for re-use is one of the key priorities of the Digital Agenda for Europe. Data available in public European organisations have an enormous potential economic growth. Nevertheless, finding and accessing environmental information isn’t always straightforward. The project will make spatial data easier to discover and use, having a positive impact on the public and standard availability of data according to the Linked Open Data Strategy for the purpose of environmental information. The target pilots will involve SMEs focusing on human activities (forestry, tourism, agriculture) in rural and protected areas such as National Parks and coastal zones. This availability will allow global

1 www.briseide.eu
environmental issues to be addressed, that are possible at this moment due to costs, efficiency and sustainability.

Producing and updating geospatial data is expensive and resource intensive. Hence, it becomes crucial to be able to integrate, repurpose and extract added value from geospatial data to support decision making and management of local, national and global resources. Spatial Data Infrastructures (SDIs) and the standardisation efforts from the Open Geospatial Consortium (OGC) serve this goal, enabling geospatial data sharing, integration and reuse among Geographic Information Systems (GIS). Geospatial data are now, more than ever, truly syntactically interoperable. However, they remain largely isolated in the GIS realm and thus absent from the Web of Data. Linked data technologies enabling semantic interoperability, interlinking, querying, reasoning, aggregation, fusion, and visualisation of geospatial data are only slowly emerging\(^2\). The vision of SmartOpenData is to leverage geospatial data as first-class citizens in the Web of Data, in proportion to their significance for the data economy.

Currently, there are three major sources of open geospatial data in the Web:

1. Spatial Data Infrastructures,
2. Open data catalogues, and
3. Crowdsourced initiatives.

Spatial Data Infrastructures (SDIs) were created to promote the discovery, acquisition, exploitation and sharing of geographic information. They include technological and organisational structures, policies and standards that enable efficient discovery, transfer and use of geospatial data using the web\(^3\). Research and development in this field is closely tied to standardisation activities led by international bodies, namely the ISO/TC 211\(^4\), OGC\(^5\) and W3C\(^6\).

In Europe, the INSPIRE Directive follows the OGC open standards, and has defined common data models for a number of application domains, such as hydrography, protected sites and administrative units, to enhance interoperability of spatial data sets of the different European countries\(^7\). It provides the legal and technical foundations to ensure member state SDIs are compatible and usable on a transnational context.

The major open standard Web services regarding discovery and querying of geospatial data in SDIs are OGCs Catalogue Service and Web Feature Service respectively. The first allows the discovery of geospatial data based on their metadata (e.g. scale, coverage) and the second enables querying of geospatial data. Additional standards provide visualisation and access to maps and tiles (Web Map Service, Web Map Tile Service, Keyhole Markup Language, Web Feature Services, Web Coverage Services) and enable developers to programmatically invoke and compose complex geospatial analysis services (Web Processing Service). Currently practically all GIS and geospatial databases are fully compatible with these standards; GIS users can consume geospatial data from SDIs and publish geospatial

---

\(^2\) See Annex A for a discussion on Linked Data.
\(^4\) ISO /TC 211 Geographic Information/Geomatics, http://www.isotc211.org
\(^6\) W3C, http://www.w3.org
data to SDIs with a few clicks. On a practical level, it is clear that SDIs must be considered as having developed over time and are now stable data infrastructures. They represent a significant investment from the public and private sectors worldwide and are the basis for interoperability among significant scientific domains. Further, they constitute the most prominent source for high-quality open geospatial data. Thus, any contribution and advancement must either be directly involved in standardization efforts, or be based solely on existing standards, without directly affecting their applications.

Open data catalogues provide open geospatial data by

a) encapsulating existing SDIs and/or

b) ad-hoc publishing available geospatial data as files.

In the latter case, geospatial data are published as regular open data. The only difference regards the use of file formats of the geospatial domain (e.g. SHP, KML) and availability of data for specific coordinate reference systems (typically national CRS). In the former case, an available national/regional SDI is exploited as a source for harvesting its geospatial data. The Catalogue Service is used to discover available data, and their metadata are added in the open data catalogue for homogenised data discovery. The actual data are available as exported file snapshots in common geospatial formats as before, or through the query services provided by the SDI. Consequently, open data catalogues typically offer geospatial data as files and at best expose any available SDI services for data access.

Crowdsourced geospatial data are emerging as a potentially valuable source of geospatial knowledge. Among various efforts perhaps OpenStreetMap, GeoNames, and Wikipedia are the most significant. GeoNames\(^8\) provides some basic geographical data such as latitude, longitude, elevation, population, administrative subdivision and postal codes. This data is available as text files and also accessible through a variety of web services such as free text search, find nearby or even elevation data services. Providing a larger variety of data, OpenStreetMap (OSM)\(^9\) has become an important platform for mapping, browsing and visualising spatial data on the Web. OSM data is available in different formats\(^10\) which can be imported into a database for its usage; it also provides web services to do search by name and inverse geocoding functionality.

The benefits of semantic technology for spatial metadata and data management are explored in a number of topics. For example, ontologies have been used in the form of taxonomies on thematic web portals (e.g. habitat or species taxonomies, categories of environmentally sensitive areas, or hierarchical land use classifications). The role of these ontologies is however limited. They provide background knowledge, but only in some experimental prototypes they are used for constructing search requests or for grouping of search results into meaningful categories. Further, in experimental settings, there are examples of using OWL for bridging differences in conceptual schemas\(^11\). The role of ontologies and knowledge engineering in these prototypes is basically to provide methodologies for integration and querying\(^12\). Ontologies have played an important role in

---

\(^8\) [http://www.geonames.org/](http://www.geonames.org/)

\(^9\) [http://www.openstreetmap.org/](http://www.openstreetmap.org/)


\(^12\) See for example Tian Zhao, Chuanrong Zhang, Mingzhen Wei, and Zhong-Ren Peng. Ontology-based geospatial data query and integration. In GIScience, volume 5266 of Lecture Notes in Computer Science, pages
structuring data of geospatial domains\textsuperscript{13}. However, semantic technology has not influenced spatial data management yet, and mainstream GIS tools are not yet extended with semantic integration functionality.

Early work included the Basic Geo Vocabulary\textsuperscript{14} by the W3C, which enabled the representation of points in WGS84, and GeoRSS\textsuperscript{15} which provided support for more geospatial objects (lines, rectangles, polygons). In addition, GeoOWL\textsuperscript{16} was developed to provide a more flexible model for geospatial concepts. Furthermore, topological modelling of geometric shapes in RDF can be done with the NeoGeo Geometry Ontology\textsuperscript{17}. However, all these ontologies only supported WSGS84, and currently offer limited support for geospatial operations required in real world GIS workloads.

GeoSPARQL has emerged as a promising standard from W3C for geospatial RDF, with the aim of standardising geospatial RDF data insertion and query. GeoSPARQL provides various conformance classes concerning its implementation of advanced reasoning capabilities (e.g. quantitative reasoning), as well as several sets of terminology for topological relationships between geometries. Therefore, different implementations of the GeoSPARQL specification are possible, depending on the respective domain/application. In addition, GeoSPARQL closely follows existing standards from OGC for geospatial data, to facilitate spatial indexing from relational databases.

\section*{1.2 SmartOpenData Requirements and Architecture}

Having the web of data enriched with spatial data implies the consideration of standards such as SDI in the creation of a Geospatial Semantic Web. Working with spatial and non spatial data is challenged by the scalability requirements due to the size of datasets, the integration of data using the different models, and the transformation of data in specialised domains. SmartOpenData aims to provide easy to use tools for non experts in cartography nor Linked Data to exploit the data and create web based geospatial enabled applications.

The main focus of the SmartOpenData project is to build a Geospatial Linked Open Data infrastructure fed by public and crowdsource data resources, existing sources for biodiversity and environment protection in European protected areas that satisfies the requirements of four kinds of target users that have been identified: (i) Public bodies, (ii) Researchers, (iii) Companies and (iv) Citizens.

To this end, work package WP2 is generating a detailed collection of requirements and use cases for the SmartOpenData infrastructure, aiming at making available a valuable quantity of open data resources through LOD and semantic approaches by easing its use and

\footnotesize
\begin{itemize}
\item http://www.w3.org/2003/01/geo/
\item GeoOWL. http://www.w3.org/2005/Incubator/geo/XGR-geo-20071023/
\item NeoGeo Geometry Ontology. http://geovocab.org/geometry.html
\end{itemize}
integration. The requirements identified in tasks T2.1 and T2.2 are together contributing to defining the overall technical architecture of the project in T2.3, specifying the interfaces between modules in order to provide a modular solution, which can be configured to provide functionalities that meet user’s needs and a commercial strategy, as well as integrating future modules.

This is the first deliverable from WP2 “Requirements and Architecture”, which aims to

1. Generate a detailed collection of requirements for the SmartOpenData infrastructure, including LOD interoperability, multilingual and semantic requirements.
2. Produce and improve the formal description of use requirements and use cases, taking into account the four kinds of target users that have been identified:
   - I. Public bodies
   - II. Researchers
   - III. Companies
   - IV. Citizens
3. Define a reference architecture model and high-level technical specification for the SmartOpenData infrastructure.

As WP2 Leader MAC led task T2.1 to coordinate with all partners in the first months of the project in identifying the basic environmental requirements of the SmartOpenData infrastructure and documenting them in this deliverable. MAC coordinated with tasks T2.2 and T2.3 as well as with WP3 (Data Modelling and Linked Open Data alignment) and WP4 (SmartOpenData Semantic Front-end Facilities), to produce this report as a key input to those tasks and workpackages. MAC worked with TRAGSA (as leader of T2.2) and HSRS (as leader of T2.3) in defining the overall technical architecture of the project in task T2.3 and their future delivery of D2.2 (User Requirements and Use Cases) and D2.3 (Architecture of SmartOpenData infrastructure) to be delivered in April 2014, as key inputs to WP5 (Demonstration Pilots) and WP6 (Evaluation, Assessment and User Groups), to complete WP2 and contribute to the first Project Milestone (Project Foundations).

This deliverable is the result of task T2.1, “Requirements of the SmartOpenData infrastructure”, which identified the basic environmental requirements for the SmartOpenData infrastructure, by addressing the following questions

1. What relevant components are in the scenarios proposed
2. What are the characteristics of these components
3. How these components can be classified or generalised
4. What legacy architectures exist
5. What relevant legal and political fundamentals are
6. What quality constraints exist and how they can be defined in the context of service level agreements (SLA).

The results are documented here as the requirements of the SmartOpenData infrastructure and define the boundaries of T2.2 and T2.3.

- **T2.2. User Requirements and Use Cases (led by TRAGSA)**

This task is collecting the requirements for the four kinds of users identified above, based on the commercial expertise of the SMEs, the requirements of their customers and an analysis of their competency. Several Use Cases are
being collected and documented with respect to the User requirements identified. The task is producing a detailed collection of user requirements and user cases for the SmartOpenData project with the aim of satisfying their needs in terms of data modelling and LOD alignment, semantic indexing infrastructure, visualisation framework, notification service and quality, and functionalities provided, that will be documented in the D2.2.

- **T2.3. Architecture of SmartOpenData (led HSRS)**

This task is researching and defining a reference infrastructure model and high-level technical specification for the project, including its main components and connection points to other tools and systems. It is using the reference model for open distributed processes (RM–ODP) methodology, to define the reference architecture to meet the technical and user requirements established throughout T2.1 and T2.2, addressing interoperability and multilingualism aspects, metrics engine and interfaces. The reference architecture will define both platform neutral components and will also provide suggestions for concrete implementation in D2.3.

The SmartOpenData infrastructure requirements definition process has carefully addressed the following risks that were identified in the DoW:

<table>
<thead>
<tr>
<th>Risk</th>
<th>Consequence</th>
<th>Plan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incomplete understanding of requirements</td>
<td>Unsatisfactory definition of ICT</td>
<td>The project is developed through iterative steps, and WP6 includes activities to assess progress and results to draw indications for problem recovery</td>
</tr>
<tr>
<td></td>
<td>platform functionality and data model</td>
<td>Platform intrinsic modularity and the foreseen activities of system integration and customisation ensure wide margins to modify the work in progress</td>
</tr>
<tr>
<td>Difficulty in “freezing” requirements</td>
<td>Specifications may change while solution design and development are in progress</td>
<td>The initial phase of requirements, use cases and architecture in WP2 is fundamental to reach a common understanding. However, the intrinsic project modularity allows development of components with different technologies.</td>
</tr>
<tr>
<td>because of the approach novelty</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lacking of consensus on the technological approach between competence areas</td>
<td>The development work could be delayed and lead to inconsistencies and integration problems</td>
<td></td>
</tr>
</tbody>
</table>

### 1.3 SmartOpenData vision and use

The vision of the SmartOpenData project is that environmental and geospatial data concerning rural and protected areas can be more readily available and re-usable, better linked with data without direct geospatial reference so different distributed data sources could be easily combined together. SmartOpenData will use the power of Linked Open Data to foster innovation within the rural economy and increase efficiency in the management of the countryside. The project will prove this in a variety of pilot programmes in different parts of Europe.
The fundamental SmartOpenData goal is to make INSPIRE/GMES/GEOSS infrastructure better available for citizens, but also mainly for SME developers. On one hand, Europe and EU invest hundreds of millions of Euros in building the INSPIRE infrastructure. On the other hand, most European SMEs and citizens use Google maps for their applications. National and regional SDIs offer information which is not available on Google, but this potential is not used.

One of the main goals of SmartOpenData is making European Spatial Data easily re-usable not only by GIS experts but also by SMEs.

In order to support Open Data Strategy for Europe and increase re-use of open public data from the European Commission, SmartOpenData will use where possible data and services from EC Open Data Portal. In addition, any application built on this data source will be registered on this portal. Similar initiatives are ongoing at national level, where SmartOpenData participants will disseminate the project outcomes in the same way.

**What does it mean “to use Linked Data for Spatial Data“?**

In the context of SmartOpenData, using linked data for spatial data means identifying possibilities for the establishment of semantic connections between INSPIRE/GMES/GEOSS and Linked Open Data spatial related content in order to generate added value. The project requirements are within the environmental research domain.

This will be achieved by making existing “INSPIRE based” relevant spatial data sets, services and appropriate metadata available through a new Linked Data structure. In addition, the proposed infrastructure will provide automatic search engines that will crawl additional available geospatial resources (OGC4 and RDF5 structures) across the deep and surface web. RDF structure is used to describe the relation between two objects (for example object A is next to object B, where object A and B could be eventually stored in different databases).

The SmartOpenData infrastructure will avoid duplicating information, and by re-using existing identifiers available in the Linked Open Data cloud, SmartOpenData immediately will have access to a lot of other data sources and these will be available through SPARQL queries. But our goal will go much further. The project will build an infrastructure of objects and relationships with the added value of further links. By associating existing geospatial data with URIs used elsewhere, recording semantic relationships and linking across different data sets, the objects will have greater context and therefore usefulness. Simple cases like the one above can be processed on the fly but the project will build computationally expensive queries in advance to aid typical analysis and will develop methods to store this added value information as triples, whether centralised, distributed, off-line, or, where possible, calculated on the fly. The aim is to achieve the highest performance in resolving the queries and delivering the required information and functionality with the minimum of data duplication. SmartOpenData will allow the interrogation of this data using linked data's query language, SPARQL directly and turning data as triples.

An example of the kind of spatial queries that SmartOpenData support is “which types of land covers are represented in specific protected areas?” Here we are working with two different sets, potentially stored in different repositories, where every set could have Gigabytes in binary form. If we would like to support generic queries we could in the worst case compare every object against every other object. However it is not possible to provide such analysis in real time online. Therefore it is necessary to build parallel RDF structures to
existing relational data. The duplication of this data in a triple store could be necessary, but would immediately introduce the problem of maintainability so data needs to be synchronised. Introducing semantic principles here will support better usability, though also will generate a list of research challenges.

**Why do we have to use Linked Data for Spatial Data?**

The main motivation to utilise the potential of Linked Data is to enrich the INSPIRE spatial content to enable improved related services to be offered and to increase the number, performance and functionality of applications.

In many cases querying data in INSPIRE (GEOSS) based data infrastructure (driven mainly by relational databases) is time consuming and often it is not sufficient and understandable for common Web users. In large databases such queries can take minutes or hours. In the case of distributed databases such a query is almost impossible or very complicated. SmartOpenData will improve this situation dramatically.

**When does it make sense to do it?**

The building of semantic Linked Data structure makes sense when we need to integrate heterogeneous data based on different domains. This structure does not depend on any formats of data, size of database or other attributes, but on common geographic concepts (e.g. river, road...) and their relations (e.g. a road crosses a river). The relevant part of real data and where needed metadata will be transformed to sub-elements of basic concepts (e.g. E55 6 is a road, Ohře is a river as well as Site of Community Importance8) and then relation between concrete object could be described (e.g. E55 crosses Ohře). This approach enables us to integrate or harmonise originally heterogeneous data based on the same concepts. An example we could mention: Protected Area XX is near road E55; in protected area XX you can see bird YY. Results of such combination could be, that near road 55 you can see bird YY.

To achieve this, the SmartOpenData approach will involve a number of key steps to be taken, each of which occurs in one or more of the planned WP5 pilots18:

1. **Describe target use case/s domain/s focused on**
   a) Agroforestry management,
   b) Environmental research and Biodiversity,
   c) Water monitoring and
   d) Forest sustainability
   e) Environmental data reuse.

   The use cases will be dealt with from end users’ (related stakeholders) point of view. Each use case will be described with a minimal essential structure and list the basic concepts to be taken into consideration. We expect a maximal re-using of existing semantic structures (controlled vocabularies, gazetteers, registers, registry services, etc.).

2. **Analyse current available data and also current data models, which are available (in the ideal case INSPIRE base).**

---

18 As described in Annex B.
3. Define new data RDF models (concepts and relations) or identify possible extensions of INSPIRE data models, which are optimal in order to fulfil all or particular use case/s.

4. Where necessary define transformation model/s (processing service) for transferring original data into RDF. This could be a complex process running for a long time and through many databases.

5. Run the transformation model (it could also include access to distributed databases).

6. Eventually store or generate on the fly new transformed data in RDF (this information needs to include also links to the original data).

7. Prepare user-friendly application interface for querying data (as a simple form to be able query data without standards experience). The queries should be divided (or fragmented) into two groups – spatial queries (processed in traditional spatial database) and semantic queries based on SPARQL.

8. Prepare visualisation of results of queries. It will include two possibilities:
   a) visualise a list of objects in some form;
   b) cartographic visualisation, which has to be provided in relation with original data. So it will be necessary to define some mechanisms, like Filter Encoding etc. that could visualise results on the base of queries.

In undertaking these steps there are several underlying questions to bear in mind:

- Which Use Cases really benefit from Linked Open Data?
- How can we guarantee consistency among RDF database and original data? The application and possible extension of existing standards, if not the creation of new ones, will be important in this respect.
- How should we define the best possible protocol for accessing original data? (WFS, SOS or querying directly database using SQL for example)
- How can we make best use of existing tools for visualisation?
- How will this new mechanism influence efficiency and speed of data processing and querying?

Why Linked Open Data and Metadata?

Metadata plays the crucial role as interface to the spatial content it describes. In the context of SmartOpenData, metadata serves as the exchange component allowing to bridge INSPIRE and other spatial worlds. Metadata will act as the entry point/interface providing essential information for transformation of spatial data to RDF structures.

It is an open question whether a file or a database should be one of the core concepts with attributes derived from metadata. But primarily we should deal with objects that can inherit some attributes from metadata.

It is a fundamental requirement that all SmartOpenData research activities will be aligned with this vision to create an Open Data infrastructure for environment protection, user engagement and creation of new business opportunities.
1.3.1 Open Data Support

SmartOpenData will work with and use the recommendations of Open Data Support, which is a 36 month project of DG RTD of the European Commission to improve the visibility and facilitate the access to datasets published on local and national open data portals in order to increase their re-use within and across borders. Open Data Support is a pan-European initiative targeting both those data publishers that are well underway but also the ones that are just starting (such as SmartOpen Data).

To achieve its objective, Open Data Support provides to (potential) publishers of open datasets, three types of services to local, regional and/or national public administrations publishing open data:

- **Data and metadata preparation, transformation and publication services** that will enable them to share the metadata of their datasets on the pan-European linked metadata infrastructure delivered by the project;
  - The common metadata vocabulary that they are using for describing datasets, is the DCAT Application Profile for data portals in Europe, which is a standard that many EU Member States are considering to adopt\(^{20}\).
  - The metadata harvesting and publishing platform that OpenDataSupport is using for collecting metadata of datasets from government data portals, transforming it into RDF, harmonising it according to the DCAT-AP, and publishing it as Linked Open Government Data (LOGD) is based on the Linked Open Data Management Suite developed in the LOD2 project\(^{21}\).
- **Training services in the area of (linked) open data**, aiming to build both theoretical and technical capacity to EU public administrations, in particular to favour the uptake of linked open data technologies, with a catalogue of online training and tests in order to further knowledge in the field of Open Data Support.
  - The training curriculum around LOGD that they have developed, which focuses on different aspects, e.g. rationale and benefits, data and metadata quality and licencing, as well as technical aspects of publishing data as LOGD. It comprises 10 self-contained training modules in 3 languages (English, French and German). It is available under an CC-BY licence at [training.opendatasupport.eu](http://training.opendatasupport.eu)
- **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.

The project is summarized in the following figure below.

---

\(^{19}\) [www.opendatasupport.eu](http://www.opendatasupport.eu) - Open Data Support is funded under SMART 2012/0107 ‘Lot 2: Provision of services for the Publication, Access and Reuse of Open Public Data across the European Union, through existing open data portals’(Contract No. 30-CE-0530965/00-17).


\(^{21}\) [https://github.com/nvdk/lodms-core/tree/virtuoso](https://github.com/nvdk/lodms-core/tree/virtuoso)
DG CONNECT is aiming to advance the Commission’s open data policy by accompanying the adoption process of the revised Directive on the re-use of Public Sector Information by Parliament and Council and by promoting open data policies across the European Union. In this vein, DG CONNECT is funding a set of initiatives, Open Data Support being one of them, for creating the basis for pan-European portals for open data.

The problem that Open Data Support is dealing with is the limited accessibility and the lack of (cross-border) awareness of open datasets published on national, regional and local data portals of European Member States. This has a negative impact on the reuse of these datasets, which remains quite low – certainly beyond expectations.

OpenDataSupport has identified the following benefits of using linked data:

- Allows for flexible integration of datasets from different sources, without needing the data to be moved.
- Fosters the reuse of information from reference/authoritative sources.
- Caters for assigning common identifiers in the form of HTTP URIs to things (e.g. people, products, business, locations...).
- Provides context to data – richer and more expressive data.
- The use of standard Web interfaces (such as HTTP and SPARQL) can simplify the use of data for machines.

But OpenDataSupport also identifies the following considerations for publishing Linked Data that the SmartOpenData infrastructure will address:

- Linked Data is high-quality data. Considerable data cleansing and curation is required.
- Managing the data lifecycle is a challenging task. Mechanisms for handling updates and deletions in the data should be devised.
- The tools and software supporting linked data solutions are still not at production level/quality.
- A central authority should take the responsibility of publishing and maintaining persistent HTTP URIs for data resources. Existing identifiers should be reused to the extent possible, especially the ones coming from reference data sources, such as company registers.

---

22 See [http://www.slideshare.net/OpenDataSupport/introduction-to-linked-data-23402165](http://www.slideshare.net/OpenDataSupport/introduction-to-linked-data-23402165)
e) Data is currently available under different licences and in most cases no licence actually exists. This hampers data reuse and integration. Possible licensing options for data and description metadata should be explored. The use of open licence, e.g. a public domain licence – CC0, is recommended, particularly for the metadata.

f) Alternative business model for publishing linked data should be further explored. The costs and benefits of the different alternatives need to identified, before governments can decide on the adoption of the linked data technological paradigm.
2 The SmartOpenData Infrastructure

SmartOpenData will be a Spatial Linked Open Data infrastructure (including software tools and data) fed by public and freely available data resources, existing sources for biodiversity and environment protection and research in rural and European protected areas and its National Parks. This will provide opportunities for SMEs to generate new innovative products and services that can lead to new businesses in the environmental, regional decision-making and policy areas among others. The value of the data will be greatly enhanced by making it available through a common query language that gives access to related datasets available in the linked open data cloud. The commonality of data structure and query language will overcome the monolingual nature of typical datasets, making them available in multiple languages.

SmartOpenData will be a sustainable Spatial Linked Open Data infrastructure to promote environmental protection data sharing among public bodies in the European Union. The SmartOpenData infrastructure will make environmental open data easy to access by:

- Using Linked Open Data for modelling, acquiring, harmonising and using data provided by sources from existing catalogues and open public data portals: GEOSS Data-CORE, GMES, INSPIRE, EC Open Data Portal, data from EC projects in FP7, eContent+ and CIP, Global Forest Observation Initiative and ISA Programme. As a result it will be easier to apply the data within specific environmental protection solutions. The infrastructure will foster integration among sources and Linked Data with regards to knowledge representation, data modelling and interfaces for publishing the resulting information, as well as contributing to the adoption of GMES, intended to be operational from 2014, and the GEOSS 2012-2015 Work Plan.

- Applying Big Data mechanisms using the SINDICE platform and technologies in order to allow the management and fast access to the large volume of data (including real-time data) generated in environmental areas with adequate protection.

- Providing 5 Star Linked Open Data according to Tim Berners-Lee’s classification and preserving data quality through best-practice guidelines, QA procedures, IPR issues, Certification and Reputation for Government Open Data modelling.

The infrastructure will enhance Linked Open Data with semantic support by integrating semantic technologies built upon connected Linked Open Data catalogues aiming at building sustainable, profitable and standardised environment protection and climate change surveillance services. This will cover the use of semantic technologies to build a new paradigm of environment protection services through the extensive use of Linked Open Data, significantly improving their accuracy, power and scope and reducing implementation costs making them affordable and sustainable for the first time.

The project will research the integration of semantic results among public open data sources available by partners and other available public data including INSPIRE, GEOSS and GMES and external semantic services such as DBPedia\(^{23}\), Freebase\(^{24}\), GeoLinkedData\(^{25}\) and the

\(^{23}\) http://dbpedia.org/About
\(^{24}\) www.freebase.com
\(^{25}\) http://datahub.io/dataset/geolinkeddata
recently released Open Data Portal by the Commission\textsuperscript{26} including valuable datasets by the EEA and Eurostat, in order to enrich environmental services, thus making available geographical and environment data also to other application and service domains. The project will research how the models and Linked Open Data components can contribute to the development of standards for the geospatial aspects of Linked Open data. Within this area, an important point will be contributing to the Shared Environmental Information System (SEIS) of the European Environment Agency\textsuperscript{27}.

The SmartOpenData infrastructure, as it will be used in the project, is depicted in the following figure.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{SmartOpenData_INF.png}
\caption{SmartOpenData Infrastructure}
\end{figure}

In the lower level the external data sources are depicted. Data sources can be grouped in two different sets.

1. The first one is composed by data sources that fulfil some of the standards supported by SmartOpenData (green boxes).
2. The second group is composed by data sources that do not fulfil those standards (blue boxes).

In the upper layer, three different scenarios have been identified:

   a) Scenario for researchers,
b) Scenario for companies  
c) Scenario for end-users.

Each scenario will focus on one specific segment using the functionalities provided by the SmartOpenData System, creating services that take advantage of such data and provide valuable services for each community illustrating how the availability of such services and the corresponding data can provide advantages for them.

Between the external data sources and the data consumer in the scenarios the SmartOpenData System provides key functionalities.

1. The most basic element of the SmartOpenData System is the **harmonisation** of data sources. This element offers an open data source layer that exposes the external data sources fully adapted to the open data standards supported by the project.

2. **Adaptation** specifically tuned for each external data source that does not provide the information according to the required standard. This is depicted in the figure as an extra box.

3. The **open data source layer** provided both semantic information of the data and data themselves. Over this open data source layer, three key functionalities are defined:
   (i) distributed semantic indexing, which provides a service for searching and locating data based on semantic information collected from all the available Data Sources;
   (ii) distributed data access, which provides data collected from external data sources, as an extra data source for easier and uniform data gathering from the users at the identified scenarios;
   (iii) administration and notification, which provides administration facilities for managing users, workflows and data to data providers.

These three functional components are coordinated inside the SmartOpenData System, creating a distributed service system which can be accessed transparently from the scenarios. It is also important to note that it will be possible for services created on the scenarios to access directly external data sources selected through the distributed semantic indexing functionality of the SmartOpenData System if they are provided using one standard as shown on the diagram.

Using this framework the project will be able to define business models specially focused on SMEs and based on innovative services as new opportunities to align research results, previous work and projects, tackling active involvement of the whole value chain in Smart Regions at policy, industry and society levels.

There is a variety of projects, initiatives and research results focused on environmental protection, so SmartOpenData will need to add to this by building new collaboration bridges among them through the establishment of semantic services, aiming at turning individual environmental innovation initiatives into a collective set of meaningful seamless policies with exponential impact, that eventually will bring more open data shared by new engaged public bodies (i.e. ESA, FAO, UNEP, etc.).

To embrace this collaboration, SmartOpenData will include a community **Governance Model** that will offer contributors and participants guidance in deploying innovative services and environment-friendly business provided in a safe and secure manner with a common legal framework. The Governance Model will be particularly sensitive to SMEs in order to promote their prosperous participation. The Governance Model will aim to establish long-term
relationships and alliances between potential stakeholders and leverage new synergies in protected areas between policy makers, citizens and companies—with special focus on SMEs—for example through the creation of public-private partnerships.

Potential alliances will aim to set up a full ecosystem of initiatives supported by Linked Open Data, covering

(i) other public-private partnerships like Parks and Benefits,
(ii) case-studies such as Forest of Bowland AONB,
(iii) LIFE projects,
(iv) Natura 2000 Network,
(v) Eurosite and Europarc
(vi) Research projects like Humboldt, Habitats, EnviroGrids BlackSee, Plan4all, Plan4business, Hlandata, Brisede, GENESIS, EuroGEOSS, and GEO AIP3, TaxonConcept, PlanetData, among others.

Based on these the SmartOpenData infrastructure and system will need to be able to demonstrate the impact of the sharing and exploiting data and information from many varied resources, in rural and European protected areas by providing public access to the data and developing demonstrators that will show how services can provide high quality results in regional development working with semantically integrated resources.

The system will need to support a bottom-up and user-driven approach with a strong initial focus on rural regions, forestry and environmental protected areas such as National Parks and Smart Regions with the aim to be able to

(i) cope with wider European areas of interest
(ii) have a broader impact by eventually involving any area with relevant data subject to join the open infrastructure, models and governance proposed.

The infrastructure and system will need to be able to demonstrate its impact through enabling the implementation of the 5 pilots described in Annex B, driven by strategic partners in Spain, Italy, Ireland, Czech Republic and Slovakia. Through these pilots the consortium will be able to cover several co-created public services as pilots for environment protection and research, such as Biodiversity protection in different European areas, eGovernment: common policy framework for the management of protected environmental areas by public administrations, Environmental Research in biodiversity and protected/invasive species, sustainable exploitation of natural resources and land use (wood, biodiesel), Climate change, emergency management and environmental alerts, Training/Learning apps. for researchers, tourism, etc., voluntary engagement in content creation, enrichment, sharing, etc., services for farmers and forest owners in sensitive areas, and services for citizens.

The SmartOpenData infrastructure will need to cover information and services for these 5 initial pilots and will demonstrate semantic services and relationships among them. In order to contribute to strengthening the role of public sector in environmental engagement collective awareness, the project will work to house more pilots, internal or external to the consortium, according to the Governance Model defined.
2.1 Open Source Licence Model

SmartOpenData will follow a licensing model compatible with the exploitation plans of the industrial partners of the consortium. The partners agree on providing an open reference implementation of the SmartOpenData infrastructure. The selected licence is Apache-2\textsuperscript{28}, since this licence:

(i) explicitly grants patent rights where necessary to operate, modify and distribute the software;
(ii) permits code that it covers to be subsumed into closed source projects;
(iii) is suitable for safeguarding the IPR of project results as well as EC investments.

Standardisation and open source are key enablers for fostering competition in the implementation of standards while enabling interoperability across vendors without vendor lock-in concerns. SmartOpenData partners will preserve their investments and will make their proprietary services compliant with the proposed open source reference implementation, and will remain on proprietary developments, according to their exploitation plans.

These developments will provide added-value solutions for the reference implementation. The access to these developments will be granted to the rest of the project partners during project life, and joint exploitation plans on these developments will be analysed in the project Exploitation plan in WP7.

\textsuperscript{28} Defined at [http://www.apache.org/licenses/LICENSE-2.0.html](http://www.apache.org/licenses/LICENSE-2.0.html)
3 Relevant SmartOpenData components in the scenarios proposed

The SmartOpenData project is taking a bottom-up and user-driven approach with a strong initial focus on rural regions, forestry and environmental protected areas such as National Parks and Smart Regions with the aim to be able to

a) Cope with wider European areas of interest and
b) Have a broader impact by eventually involving any area with relevant data subject to join the open infrastructure, models and governance proposed.

Three different scenarios have been identified in the top layer of the SmartOpenData infrastructure:

1. Scenario for researchers,
2. Scenario for companies,
3. Scenario for end-users.

Each scenario will focus on one specific segment using the functionalities provided by the SmartOpenData System, creating services that take advantage of such data and provide valuable services for each community illustrating how the availability of such services and the corresponding data can provide advantages for them.

Demonstration of impact will be carried out through the implementation of 5 pilot scenarios in WP5, driven by strategic partners in Spain, Italy, Ireland, Czech Republic and Slovakia. Through these pilots the consortium will be able to cover several co-created public services as pilots for environment protection and research, including:

1. Biodiversity protection in different European areas
2. eGovernment: common policy framework for the management of protected environmental areas by public administrations
3. Environmental Research in biodiversity and protected/invasive species
4. Sustainable exploitation of natural resources and land use (wood, biodiesel)
5. Climate change, emergency management and environmental alerts
6. Training/Learning apps. for researchers, tourism, etc.
7. Voluntary engagement in content creation, enrichment, sharing, etc.
8. Services for farmers and forest owners in sensitive areas, and services for citizens.

The SmartOpenData infrastructure will cover information and services for these 5 initial pilot scenarios and will demonstrate semantic services and relationships among them. In order to contribute to strengthening the role of the public sector in environmental engagement collective awareness, the project will work to support more pilots, internal or external to the consortium, according to the SmartOpenData Governance Model that will be defined.

3.1 Characteristics of the SmartOpenData components

The broad strategic requirements of the SmartOpenData infrastructure and system include the following:

- To enable wide access to scientific data to allow researchers in different domains to collaborate on the same data sets:
SmartOpenData will need to provide a homogeneous Linked Open Data access framework for modelling, acquiring, harmonising and using geographic and environment data. This framework will thus allow researchers to perform cross-collaborations with public catalogues thanks to the semantic alignment with a wide range of catalogues and geoportals of INSPIRE, GEOSS Data-CORE, GMES, EC Open Data Portal, FP7 projects and a wide range of initiatives about environment protection, giving access in an easy, clear and affordable way to a new, better and greater amount of qualified data.

- To enable engagement in entirely new forms of scientific research and to explore correlations between research results:

  o Linked Open Data stimulates semantic correlation among repositories and data sources beyond organisational boundaries. SmartOpenData will need to cover this correlation in the environmental protection scope, unlocking new ways of research by leveraging the semantic use of scientific data: the huge amount of linked information made available will allow more powerful computation studies, data assessment and connecting public bodies’ researchers, companies and citizens in an improved effort for addressing research topics.

- To enable the use of models, innovative environmental tools and information products, based on accepted standards:

  o Linked Open Data and semantic technologies applied to geographic and environmental information leverages and demands a new set of advanced solutions for public and private sector, researchers and final users. SmartOpenData will need to be able to address a representative set of these solutions through the pilots that will be deployed during the project, i.e. biodiversity protection, eGovernment, Environmental research, sustainable exploitation, climate change or Emergency management. The project will also aim at improving spatial data visualisation and interfaces for publishing results according to user requirements.

- To enable delivery of benefits to researchers, European end-user agencies, the industrial sector, policy makers, and citizens, across environmental knowledge domains:

  o the SmartOpenData framework will need to directly benefit the value chain of environment protection by tackling effective and cost saving semantic services: According to the OECD Principles and Guidelines for Access to Research Data from Public Funding, making access easier and more affordable increases the returns from public investment and reinforces open scientific inquiry. The project will also leverage the establishment of public-private partnerships and user-centric services for engaging citizens in environmental protection policies.

- To ensure seamless interoperability of data catalogues:

  o The experiences from previous projects such as EnviroGRIDS and implementation of the Czech and Slovakian INSPIRE geoportal demonstrated

29 Available at http://www.oecd.org/sti/sci-tech/38500813.pdf
30 http://envirogrids.net/
serious concerns with services in different catalogues, with more than 60% of GEOSS registered services not working according to standards. In this respect, SmartOpenData will need to have a strong focus on allowing semantic interoperability among geographic and environment Open Data catalogues mentioned above.

- SuperCat offers now seamless access to INSPIRE services.
- HSCat is harvesting metadata from many European INSPIRE portals, and GEOSS portals.
- Linked Open Data and the semantic technologies approach will support fluent and flexible interaction among repositories.

At the same time, new catalogues and data sources will be included without additional effort, as long as they are Linked Open Data compliant to Tim Berners-Lee's classification and the SmartOpenData Governance Model. The W3C's DCAT standard and the Open Knowledge Foundation's application profile of that vocabulary for use in its CKAN software provide important reference points.

- SmartOpenData will need to advance the state of the art in the creation of a sustainable Linked Open Data infrastructure in order to promote environmental protection data sharing among public bodies in the European Union
  - Integration and interoperability has been a common concern in geospatial data management in recent years, to which Linked Open Data may be the solution. In this respect, INSPIRE provides Linked Open Data connections through Legal Acts, but GMES and GEOSS lack these kinds of models. Furthermore, the Linked Data approach is not yet fully embedded in INSPIRE. Other initiatives like the Geo Linked Data initiative\(^\text{31}\) has researched how to relate semantic and geospatial depictions of a real world entity in the Semantic Web. LinkedGeodata\(^\text{32}\) also followed a similar approach by using OpenStreetMap data. Other initiatives are UK Ordnance Survey OpenData\(^\text{33}\) and GeoNames\(^\text{34}\). At the same time, GEOSS has also explicit goals for sharing environmental data by services linked together (GEO data-sharing principles). However, though some initiatives have started and some geospatial data catalogues have been exposed with Linked Open Data technologies, they are still incomplete and very focused on specific and divergent domains, lacking mechanisms and standards for correlation amongst them.

At the framework and services level, the United Nations initiative on Global Geospatial Information Management (GGIM)\(^\text{35}\) stated the need of developing common frameworks and methodologies “to promoting common principles, policies, methods, mechanisms and standards for the interoperability and interchangeability of geospatial data and services”. Addressing this,
SmartOpenData proposes to advance in the state of the art building integration and semantic interoperability amongst open public data portals, INSPIRE, GMES and GEOSS large data catalogues, providing a sustainable framework for building environment protection services.

- SmartOpenData will need to support the UNSDI (United Nation Spatial Data Infrastructure) vision:\(^{36}\)
  - to build a comprehensive, decentralised geospatial information framework that facilitates decision-making at various levels by enabling access, retrieval and dissemination of geospatial data and information in a rapid and secure way, avoiding duplication in data collection and management within the United Nations, and with and between its member states and partners.

- SmartOpenData will need to enhance Linked Open Data with semantic support by integrating semantic technologies built on connected Linked Open Data catalogues aiming at building sustainable, profitable and standardised environment protection and climate change surveillance services
  - Management of Geospatial information is a worldwide concern. In this respect, GGIM has noted the importance of increasing Linked Data adoption at mid-term. Thus, semantic technologies are emerging as a key point for solving current issues of this management. The development of standards for the geospatial aspects of Linked Open Data is an open concern in OGC (ISO/TC211) as INSPIRE stakeholders, for which an ad hoc group for analysing Linked Open Data for geographical services was constituted. Within this, the Delft Report\(^ {37}\) notes several challenges not covered at this moment by ISO/TC 211 in terms of compatibility with Linked Open Data. For instance, the Tutorial “Lift your data” presented at the Istanbul conference remarks how INSPIRE data is only 2+3 stars in the Linked Data scale and the need for links to geo features from other entities on the web. There is much important work ahead to bring the standardisation of semantic technologies to the level of enabling integration of GMES, GEOSS and INSPIRE catalogues. Some of this work has already begun at W3C which enjoys good relations, both formal and informal, with OGC, ISO and other standardisation bodies. All of these advantages are available to SmartOpenData through ERCIM, the host organisation of W3C.

- SmartOpenData will need to go beyond the state of the art leveraging the adoption of new standards regarding
  - alignment and compatibility of semantic technologies,
  - geospatial data interoperability,
  - new RDF-based vocabularies for integrated location description and
  - Open Data Quality, Certification and Reputation.
  - This work will involve the semantic harmonisation of different datasets of INSPIRE -INSPIRE compliance does not guarantee harmonisation-, GEOSS and GMES and will be consistent with the ISO/TC 211. Where no such standard currently exists, suitable work will be promoted at W3C.

\(^{36}\) [http://www.ungiwg.org/content/united-nations-spatial-data-infrastructure-unsdi](http://www.ungiwg.org/content/united-nations-spatial-data-infrastructure-unsdi)

- By bringing semantic technologies to the environment protection area as an umbrella for public data portals and INSPIRE, GEOSS and GMES datasets, SmartOpenData will need to provide a new set of services that efficiently handle the large amount of real-time environment data produced, its quality and reputation, leveraging lower response times and a much more effective surveillance of climate change.
  o In addition, SmartOpenData will contribute to leverage Linked Open Data and semantic compatibility in ISO/TC 211, contributing to LOD-INSPIRE convergence promoting improved standards for integrated location description and environmental data interoperability.

- SmartOpenData will need to enable the definition of business models specially focused on SMEs and based on innovative services as new opportunities to align research results, previous work and projects, tackling active involvement of the whole value chain in Smart Regions at policy, industry and society levels.
  o Promoting the adoption of new policies regarding environmental Open Data sharing in public authorities brings a new set of collaboration opportunities with other public, private and research sectors, allowing the emergence of a new dimension of services regarding environmental protection and sustainable use. In this regard, the Delft report\textsuperscript{38} states that “the addition of Linked Data to the geographic standards may produce effective cost savings in spatial data production and use less investments.” However, this kind of service needs a new collaboration framework in which subjects such as (i) Licensing harmonisation, (ii) IPR issues and (iii) Rights management have to be addressed in order to really allow a faster and easier access and use of Open Data.

- SmartOpenData will need to advance the state of the art by researching a framework for use of public Open Data by means of consistent Open Government licenses covering both IPR, Rights management, Certification and Reputation issues.
  o The aim is (i) to establish a clear set of policies that fully leverages the rise of fruitful public-private partnerships and, (ii) having in mind the simpler and faster discovery and use of environmental data, also smooth the way for establishing simpler and faster collaboration policies. During the project the system will need exploration of research indicators, such as new PPP policy framework tested through the SmartOpenData infrastructure and pilots.

- SmartOpenData will need to be able to demonstrate the impact of sharing and exploiting data and information from many varied resources, in rural and European protected areas by providing public access to the data and developing demonstrators that will show how services can provide high quality results in regional development working with semantically integrated resources during the project.
  o In recent years there has been a significant increase in the number of initiatives in the environmental protection area. In general, these initiatives have supported actions in very focused yet disaggregated scopes, resulting in very attractive efforts that, however, could not be easily replicated and fully adopted in other analogous geographical areas.

\textsuperscript{38} The Delft Report: Linked Data and the challenges for geographic information standardization, Revista Catalana de Geografia IV época / volum XVII / núm. 44 / febrer 2012, \url{http://www.rcg.cat/articles.php?id=220}
- SmartOpenData will need to be able to test the semantic technologies developed in 5 pilots that will be fully interoperable at data, metadata, semantic and legal levels.
  - To enable this, the project will build an architecture of linked data clouds leveraged by SINDICE. The objective is to achieve a generic approach, managed under a clear Governance Model, valid not only at pilot level but including interoperability principles that support and boost the adhesion of other initiatives and pilots during the project. And explore research indicators such as public geoportal available covering the 5 interoperable pilots developed into a single access point and including a Governance Model, framework and procedures allowing at least 5 other environment protection initiatives to join SmartOpenData.

### 3.2 Research issues that SmartOpenData needs to address

The research focus for the SmartOpenData project will address how to use existing GI data within an RDF framework, or alternatively, how existing GI data can be accessed as part of linked data. To achieve this, the project will develop algorithms that expose the wealth of environmental data held by the partners as linked data. This may require some human intervention in some cases but such intervention will be minimised with a view to making it repeatable and scalable. For example, the Open Refine tool can be carried out on tabular datasets of unlimited size and is likely to be useful in this task, perhaps supported by a SmartOpenData reconciliation API. In a linked data environment, the definition of points, lines and polygons remains untouched but the relationships between features, the names of places and, in particular, the identifiers, are handled differently. Separating those elements out and encoding them as linked data, and doing so at scale, will be a significant challenge.

Creating the data as RDF and adding it to a triple store is only the first step, however. More difficult is the discovery of links to data already available in the linked open data cloud, such as GeoSpecies.

Challenges that SmartOpenData must address in Open Data include:

- **Discoverability**
  - In order for data to be useful, it first needs to be discoverable. Building strong catalogues of metadata from numerous sources is one of the best ways this can be achieved. In addition, building in referenceability into data catalogues is also an effective way of tracking how data propagates through different work products from a raw dataset to an API to a mashup, etc.

- **Federation**
  - As open data becomes part of the day to day business of these organizations, the work of cultivating, publishing, and maintaining datasets and data catalogues will become more decentralized. These decentralized catalogues still need to be aggregated into combined organizational catalogues, but their maintainers should still be able to pick and choose which open data

40 [http://lod.geospecies.org/](http://lod.geospecies.org/)
41 Adapted from "Our Next Challenge: Open Data Platforms That Talk to Each Other" Chris Metcalf, Socrata, [http://www.w3.org/2013/04/odw/odw13_submission_59.pdf](http://www.w3.org/2013/04/odw/odw13_submission_59.pdf)
technologies are most appropriate to their needs. In time, these data
catalogues will stop resembling the big, centralized catalogues being built
now, and will morph into an ecosystem of distributed, federated data
catalogues that need an effective and efficient way for catalogues to be
propagated for discovery.

- Interoperability
  - As catalogues from multiple sources are composed by federation, it becomes
    more and more important for the platforms that these data catalogues on
    which they are built be compatible, even if they are built by different
    providers.

SmartOpenData must work together with and help to develop build cross-platform data
catalogue standards that are implemented consistently between different catalogue
platforms such as those from open data platform providers such as CKAN, Junar, and
Socrata\textsuperscript{42}. Developing practical, portable data catalogue standards will:

- Allow data catalogues to be easily federated and shared between organizations. This
  will, for example, allow a parent organization to easily consume and aggregate
catalogues from different sub-organizations, regardless of what software those
  catalogues are hosted on.
- Allow catalogues to be portable between providers. As organizations change and
grow, they may also want to move from one provider to another. Standardized
metadata schemas make that a more straightforward process.
- Allow the catalogue to track reuse and propagation of data sets through the
catalogue. By allowing catalogue entries to be related to one another, the catalogue
will morph from a simple listing of data sets into a network of related entries,
supporting reuse and discovery.
- Allow data catalogues to be consistently indexed by search providers. Google and
other search engines will be able to consume structured data catalogues and provide
better insight and discovery into what data sets are available.

Work is already under way on efforts to build these standards. Through projects like DCAT
(Data Catalogue Vocabulary)\textsuperscript{43}, Schema.org\textsuperscript{44}, and the Open Data Substrate\textsuperscript{45}, headway
is being made towards making the hard decisions necessary to build the required
interoperability standards.

However simply having standards is not enough. The world doesn’t need more
unimplemented standards. The best standard is the one with actual running, production
level code. So SmartOpenData must solve the real-world problems, and hopefully walk away
with a more connected open data ecosystem that includes both the open data and GI
worlds.

\subsection*{3.2.1 Current Shortcomings to be addressed by SmartOpenData}

To a linked data engineer, geospatial data is a lump of indigestible literal. To a geospatial
engineer, linked data is a crazy open world that just doesn’t follow the rules. The most

\textsuperscript{42} These are discussed in section 5.
\textsuperscript{43} http://www.w3.org/TR/vocab-dcat/
\textsuperscript{44} http://schema.org
\textsuperscript{45} http://open-data-standards.github.io/
advanced effort to reconcile the two worlds is embodied by OGC's GeoSPARQL standard. This merges the two technologies, with the GeoSPARQL engine translating queries back and forth between RDF and geospatial engines. The number of implementations of GeoSPARQL is growing but there remains some debate as to whether it is the best approach. The NeoGeo vocabulary is favoured by French mapping agency IGN and handles geospatial data differently by linking to it from the RDF, rather than transporting large literals. The INSPIRE standards have been developed entirely in an XML-centric manner and the European Commission's JRC is now working on making better use of Linked Data. This is being done in a W3C Community Group focussing on locations and addresses. A related, but separate, Community Group is also considering better interplay between Web and geospatial technologies.

What these activities all suggest is that SmartOpenData has work to do to allow geospatial and linked data specialists to communicate easily, avoiding the so-called religious wars. SmartOpenData brings together specialists in both disciplines: RDF to describe a location or point of interest, GI to define where it is on the Earth's surface.

3.2.2 Machine translation in Spatial Data and Metadata

There are two principal approaches to machine translation: rule-based and statistical. Current state-of-the-art machine translation (MT) technology is based on the SMT (statistical MT) paradigm, which assumes the application data to match the training data, used during the learning phase to extract and generalise the parameters of the system. Combined methods are also being investigated currently, bringing together the linguistic and translation knowledge accumulated over the last 40 years with the SMT systems as deployed today. For SMT systems, the more distant the actual data is from the data used for training, the worse the results are. As SmartOpenData is concerned with environmental and geographical data, the project will explore resource-limited adaptation to those domains.

Machine translation Language Resources (LR) mainly include:
- Lexicons and bilingual dictionaries giving correspondences between words or sequences of words in different languages. Here the project will make use of the GEMET\textsuperscript{46} and AgroVoc\textsuperscript{47} thesauri.
- Parallel corpora: large amounts of parallel texts.

The lack of LR coverage constitutes a severe limitation for the development and proper functioning of language technologies and particularly for MT systems. Moreover, LRs cannot be ever considered complete for any particular language because of language change, domain tuning and neologisms that are characteristic properties of human languages. The current solution to the problem of language resource coverage in industrial applications is by means of a continuous supply of hand-made components. Thus, updating, tuning and maintenance of language resource components (annotated texts, lexica, language models, etc.) is the most laborious and expensive part of NLP-based applications such as MT, in both statistical and rule-based paradigms. The acquisition of resources for each domain and each language is very expensive in terms of human resources because it has to be conducted mostly manually by highly skilled people. Some components are already available (text

\textsuperscript{46} www.eionet.europa.eu/gemet/
\textsuperscript{47} http://aims.fao.org/standards/agrovoc/
processing and text annotation package tools as FreeLing\textsuperscript{48}, CorpusBootstrap\textsuperscript{49}, TreeTagger\textsuperscript{50}, or Aligners such as Giza++\textsuperscript{51}). Other components will need to be developed and made available as open source software during the project (lexical acquisition technologies, and some parallel corpora processing components). SmartOpenData will reuse existing Open Source software packages, like for example Moses\textsuperscript{52}. Two types of components will be of particular interest for the project:

- Parallel corpus technologies that concern the handling and induction of resources from parallel texts;
- Lexical acquisition technologies where a key challenge remains to produce new lexicons or tune existing lexical resources for domain specific tasks.

An important challenge will be the integration of MT modules with spatial (data and metadata services). It will be necessary to develop a specific proxy mechanism allowing the caching of data and external services (For external services are not considered WMS and WCS, due the fact that these services transfer only images). This proxy will be connected with an MT module, providing translation. For client applications data will be available through standard OGC services. For this purpose SmartOpenData will apply experience with HSProxy\textsuperscript{53}.

Through ERCIM, as a hosting organisation of W3C, SmartOpenData has excellent links with the JRC INSPIRE team and so the project outcomes will feed directly into the ongoing effort to make use of the relevant standards on the semantic web.

### 3.2.3 Large Volumes of Real Time Data.

Another area of research for the project will be the handling of large volumes of real time data. Activities of partners within the consortium include real time monitoring of environmental conditions that generate data that needs to be added to the dataset using SPARQL Update. This puts a strain on the infrastructure and so methods to reduce that stress will need to be researched, perhaps using, through ERCIM, the W3C POWDER Working Group\textsuperscript{54} recommended technology as a data compression tool, as currently being tested in the SemaGrow project\textsuperscript{55}. Tracking the provenance of any data is important, but as yet there is no (standardised) linkage within the Semantic Web technology stack between Provenance and SPARQL Update.

### 3.2.4 Multilingualism

Another important problem that SmartOpenData will need to address is multilingualism. The problem of translating geographical data and metadata has not yet been solved inside INSPIRE or GEOSS. It brings problems of global data utilisation by local communities and local data by foreigners. Translation of geographical data is a big challenge for everyone within the SDI community and its importance will grow in relation with growing of SDI. The

---

\textsuperscript{48} \url{http://nlp.lsi.upc.edu/freeling/}
\textsuperscript{49} See \url{http://text-processing.com/}
\textsuperscript{50} \url{http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/}
\textsuperscript{51} \url{http://code.google.com/p/giza-pp/}
\textsuperscript{52} \url{www.statmt.org/moses/}
\textsuperscript{53} \url{http://hslayers.org/introduction/proxy.html}
\textsuperscript{54} \url{http://www.w3.org/2007/powder/}
\textsuperscript{55} \url{www.semagrow.eu}
implementation of RDF should help ease the translation of geographic names or keywords from vocabularies like GEMET or AgroVoc.

3.3 Required impacts of SmartOpenData

The European Environment Agency (EEA) states that Europe’s protected areas should play a key role in protecting biodiversity and in Europe’s economy, contributing over EUR €15B a year in jobs, food and other services for the people of Europe.

Currently there exists many initiatives at European level in the environmental protection and climate change domain. However, these initiatives, even if aligned to standards, remain focused on trying to solve very specific and location-dependent issues. In this regard, SmartOpenData will need to enable the integration of currently decoupled systems and building new services by means of interoperability through Linked Open Data and Semantic technologies.

- SmartOpenData will support the creation of opportunities in global markets, new businesses, jobs and growth, through the delivery of a wide range of innovative products and services. Starting from the initial idea that the key step for running the wheel that turns outcomes into impact is to truly engage public bodies at European level in sharing valuable, quality and updated Open Data on a regular basis. To achieve this, public bodies and governments have to go beyond “sharing as the goal” and foresee a real impact in this sharing. Instead of just a pure volunteer form or something that “has to be done”, to walk the path of “Publish once, use many times”.

- SmartOpenData will need to have real engagement of public bodies through the establishment of a Governance Model that leverages new business and public-private partnerships giving benefits to the whole chain of environment protection services. Alternatives for these businesses are rich and diverse and will be studied within the project. Succeeding in the inclusion of public bodies is the key to establish a virtuous circle around incremental sharing of quality Open Data to allow for offer and demand to meet, truly allowing pilot providers to exploit their results and prompting the commercial sector to create value added products and services, thus creating a new global environmental protection market with final user involvement.

- SmartOpenData will enable improved public services, e.g. the provision of information to address crisis situations and disasters with cross-boundary impact SmartOpenData will achieve real impact on cost savings in the deployment of Linked Open Data services. This idea is supported by The Delft report\(^\text{56}\): “The addition of Linked Data to the geographic standards may produce effective cost savings in spatial data production and use by improving some issues relevant to Spatial Data Infrastructures (SDI)”. Regarding disaster management, SmartOpenData can widen the GEOSS information level scope by means of new interoperable disaster and crisis services that go beyond GEOSS and also integrate INSPIRE, GMES the many Linked Open Data initiatives.

- SmartOpenData will need to foster integration between INSPIRE and Linked Data, making LOD implementations in the EU dealing with location follow the semantics and ontology of the INSPIRE Legal Acts, thus helping to address the existing “discrepancies between INSPIRE and LOD with regards to knowledge representation and web service interfaces”

- SmartOpenData will ensure greater transparency in public administration through the improved visibility of information, informing citizens and business about policies, public spending and outcomes. Open Data is a motor and main driver to public sector transformation, increasing awareness, engaging stakeholders and encouraging reuse of data, thus enabling the commercial sector to use this data to create better products and services. Environmental agencies are aware of this and are increasingly going open data as part of a global transparency effort.

- Beyond Open Data, SmartOpenData will apply Linked Open Data technologies to build more powerful environmental protection services in the short term, as an opportunity to take more efficient and cost saving decisions and working practices. Paradoxically, evaluating the impacts of these policies requires robust information on a range of social, economic and environmental conditions. That is one of the main reasons why SmartOpenData understands that using Linked Open Data for environment protection is a key factor in order to start and accelerate the engines for a real transformation of environmental information systems. For instance, regarding geographical information, SmartOpenData will deliver real benefit from INSPIRE and Linked Open Data sharing, promoting open government policies.

- SmartOpenData must empower citizens and citizens' associations, enabling them to contribute to environmental governance processes in the domains of transparency, knowledge management, accountability and responsiveness. Linked Open Data highlights the need to adapt services to fit the real needs of final users, whether service developers, policy makers or common citizens. Benefits of Linked Data for users seem obvious; for instance, they can build mash-ups and applications on top of INSPIRE data more easily. Real impact of environmental protection policies has to extend this approach. SmartOpenData will need to go further because, apart from fitting user needs, it is necessary to tackle an essential factor for a successful impact in the area of environment protection, which is user engagement.

  In the end, only user engagement guarantees the continuity and growth of the environmental policies applied at European level and the environment sustainability itself. Achieving real user engagement will empower the further adoption of Open Data policies, closing the virtuous circle. In other words, Open Data sharing and semantic services become requisites; final user intrinsic motivation and response to calls to action is the success and the real impact.

  Building engagement is also building trust. Thus, user privacy and ethical issues will need to be primary concerns and will be taken into account from the very beginning of the activities.

  The SmartOpenData user-driven engagement approach will be sparked off by the 5 pilot service scenarios and their ability to attract new participants, being (i) new user communities and (ii) existing user associations from other projects and initiatives.
The degree of maturity of the user communities and stakeholders of the HABITATS project is an excellent starting point for this.

- **SmartOpenData** will encourage and enable cooperation with other national and international research activities. Specifically, all beneficiaries of the SmartOpenData project will adhere the GEOSS Data Sharing Principles:
  - There will be full and open exchange of data, metadata and products shared within GEOSS, recognizing relevant international instruments and national policies and legislation;
  - All shared data, metadata and products will be made available with minimum time delay and at minimum cost;
  - All shared data, metadata and products being free of charge or no more than cost of reproduction will be encouraged for research and education.

Consequently, the consortium will undertake the registration of the components resulting from the project in the GEOSS.

The approach of SmartOpenData brings a wide range of opportunities for cross-collaborations with public catalogues and geoportals of INSPIRE, GEOSS Data-CORE and GMES. Also EC Open Data Portal, FP7 projects and a wide range of initiatives about environment protection brings a promising horizon of collaboration, as well as lean and support the results and outcomes obtained within these initiatives.

The broad experience of the project partners will enable the establishment of an important framework of collaboration.

- **SmartOpenData** must contribute to standards. The standards development process, particularly at W3C, is very much centred on creating consensus among an interested group of individuals and companies. New W3C Recommendations are created by working groups that are chartered through a process in which W3C members are crucial. A more lightweight mechanism, Community Groups, offers a forum for discussion and can publish its own documents. These will be exposed to the Web community and may then lead to the creation of a Working Group in future. As a minimum therefore, SmartOpenData must publish Community Group documents by ERCIM, particularly around the data model. This will be strongly influenced and informed by the wider Web community. As a maximum, the project may lead to the development of new formal standards.

ERCIM, as part of W3C, has excellent relations with the Open Geospatial Consortium which is the primary standards body for geospatial data and is already actively cooperating in the Locations and Addresses Community Group, chaired by the JRC INSPIRE team. Besides, the general issue of handling geospatial data on the Web is an active one and the focus of attention, particularly among the search engines who are keen to offer location-based search and related services. In essence, OGC provides a range of highly respected and well implemented standards around geospatial data.

W3C is in the same position for the Web. The research undertaken within SmartOpenData offers real possibilities for future standards work by either or both standards bodies.

Factors to impact SmartOpenData’s achievement, will be determined by the following assumptions and external factors:

---

• Data shared by public bodies and final users:
  • Achieved impact will greatly depend on the amount and quality of shared open data by public bodies. Public bodies present in SmartOpenData guarantee an optimal starting point to build this engagement, from which the Governance Model will serve as a mechanism to encourage new public bodies to join allowing new businesses and public-private partnerships. Cost savings embracing SmartOpenData is a key factor to achieve a broad deployment of LOD services.

• User engagement:
  • User engagement is a key factor to get real impact. In order to promote it, SmartOpenData will set up and maintain a User group that will provide frequent feedback from early in the project, establishing a customer discovery process that will allow solutions to be built, that really fit the needs of the
    ▪ end users (public bodies, companies, researchers, citizens),
    ▪ the infrastructure
    ▪ services.

• Alignment with European environmental protection trends and standards
  • The SmartOpenData data model and infrastructure will be built upon and use existing standards regarding geographic, spatial, LOD and semantic standards.

• Open source ecosystem
  • The SmartOpenData impact will depend largely on the engagement of early adopters and the adoption of project outcomes by the community. In order to achieve this, the consortium must leverage third parties and its own customers to participate and benefit from SmartOpenData outcomes from the beginning of the project.

3.4 Specific requirements for each of the SmartOpenData Pilots

The 5 SmartOpenData pilots\(^{58}\) will be focused on different environmental scenarios, targeting different users, in various European areas as follows:

<table>
<thead>
<tr>
<th>Pilot</th>
<th>Scenario / Users</th>
<th>Public Sector</th>
<th>Research</th>
<th>Enterprises</th>
<th>Citizens</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Spain</td>
<td>Agroforestry Management</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Ireland</td>
<td>Environmental Research and Biodiversity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>3. Italy</td>
<td>Water Monitoring</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>4. Czech Republic</td>
<td>Forest Sustainability</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>5. Slovakia</td>
<td>Environmental Data reuse.</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 1 SmartOpenData Pilot Environmental Scenarios

Based on the previous discussion and a discussion amongst the partners and potential users, the following table maps the specific functional requirements of the SmartOpenData components for each of the pilot scenarios that will be implemented in the project, as described in Annex B.

\(^{58}\) These are described in Annex B.
<table>
<thead>
<tr>
<th>Required Functions / Pilot Scenarios</th>
<th>Agroforestry Management</th>
<th>Environment Rsch &amp; Biodiversity</th>
<th>Water Monitoring</th>
<th>Forest Sustainability</th>
<th>Environment Data reuse.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discovery &amp; Referenceability definition</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>View</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Distributed Data services</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Transformation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Analysis &amp; Analytics</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Monitoring</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>External services</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Applications</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>WMC</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>RSS/GeoRSS</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>KML/KMZ</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>CMS</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Public Geoportal access to all pilots</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Pilots interoperable at data, metadata, semantic &amp; legal levels.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Crowdsourcing/Co-creation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Share/engage thru Social Media</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Workflow management</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Smart Phone Apps</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Multilingual Access</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Multilingual data harvesting</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Data standardisation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Federation of decentralised data catalogues</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Suite of GeoSPARQL GeoSpatial LOD tools</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Efficient GeoSpatial RDF querying</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Fuse/aggregate GeoSpatial RDF data</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Selection of ontologies</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Harmonisation of data sources,</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Easy Visualisation &amp; Authoring</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Administration &amp; notification</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Repeatable &amp; scalable.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Easy data publishing/interfacing</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Impact measures &amp; evaluation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Governance &amp; Curating Management</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Mobile, Fixed &amp; Distributed</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Required Functions / Pilot Scenarios</td>
<td>Agroforestry Management</td>
<td>Environment Rsch &amp; Biodiversity</td>
<td>Water Monitoring</td>
<td>Forest Sustainability</td>
<td>Environment Data reuse.</td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>-------------------------</td>
<td>---------------------------------</td>
<td>-----------------</td>
<td>----------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>Access</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mashability/Web2 interactive</td>
<td></td>
<td>✅</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geoblogging</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data market</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSW interface</td>
<td>✅</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access to deep web</td>
<td></td>
<td>✅</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantic Search</td>
<td>✅</td>
<td>✅</td>
<td></td>
<td></td>
<td>✅</td>
</tr>
<tr>
<td>Semantic interoperability of decentralised sources</td>
<td>✅</td>
<td>✅</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Public-private geospatial data</td>
<td>✅</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Acquisition &amp; Organisation</td>
<td></td>
<td>✅</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Composition/Bundling</td>
<td></td>
<td>✅</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access control address user privacy &amp; ethical issues</td>
<td>✅</td>
<td>✅</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Licence, IPR &amp; Rights Management</td>
<td>✅</td>
<td></td>
<td></td>
<td></td>
<td>✅</td>
</tr>
<tr>
<td>Data cleansing / editing</td>
<td>✅</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preserve data quality</td>
<td>✅</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QA procedures/ 5 Star LOD</td>
<td>✅</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Certification &amp; Provenance</td>
<td>✅</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adaptation</td>
<td>✅</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Process large volumes of real time data</td>
<td></td>
<td>✅</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Use INSPIRE, GEOSS, GMES &amp; Geonames datasets</td>
<td>✅</td>
<td>✅</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2 Functions required in each of the Pilot Scenarios

Further useful case studies against which the SmarOpenData functionality will be tested and validated will be drawn from other sources such as the SMEspire Best Practice Catalogue, as listed in Annex D.
4 How the SmartOpenData components can be classified or generalised

Linked Open Data is emerging as a source of unprecedented visibility for environmental data and will enable the generation of new businesses as well as a significant advance for research in the environmental area. In order for this envisioned strategy to become a reality, it is necessary to advance the publication of existing environmental data, most of which is owned by public bodies. How Linked Open Data can be applied generally to spatial data resource and specifically to public open data portals, GEOSS Data-CORE\(^{59}\), GMES (now known as Copernicus)\(^{60}\), INSPIRE and voluntary data (OpenStreetMap, GEP-WIKI\(^{61}\), etc.), and how it can impact the economic and sustainability progress in European Environmental research and Biodiversity protection are open questions that need to be addressed to better benefit from an improved understanding and management of environmental data. The SmartOpenData project must address these questions by defining mechanisms for acquiring, adapting and using Open Data with a particular focus on biodiversity and environmental protection in rural and European protected areas and its National Parks.

The SmartOpenData vision is that environmental and geospatial data concerning rural and protected areas can be more readily available and re-usable, better linked with data without direct geospatial reference, so different distributed data sources can be easily combined together.

SmartOpenData will use the power of Linked Open Data to foster innovation within the rural economy and increase efficiency in the management of the countryside. The project will prove this in a variety of pilot programmes in different parts of Europe (in WP5). The SmartOpenData goal is to make INSPIRE/GMES/GEOSS infrastructures better available for citizens, as well as for public and private organization. On one hand, Europe and the EU invests hundreds of millions of Euros in building the INSPIRE infrastructure. On the other hand, public and private organizations, as well as citizens, use their applications based on Google maps. National and regional SDIs offer information which is not available on Google, but this potential is not used. One of the main goals of SmartOpenData is making European Spatial Data easily re-usable not only by GIS experts but also by various organizations and individuals at a larger scale. To realize this, on a technical level, the project will

1. harmonise geospatial metadata (ISO19115/19119 based) using the principles of the Semantic Web,
2. provide spatial data fusion introducing principles of Open Linked Data,
3. improve spatial data visualisation of Geospatial Open Linked Data and
4. publish the resulting information according to user requirements and Linked Open Data principles.

In the context of the SmartOpenData project, using linked data for spatial data means identifying possibilities for the establishment of semantic connections between

\(^{59}\) [http://www.earthobservations.org/geoss_dsp.shtml](http://www.earthobservations.org/geoss_dsp.shtml)

\(^{60}\) [www.copernicus.eu/](www.copernicus.eu/)

INSPIRE/GMES/GEOSS and Linked Open Data spatially related content in order to generate added value. The project requirements are within the environmental research domain. This will be achieved by making existing “INSPIRE based” relevant spatial data sets, services and appropriate metadata available through a new Linked Data structure. In addition, the proposed infrastructure will provide automatic search engines that will crawl additional available geospatial resources (OGC and RDF structures) across the deep and surface web.

The main motivation to utilise the potential of Linked Data is to enrich the INSPIRE spatial content to facilitate improved related services to be offered and to increase the number, performance and functionality of applications. In many cases querying data in INSPIRE (GEOSS) based data infrastructures (driven mainly by relational databases) is time consuming and often it is not sufficient and understandable for common Web users. In large databases such queries can take minutes or hours. In the cases of distributed databases such a query is almost impossible or very complicated. SmartOpenData aims to improve this situation dramatically.

The research focus for ¡Error! No se encuentra el origen de la referencia. will address how to use existing GI data within an RDF framework, or, from the other direction, how existing GI data can be accessed as part of linked data. To achieve this, new algorithms will be developed that expose the wealth of environmental data as linked data. This may require some human intervention in some cases but such intervention will be minimised with a view to making it repeatable and scalable. For example, the Open Refine tool62 allows the same operation to be carried out on tabular datasets of unlimited size and is likely to be useful in this task, perhaps supported by a ¡Error! No se encuentra el origen de la referencia. reconciliation API. In a linked data environment, the definition of points, lines and polygons remains untouched but the relationships between features, the names of places and, in particular, the identifiers, are handled differently. Separating those elements out and encoding them as linked data, and doing so at scale, will be a significant challenge.

Creating the data as RDF and storing it in dedicated triple stores is only the first step, however. More difficult is the discovery of links to data already available in the linked open data cloud, such as GeoSpecies. The example given on the GeoSpecies Web site63 shows the detail of the Cougar including links to where it can be expected to be found. It is links between datasets that makes linked open data so powerful and forging those links is an essential aspect of realising the objectives of ¡Error! No se encuentra el origen de la referencia..

4.1 Generic Components Interfaces

SmartOpenData will provide the following generic open data component interfaces similar to those defined in the SemaGrow project64:

- The SPARQL endpoint:
  - Receives from client applications SPARQL queries and reactivity parameters.
  - Receives from the Federated Endpoint Wrapper SPARQL query results.
  - Sends to client applications SPARQL query results.

---

62 http://openrefine.org (formerly Google Refine)
63 http://datahub.io/dataset/geospecies/resource/47e71c4c-9565-4185-b8c0-bdef6449278e
64 www.semagrow.eu and see section 5.
Sends to Query Decomposition SPARQL queries.
- Sends to Query Decomposition and to the Federated Endpoint Wrapper reactivity parameters.

- **Query Decomposition:**
  - Receives from the SPARQL endpoint SPARQL queries and reactivity parameters.
  - Queries Resource Discovery about candidate data sources, relevant to query patterns, with instance statistics, load info, and alignment proximity annotations.
  - Sends to the Federated Endpoint Wrapper a list of SPARQL queries matched against the endpoint they should be dispatched to.
  - Receives from the Federated Endpoint Wrapper a control signal regarding unavailable sources or other reasons to compute a different decomposition.

- **Resource Discovery:**
  - Receives run-time load information from the Federated Endpoint Wrapper.
  - Queries Data Summaries according to the query patterns received by Query Decomposition.
  - Queries the Schema Mappings Repository about known alignments between entities.
  - Is queried by Query Decomposition about candidate data sources.

- **The Federated Endpoint Wrapper:**
  - Receives from Query Decomposition a list of SPARQL queries matched against the endpoint they should be dispatched to.
  - Receives from the SPARQL endpoint SPARQL reactivity parameters.
  - Sends to Query Decomposition a control signal regarding unavailable sources or other reasons to compute a different decomposition.
  - Queries Schema Mappings Repository about known alignments between entities.
  - Queries the endpoints of the federation.
  - Sends to the SPARQL endpoint SPARQL query responses.

- **The Data Summaries endpoint:**
  - Is populated by the source metadata authoring tool.
  - Is queried with RDF resources ad responds with relevant endpoints and known alignments.

- **The Schema Mappings Repository:**
  - Is populated by the ontology alignment tool.
  - Is queried about known alignments between entities.

## 4.2 Multilingualism

An important problem to be addressed in the context of SmartOpenData is multilingualism. The problem of translating geographical data and metadata has not yet been solved inside INSPIRE or GEOSS. It brings problems of global data utilisation by local communities and local data by foreigners.
Translation of geographical data is a big challenge for everyone within the SDI community and its importance will grow in relation with the growth of SDI. The implementation of RDF should help ease the translation of geographic names or keywords from vocabularies like GEMET or AgroVoc.

4.3 Machine Translation

There are two principal approaches to machine translation: rule-based and statistical. Current state-of-the-art machine translation (MT) technology is based on the SMT (statistical MT) paradigm, which assumes the application data to match the training data, used during the learning phase to extract and generalise the parameters of the system.

Combined methods are also being investigated currently, bringing together the linguistic and translation knowledge accumulated over the last 40 years with the SMT systems as deployed today. For SMT systems, the more distant the actual data is from the data used for training, the worse the results are. As we are concerned with environmental and geographical data, resource-limited adaptation to those domains will be explored in the context of SmartOpenData.

4.4 Large volumes of real-time data

Another area of generic research for the project will be the handling of large volumes of real time data. This puts a strain on the infrastructure and so methods to reduce that stress will need to be researched, possibly using the W3C POWDER technology\textsuperscript{65} as a data compression tool. Tracking the provenance of any data is important of course but as yet there is no (standardised) linkage within the Semantic Web technology stack between Provenance\textsuperscript{66} and SPARQL Update.

\textsuperscript{65} http://www.w3.org/standards/techs/powder#w3c_all
\textsuperscript{66} http://www.w3.org/standards/techs/provenance#w3c_all
5 Legacy Architectures and existing Systems

There exist many different open data sources for protecting biodiversity and environmental research in Europe – in coastal zones, agricultural areas, forestry, etc., mainly focused on the Natura 2000 network\(^{67}\), and areas where environmental protection and activities like agriculture, forestry or tourism need to be balanced with the Habitats Directive\(^{68}\) and the European Charter for Sustainable Tourism in Protected Areas\(^{69}\). Better understanding and management of these data not only can provide economic value for these areas (value currently largely unknown), but will enable organizations to develop new services based on these data and open up new possibilities for public bodies and rural and protected areas to benefit from using data in novel ways, improving their knowledge and environment protection through new innovation ecosystems.

In this context, the SmartOpenData project has set its goals to:

1. Create a sustainable Linked Open Data\(^{70}\) infrastructure in order to promote environmental protection data sharing among public bodies in the European Union;
2. Enhance Linked Open Data with semantic support by integrating semantic technologies built upon connected Linked Open Data catalogues aiming at building sustainable, profitable and standardised environment protection and climate change surveillance services;
3. Define business models specially focused on SMEs and based on innovative services as new opportunities to align research results, previous work and projects, tackling active involvement of the whole value chain in Smart Regions at policy, industry and society levels;
4. Demonstrate the impact of the sharing and exploiting data and information from many varied resources, in rural and European protected areas by providing public access to the data and developing demonstrators that will show how services can provide high quality results in regional development working with semantically integrated resources.

In the context of the SmartOpenData project, in addition to using tools (such as those available on the Plan4Business Open Data Repository\(^{71}\), HABITATS Reference Laboratory\(^{72}\) and EnviroGRIDS portal\(^{73}\)) use of linked data for spatial data means identifying possibilities for the establishment of semantic connections between INSPIRE/GMES/GEOSS and Linked Open Data spatially related content in order to generate added value. The project requirements are within the environmental research domain. This will be achieved by making existing “INSPIRE based” relevant spatial data sets, services and appropriate metadata available through a new Linked Data structure. In addition, the proposed infrastructure will provide automatic search engines that will crawl additional available

\(^{67}\) http://www.natura.org/
\(^{68}\) http://ec.europa.eu/environment/nature/legislation/habitatsdirective/index_en.htm
\(^{69}\) http://www.european-charter.org/home/
\(^{70}\) http://linkeddata.org/
\(^{71}\) www.plan4business.eu
\(^{72}\) www.habitats.cz
\(^{73}\) Available at http://envirogrids.net/index.php?option=com_content&view=article&id=68&Itemid=81
geospatial resources (OGC and RDF structures) across the deep and surface web. The main motivation to utilise the potential of Linked Data is to enrich the INSPIRE spatial content to facilitate improved related services to be offered and to increase the number, performance and functionality of applications. In many cases querying data in INSPIRE (GEOSS) based data infrastructures (driven mainly by relational databases) is time consuming and often it is not sufficient and understandable for common Web users. In large databases such queries can take minutes or hours. In the cases of distributed databases such a query is almost impossible or very complicated. SmartOpenData aims to improve this situation dramatically.

The most advanced technical effort to reconcile the Linked Data and Geospatial Data worlds is embodied in OGC's GeoSPARQL standard. This merges the two technologies, with the GeoSPARQL engine translating queries back and forth between RDF and geospatial engines. The number of implementations of GeoSPARQL is growing but there remains some debate as to whether it is the best approach. The NeoGeo vocabulary is favoured by the French mapping agency IGN and handles geospatial data differently by linking to it from the RDF, rather than transporting large literals. The INSPIRE standards have been developed entirely in an XML-centric manner and the European Commission's JRC is now working on making better use of linked data. This is being done in a W3C Community Group focussing on locations and addresses. A related, but separate, Community Group is also considering better interplay between Web and geospatial technologies. What these activities all suggest is that there is work to be done to allow geospatial and linked data specialists to communicate easily, avoiding the so-called “religious wars”. ¡Error! No se encuentra el origen de la referencia. brings together specialists in both disciplines: RDF to describe a location or point of interest, GI to define where it is on the Earth's surface.

Current open data platform providers include CKAN, Junar, and Socrata, which are outlined in the following subsections.

5.1 CKAN data portal

CKAN, is a leading open-source data portal platform and powerful data management system that makes data accessible — by providing tools to streamline publishing, sharing, finding and using data. CKAN is aimed at data publishers (national and regional governments, companies and organizations) wanting to make their data open and available.

CKAN is open source and can be downloaded and used for free. Users can also get hosting and support from a range of suppliers. A full-time professional development team at the Open Knowledge Foundation maintains CKAN and can provide full support and hosting with SLAs.

CKAN Feature Overview:
- Complete catalogue system with easy to use web interface and a powerful API

http://www.opengeospatial.org/standards/geosparql
http://geosparql.org, https://twitter.com/marin_dim/status/271573164268609536,
http://www.strabon.di.uoa.gr
http://geovocab.org/doc/neogeo
http://www.w3.org/community/locadd
http://www.w3.org/community/geosemweb
http://ckan.org
http://ckan.org/#sthash.NT5L1Otg.dpuf

See more at http://ckan.org/#sthash.NT5L1Otg.dpuf
- Strong integration with third-party CMS’s like Drupal and WordPress
- Data visualization and analytics
- Workflow support lets departments or groups manage their own data publishing
- Fine-grained access control
- Integrated data storage and full data API
- Federated structure: easily set up new instances with common search

CKAN is a complete out-of-the-box software solution that makes data accessible — by providing tools to streamline publishing, sharing, finding and using data. CKAN is currently used by governments and user groups worldwide to power both official and community data portals. CKAN was developed by the non-profit Open Knowledge Foundation to run TheDatahub.org\(^81\), a public registry of open knowledge datasets. It now powers more than 40 data hubs around the world, including portals for local, national and international government, such as the UK’s data.gov.uk\(^82\) and the European Union’s publicdata.eu\(^83\).

**Technology**

CKAN is built with Python on the backend and Javascript on the frontend, and uses The Pylons web framework and SQLAlchemy as its ORM. Its database engine is PostgreSQL and its search is powered by SOLR. It has a modular architecture that allows extensions to be developed to provide additional features such as harvesting or data upload. CKAN uses its internal model to store metadata about the different records, and presents it on a web interface that allows users to browse and search this metadata. It also offers a powerful API that allows third-party applications and services to be built around it.

CKAN is a complete open source software solution for data publishers (national and regional governments, companies and organizations) that makes data accessible, by providing tools to streamline publishing, sharing, finding and using data.

![CKAN High Level Architecture](http://datahub.io/)

Figure 3 CKAN High Level Architecture

---

81 [http://datahub.io/](http://datahub.io/)
83 [http://www.publicdata.eu/](http://www.publicdata.eu/)
5.2 Junar platform

Junar delivers all the benefits of SaaS (Software-as-a-Service) to help organizations use Open Data to spur innovation. Junar makes it easy to deal with complex end-to-end Open Data projects and turns the difficult task of opening data into a secure and controlled process. This allows the data provider to focus on transforming data into readable, searchable, and usable insights that everyone can share.

Junar delivers an easy-to-use, cloud-based open data platform that enable businesses, governments, NGOs, and academia to free their data to drive new opportunities, collaboration, and transparency. Some of the world's leading companies trust Junar with their most valuable assets: their data and the end users who are using it.

With Junar users can open data fast, decide what is kept for internal use and what is made public, and optimize use of the data by adding meta information which drives search engine optimization.

Junar enables organizations to publish their data as follows:

- **Collect** - Identify and organize data quickly and easily using Junar’s intuitive discovery tool
- **Enhance** - Transform data into visually-appealing tables, charts, and maps
- **Publish** - Optimize the data publishing process with Junar’s workflow engine
- **Share** - Engage stakeholders in new ways through social media
- **Analyse** - Track how data is being viewed and used and measure its impact

Access to data leads to breakthrough innovation and Junar powers the emerging Data Economy by delivering a leading cloud-based Open Data platform. The platform enables businesses, governments, and other organizations to free their data to drive new opportunities, collaboration, and transparency.

Junar provides everything needed to open data with confidence. It's built for massive scale and supports local and global data-driven organizations. With Junar, organisations can easily choose what data to collect, how to present it, and when it should be made publicly available. Users can also determine which datasets are made available to the public and which datasets are available only for internal use. A next generation data management system for sharing information.

Junar can integrate data into the user’s Web site or host it. It provides a simple workflow that allows collaboration with colleagues and manages the entire Open Data project. The platform also allows users to drive social conversations with end users, understand what data they want and find valuable. And because the provider can add meta data to the information it publishes, it will likely see improved use of its data due to search engine optimization.

5.3 Socrata suite of products

Socrata is a Seattle-based cloud software company, focused exclusively on democratizing access to government data. They help public sector organizations improve transparency,
citizen service and fact-based decision-making by efficiently delivering data to citizens, employees and developers in a user-friendly experience on web, mobile and machine-to-machine interfaces.

Innovators like the World Bank, Medicare, Data.gov, EnergyStar, New York City, Chicago, San Francisco, Oregon and Maryland, have all chosen Socrata for its turnkey and low-cost cloud delivery of easy-to-use open data products.

The Socrata suite of products helps its customers unlock the full value of their data:

- Share public data with citizens and developers in the most engaging experience, with a global open data solution, Socrata Open Data Cloud.
- Empower employees to deliver amazing citizen self-service experiences with turnkey Quality of Life Apps from Socrata and the developer community.
- Save time and money by deploying cloud APIs that connect ecosystem partners and mobile app developers to data with Socrata API Foundry.
- Achieve higher levels of performance and shift to data-driven decision making with Socrata GovStat.
- Appify their data and bring the citizen experience to the modern era with Socrata DataSlate.
- Leverage the Socrata Data Consumerisation Platform and tools to modernize their entire capture-to-consumption data infrastructure, and develop custom data products using professional services from Socrata and its partners.

The Socrata Open Data Field Guide is a particularly useful comprehensive guide to ensuring that an open data program serves both the provider and their citizens.86

Socrata enables:

- Federation of User Data With Other Organizations - When it comes to open data, value grows exponentially the more data is shared between government organizations. Socrata Data Federation makes that a one-click process. Easily create a regional data hub with neighboring cities and counties in the user’s region. Constituents find relevant data based on topics or themes, regardless of where the data originates from. Federate seamlessly with other Socrata customers, or any portal that supports DCAT standards.

- Measuring the Success of an User’s Initiative in Real-time - Find out what’s working and what’s not based on user interactions, in addition to real-time consumption and distribution of your data and APIs.
  - Data Publishers track which data is popular with their constituents, how it is being consumed, and where it is shared on the web.
  - Real-time reporting allows data publishers to monitor poignant (“hot”) datasets, trending keyword searches, and influential traffic sources.
  - Track usage of your APIs and follow which apps are relying on your data.

- Create Beautiful Citizen Information Sites - Citizens, developers, reporters and researchers are looking for information they can use in their daily lives, not for a catalogue of datasets.

Configure the site any way you like, and create a look and feel that’s consistent with your organization’s visual identity

- Bring your site to life, and provide a rich context to help people relate to the information
- Highlight topical areas and align your data releases with your organizational goals and objectives

**Advanced Features for Technically-savvy Users include:**
- Researchers, analysts and other technically-savvy constituents need more advanced data analysis tools in addition to bulk downloads in multiple formats.
- Technically-savvy data consumers can perform advanced filtering, and sorting operations
- Additional analytical capabilities such as roll-up and group-by operations are available natively
- Advanced users can download data in a multitude of formats of their choosing including CSV, JSON, PDF, RDF, RSS, XLS, XLSX, or XML

**Making the data Social:**
- Civic engagement, or citizen participation, is a stated goal for most open data initiatives in government. Socrata gives constituents a rich social experience around data that promotes participation and expands reach.
- Constituents contribute to creating a rich experience around data by providing ratings and comments. The social feedback loop in turn drives more adoption and consumption of government data
- Constituents help propagate interesting data on social networks creating a social distribution channel for government data
- Co-creation, or Crowdsourcing Insight
- Interesting government data often tells stories that engaged constituents care about. Citizen journalists, bloggers and professional media are not content to just consume data. They want to contextualize it and use it to share their stories. The Socrata Data Player makes embedding data on blogs and web sites a snap.
- Anyone can easily embed a dataset or a user-created visualization on the web
- Helps propagate data on blogs and media sites, beyond the government’s datasite, increasing its reach and impact
- The Socrata Data Player provides a fully interactive experience that maintains the integrity and freshness of the source data

**Anyone Can Create Maps and Charts**
- Maps and charts are useful and friendly ways to transform millions of rows of data into information most people can relate to. Best of all, Socrata makes it easy for both publishers and constituents to create them.
- Everyone can easily create an interactive map from location data, or GIS files like Esri shapefiles, KML/KMZ files, using either Google Maps, Bing Maps or ESRI
- Chart creation capabilities includes various chart types such as Area, Bar, Column, Donut, Line, Pie, Time Line, Tree Map and Heat Map
- User-created visualizations become part of the site, allowing everyone to view and build on them, contributing to a process that makes data more useful with every contribution
5.4 SemaGrow project

SemaGrow (Data Intensive Techniques to Boost the Real-Time Performance of Global Agricultural Data Infrastructures) is a current FP7 project\(^{88}\) that aims to develop the scalable, efficient, and robust data services needed to take full advantage of the data-intensive and inter-disciplinary Science of 2020 and to re-shape the way that data analysis techniques are applied to the heterogeneous data cloud.

As the trend to open up data and provide them freely on the Internet intensifies, the opportunities to create added value by combining and cross-indexing heterogeneous data at a large scale increase. To seize them, SemaGrow aims to provide an infrastructure that is not only efficient, real-time responsive and scalable but is also flexible and robust enough to welcome data in any schema and form and to transparently relegate and translate queries from a unifying end-point to the multitude of data services that make up the open data cloud.

In order to achieve this ambitious vision and solve a difficult data management problem, SemaGrow is addressing the following key challenges:

- Develop novel algorithms and methods for querying distributed triple stores that can overcome the problems stemming from heterogeneity and from the fact that the distribution of data over nodes is not determined by the needs of better load balancing and more efficient resource discovery, but by data providers.
- Develop scalable and robust semantic indexing algorithms that can serve detailed and accurate data summaries and other data source annotations about extremely large datasets. Such annotations are crucial for distributed querying, as they support the decomposition of queries and the selection of the data sources which each query component will be directed to.
- Since it is, in the general case, not possible to align schemas and vocabularies so perfectly that there is no loss of information, investigate how to minimize losses and how to not accumulate them over successive schema translations.

\(^{87}\) At [http://github.com/socrata](http://github.com/socrata)

\(^{88}\) [www.semagrow.eu](http://www.semagrow.eu)
To address these challenges, SemaGrow carries out fundamental databases research and develops methods and infrastructure that will be rigorously tested on three large-scale current use cases as well as on their projected data growth beyond project’s end: we are laying the foundations for the scalable, efficient, and robust data services needed to take full advantage of the data-intensive and inter-disciplinary Science of 2020.

At the core of the SemaGrow architecture is the concept of efficiently identifying the optimal way to distribute a query among the nodes of a federation of SPARQL endpoints over heterogeneous data sources. The SemaGrow architecture proposes that this is achieved by means of collecting and indexing meta-information about the data stored in each data source; in this manner the data sources do not need to be cloned and re-hashed, and the way data is distributed among them does not need to be centrally controlled.

The SemaGrow Large Scale Distributed Architecture Stack, is as follows:

![Figure 4 SemaGrow Architecture](http://www.semagrow.eu/sites/default/files/SemaGrow_D2.3.1-Large%20Scale%20Distributed%20Architecture.pdf)

SmartOpenData will provide generic open data component interfaces similar to those defined in the SemaGrow project, as discussed in section 4.
5.5 GeoKnow project

GeoKnow is a recently established EU FP7 research project, motivated by previous work in the LinkedGeoData project (LGD), which makes OpenStreetMap (OSM) data available as an RDF knowledge base. As a result, OSM data were introduced in the LOD cloud and interlinked with GeoNames, DBpedia, and multiple other data sources. LGD intended to simplify information creation and aggregation related to spatial features. During this exercise, several research challenges were found such as scalability with spatial data, query performance, spatial data modelling, flexible transformation of special data, as well data operations such routing data. It was realised that geospatial data, specially scientific data, available on the web can open new opportunities to improve management and decision making applications.

The GeoKnow vision is to make geospatial data accessible on the web of data and turn the Web into a place where geospatial data can be published, queried, reasoned, and interlinked, according to the Linked Data principles. This will move geospatial data beyond syntactic interoperability to actual semantic interoperability, and to services that can geospatially reason on the Web. Linked data will not only be extended with spatial data to be able to improve information retrieval based on geospatial data, or to answer questions that were not possible with isolated geospatial data, but also represents a step towards the discoverability of data that share geospatial features (i.e. supported by querying and reasoning), and a boosting for the geospatial data integration through geospatial data merging and fusing tools.

GeoKnow aims to repurpose SDI standards, enabling the existing vast body of geospatial knowledge to be introduced in the Data Web. Further, it will apply the RDF model and the GeoSPARQL standard as the basis for representing and querying geospatial data. In particular, GeoKnow contributions will be in the following areas:

- Efficient geospatial RDF querying.
  - Existing RDF stores lack performance and geospatial analysis capabilities compared to geospatially-enabled relational DBMS. GeoKnow will focus on introducing query optimisation techniques for accelerating geospatial querying at least an order of magnitude.

- Fusion and aggregation of geospatial RDF data.
  - Given a number of different RDF geospatial data for a given region containing similar knowledge (e.g. OSM, PSI and closed data) and will devise automatic

---

92 www.geonames.org
96 http://ec.europa.eu/information-society/policy/psi/indexen.htm
fusion and aggregation techniques in order to consolidate them and provide a data set of increased value and quantitative quality metrics of this new data resource

- Visualisation and authoring.
  - GeoKnow will develop reusable mapping components, enabling the integration of geospatial RDF data as an addition data resource in web map publishing. Further, it will support expert and community-based authoring of RDF geospatial data within interactive maps, fully embracing crowdsourcing.

- Public-private geospatial data.
  - To support value added services on top of open geospatial data, GeoKnow will develop enterprise RDF data synchronisation workflows that can integrate open geospatial RDF with closed, proprietary data. It will focus on the supply chain and e-commerce use cases.

- GeoKnow Generator.
  - This will consist of a full suite of tools supporting the complete lifecycle of geospatial linked open data. The GeoKnow Generator will enable publishers to triplify geospatial data, interlink them with other geospatial and non-geospatial Linked Data sources, fuse and aggregate linked geospatial data to provide new data of increased quality, visualise and author linked geospatial data in the Web.

GeoKnow aims to contribute to the following areas concerned with geospatial data:

- Creation and maintenance of qualitative geospatial information from existing unstructured data such as OpenStreetMap, Geonames and Wikipedia, anticipating geospatial search and acquisition and aggregation of information resources.

- Reuse and exploitation of unforeseen discoveries found in geospatial data. GeoKnow will provide methods to acquire, analyse and categorise data that is rapidly evolving, immense, incomplete and potentially conflicting with: Tools and methodologies for mapping and exposing existing structured geospatial information on the web of data, considering comprehensive and qualitative ontologies and efficient spatial indexing.

- Automatic fusing and aggregation of geospatial data by developing algorithms and services based on machine learning, pattern recognition and heuristics.

- Tools for exploring, searching, authoring and curating the Spatial Data Web by using Web 2.0 and machine learning techniques based on scalable spatial knowledge stores.

All these contributions are integrated in the open source GeoKnow Generator framework developed by the consortium.
The GeoKnow Generator will provide a comprehensive toolset of easy-to-use applications covering a range of possible usage scenarios (e.g. mobility/traffic, energy/water, culture, etc).

The GeoKnow project contemplates two application scenarios:
- A travel e-commerce data management tool
- A spatial-semantic collaboration and data integration tool along value-chains in supplier and customer networks.

5.6 HABITATS Reference Laboratory

HABITATS (Social Validation of INSPIRE Annex III Data Structures in EU Habitats) was a CIP PSP ICT project that focussed on the adoption of INSPIRE standards through a participatory process to design and validate Data, Metadata and Services Specifications with real citizens and business.

The HABITATS pilots and implementation were based on
1. The HABITATS Social Media
2. The HABITATS Reference Laboratory

The HABITATS Reference Laboratory (RL) portal, which was continuously developed to evolve with the pilots’ needs, was a core enabler in the HABITAT project. It enabled sharing of data from different pilot localities, provided support for pilots with pan-European (and Global) data discovery services, and validated the INSPIRE compliance of the various data.

97 www.inspiredhabitats.eu
98 At http://www.inspiredhabitats.eu/Habitats/inspiredhabitats.ning.com/index.html
99 At www.habitats.cz
The set of services as implemented on the HABITATS Reference Laboratory (RL) geoportal platform include both interoperability services and enabling services, such as the:

a) visualisation of information layers,

b) overlay of information from different sources,

c) spatial and temporal analysis etc..

The RL acted as a client of the seven HABITATS pilots and provided a very rich set of cross-pilot, inter-regional and enabling services and tools that were validated by users on the basis of concrete implementations in phase 2 of the project.

The RL networking architecture is as follows:

---

### Figure 6 HABITATS Reference Laboratory Architecture

The HABITATS Reference Laboratory extends user-centric, co-design approaches into the arena of standards design and adoption processes, considering standards initiatives such as INSPIRE, OGC, UNSDI to be significant social, economic and institutional innovations. The elements of approach are maintained, applying the model at all levels from the global scale of the to the local and regional policies that framed many HABITATS validation pilots. Community building activities follow a Web 2.0 approach to capture the knowledge in active user communities with a strong

---

100 HABITATS D4.3.2 Networking services and service toolkit, March 2013, [www.habitats.cz/gallery](http://www.habitats.cz/gallery)
interest in contributing to the standards development process. By inviting a broad multi-sectoral and inter-disciplinary range of concerned stakeholders to participate in the HABITATS network, a viral motivation spiral was set off. A peer-to-peer approach to opening up information sources and providing access to content ensure a rapid extension of the critical mass of environmental data was established by project partners.

The RL allowed deployment of a state of the art technological solution, which was tested and adopted by HABITATS partners and user partners. It allowed testing of existing technology and generating of further research tasks driven by users.

The RL also collected information coming from other projects, which was an important input for HABITATS analysis and public discussion. The social assessment methodology was an important part of the Reference Laboratory

### 5.7 Plan4Business Open Data Repository

Plan4Business (A service platform for aggregation, processing and analysis of urban and regional planning data)\(^{101}\) is an FP7 project that has developed an aggregation platform serving multiple providers and thus offering users a full catalogue of planning data such as transport infrastructure, regional plans, urban plans and zoning plans. The platform offers clients the data itself in integrated, harmonised and thus ready-to-use form, but also provides rich analysis and visualisation services. Such services are offered via different interfaces, such as an API and an interactive web front-end (WebGIS).

The Plan4business platform is summarized in the following table\(^{102}\):

<table>
<thead>
<tr>
<th>Aspect</th>
<th>plan4business summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data types in the pool</td>
<td>Urban and regional planning data, basic geodata such as transport networks,</td>
</tr>
<tr>
<td></td>
<td>geostatistical data, land use</td>
</tr>
<tr>
<td>Similar pools</td>
<td>Only national or regional available, with limited online services, such as simple</td>
</tr>
<tr>
<td></td>
<td>searching</td>
</tr>
<tr>
<td>Pool size (end of the project)</td>
<td>25% of valid spatial plans in Europe, accessible without legislative restrictions</td>
</tr>
<tr>
<td>Curation cost change and</td>
<td>Improvement of Integration and Harmonisation Engines in project and pool of</td>
</tr>
<tr>
<td>monitoring</td>
<td>integrated data lowers effort</td>
</tr>
<tr>
<td>Main technical innovations</td>
<td>Collaborative integration and automated data harmonisation tools, analytical</td>
</tr>
<tr>
<td></td>
<td>processing and conversion engines</td>
</tr>
<tr>
<td>Services offered on the pool</td>
<td>Integration/Conversion, Data Management (Versioning, History), Analysis,</td>
</tr>
<tr>
<td></td>
<td>Processing, Simulation and Visualisation</td>
</tr>
<tr>
<td>API or Services</td>
<td>APIs will be offered as well as applications.</td>
</tr>
<tr>
<td>User sophistication</td>
<td>Professional technical and non-technical users</td>
</tr>
<tr>
<td>Owner of the data pool</td>
<td>Original data providers; licensing concept is in preparation for this purpose.</td>
</tr>
</tbody>
</table>

**Table 3 Plan4Business Characteristics**

The Plan4business Open Data Repository uses the Micka catalogue developed by HSRS, also a partner in SmartOpenData. The catalogue enables storage not only of metadata about existing datasets, but also about analyses, map compositions and integration services that can be performed. The Integration Engine accessing and harmonising data in the Storage

---

101 [www.plan4business.eu](http://www.plan4business.eu)
102 From [http://www.plan4business.eu/about/platform.html](http://www.plan4business.eu/about/platform.html)
Engine is supported by the HUMBOLDT Alignment Editor (HALE)\textsuperscript{103}. The Analysis Engine processes the requests given by users through the plan4business portal. The query is processed and the Analysis Engine accesses the data storage and retrieves query results that are then provided to users in standardised form. Its architecture is as follows\textsuperscript{104}:

![Plan4business Open Data Repository Architecture](image)

The relationships between plan4Business and the INSPIRE data models for vector and metadata are as follows:

![Plan4Business and INSPIRE data models](image)

\textsuperscript{103} www.esdi-community.eu/projects/hale

\textsuperscript{104} From http://www.google.ie/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=2&cad=rja&ved=0CDMQFjAB&url=http%3A%2F%2Fwww.plan4business.eu%2Fdep-bin%2Fdownload.pl%3Ff%3D153.pdf&ei=1XuMUrG8JuiS7Qao64CgCg&usg=AFQjCNF3-xcWC2QORC199KuZcKVI_Nckdflg
The Plan4business Open Data Repository builds on and extends the HABITATS Reference Laboratory tools, with two major additions that will benefit SmartOpenData

1. It has additional tools that partly provide LOD functionality.
2. It is collecting large scale heterogeneous data. It has a set of operational data from Europe, basic data, statistical data, data which was in HABITATS, OpenStreetMap transformed to INSPIRE models. Some country transport networks, for example Norway. Spatial plans from some countries, for example the Irish Open Land Use data, Czech cadaster maps, etc.

As the project has common partners with SmartOpenData, the repository could be used early in the project for fast prototyping with LOD tools of large-scale implementation in WP3 (Data modelling and LOD alignment) by transforming it to Open Data.

5.8 InGeoCLOUDS project

InGeoCLOUDS (INSPIREd GEOdata CLOUD Services) is a CIP PSP ICT project that aims to demonstrate the feasibility of employing a cloud-based infrastructure coupled with the necessary services to provide seamless access to geospatial public sector information, especially targeting the geological, geophysical and other geoscientific information. Geodata information possesses interesting characteristics such as the size of the available data, the existing metadata descriptions (mostly according to the European Directive INSPIRE) and the current availability of related services that can be moved to the cloud.

Project partners' data and services, available under more traditional infrastructures can be easily deployed to the cloud. One of the project challenges is the linking of the partners' data among themselves and with relevant external datasets.

InGeoCLOUDS provides the following services:

<table>
<thead>
<tr>
<th>User Group</th>
<th>Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consumers</td>
<td>Scientific / Data Services</td>
</tr>
<tr>
<td>Data providers</td>
<td>Consultancy for GSOM (GeoScientific Observation Model). Mapping of one's data model to GSOM.</td>
</tr>
<tr>
<td></td>
<td>INSPIRE consultancy</td>
</tr>
<tr>
<td></td>
<td>GSOM Service (Data Import; SPARQL Endpoints)</td>
</tr>
<tr>
<td></td>
<td>INSPIRE Compliant Data Publication / Services (Data Import; Catalogues; INSPIRE Services (WFS, WMS))</td>
</tr>
<tr>
<td>Application providers</td>
<td>Deployment of new applications</td>
</tr>
<tr>
<td></td>
<td>Simple GeoProcessing Facilities</td>
</tr>
<tr>
<td></td>
<td>Web WebGIS client</td>
</tr>
<tr>
<td></td>
<td>Monitoring and Billing Services</td>
</tr>
<tr>
<td>IT providers</td>
<td>InGeoCloudS package for private deployments</td>
</tr>
<tr>
<td></td>
<td>Elastic/QoS/Secure/Performance Infrastructure (PaaS)</td>
</tr>
</tbody>
</table>

Table 4 InGeoCLOUDS services

[105] [www.ingeoclouds.eu]
InGeoCLOUDS is based on Cloud computing because:

1. From an IT resource management point of view, hardware and network resources have become a critical bottleneck and major cost item. Of the main characteristics of the Cloud, perhaps the most essential from the user viewpoint are unlimited resources in terms of storage and computing power – it scales transparently and in a semi-automated manner, while offering up-to-date underlying technology – and its pay-as-you-go/pay-per-use delivery model with potential reduction of traditional IT infrastructure costs.

2. From a scientific point of view, rising data quantity and quality has not been accompanied by an equivalent increase in visibility, accessibility and sharing or better formal descriptions and standards. With its services for smooth data publication in line with Open Geospatial Consortium (OGC)/INSPIRE recommendations and integrating formalised conceptual models, InGeoCloudS provides a solution.

The InGeoCLOUDS architecture\textsuperscript{106} is as follows:

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure9.png}
\caption{InGeoCLOUDS architecture}
\end{figure}

The main roles of the architecture components are as follows:

\textsuperscript{106}From \url{http://www.ingeoclouds.eu/?q=wiki/architecture-logical-view} where much more information is provided.
1. Cloud computing platform - Every cloud service provider (e.g., Amazon) exposes a specific set of services to exploit the facilities of the cloud platform. For Pilot1 they intensively use West European data center (in Ireland) and developers/integrators also benefit from the Amazon AWS platform consoles that offer comprehensive view on the platform.

2. InGeoCLOUDS middleware - The InGeoCLOUDS (IGC) middleware exploits the underlying cloud computing platform to provide reliable and scalable storage and computing facilities to the other components of the architecture. There is only one component, named Elastic Compute, which directly interacts with the cloud platform, and providing to the IGC platform transparent access to the cloud services. The Elastic File Server and the Elastic Database Server provide scalable services through standard mechanisms, such as ODBC/JDBC and NFS. These two components are at the core of the architecture as they are used by every other component, and specifically by the data providers' applications to store and manage their data. The use of such standard technical solutions makes it possible to have a smooth adaptation of the existing services from the data providers’ infrastructures.

3. InGeoCLOUDS Management - Software components of the architecture are coordinated and monitored by the IGC Management component. This component is aware of the whole IGC architecture and coordinates the deployment of the various components and any other actions that need some sort of cooperation among software modules (e.g., creation of users accounts for the database server and the file server). All the management operations are handled by this module, for instance to create a new account, and they are forwarded as needed to the other components of the platform, e.g. create a new database for the given account.

4. InGeoCLOUDS Administration - The IGC Administration component provides administration tools for the whole system such as project’s management (corresponding to scientific domains), user management, portal configuration. It also includes billing and monitoring tools enabling the supervision of the whole platform and taking business-level decisions.

5. Data Publication - Data Publication is an important service of the platform that allows publishing of geo-dataset as a service in compliance with INSPIRE directive requirements, in particular with the use of the OGC standards services: Web Map Service (WMS) and Web Feature Service (WFS). The data Publication component is in charge of the creation of OGC service (“Create, Edit, Delete a service”) and to manage the access to these services for the others components of InGeoCLOUDS (portal) but also more open and external access to these services). The data publication is based on a web-dedicated mapping HTTP server and it provides all operations for the publication of geospatial data with the use of HTTP protocol.

6. Data Management - Integration of data providers’ services and applications into the InGeoCLOUDS architecture is also achieved through interactions with Data Import and Data Integration & Linking components. The two have very different goals. The Data Import component allows data providers importing data from their in-premises infrastructure. These data are transferred using FTP/FTPS or RESTful web services. The Data Integration & Linking module supports one of the most challenging objectives of this project: to allow the seamless integration of the disparate data
sources of the data providers in the form of Linked Data. This module provides the functionality of directly and indirectly creating Linked Data (through R2RML mapping specifications) from the providers’ respective data sets as well as querying and exporting the providers’ Linked Data in various formats.

7. Portal and Tools - Finally, the architecture provides some tools to facilitate the visualization and navigation of the information managed by the IGC platform. These include the Web GIS Client component, which helps the data provider to create interfaces and services responding to the objectives of different use cases. The Web GIS client is in charge of facilitating the integration of their applications at the GUI level and interaction between the data publication component and the web client. All uses cases require to propose a mapping interface available in the client browser, with very generic functions: zooms, pan, list of layers, etc. The Web GIS Client component includes all these functionalities (in javascript) and ensures it operates correctly with the OGC services available in the data publication component. The current implementation of the Web GIS Client is exploited by the portal of the InGeoCLOUDS platform. Data provider services are thus integrated in the cloud infrastructure and they make use of the facilities provided by the platform in terms of data storage, conceptual models, and publication services through Web maps and standardized services.

5.9 LOD2 project.

The Commission’s Open Data Support European initiative is using the DCAT Application Profile common metadata vocabulary to describe datasets for data portals in Europe\(^\text{107}\). While the metadata harvesting and publishing platform that they are using for collecting metadata of datasets from government data portals, transforming it into RDF, harmonising it according to the DCAT-AP, and publishing it as Linked Open Government Data (LOGD) is based on the Linked Open Data Management Suite developed in the LOD2 project\(^\text{108}\).

LOD2 (Creating Knowledge out of Interlinked Data) is a large scale FP7 Integrating Project\(^\text{109}\) that is developing:

- enterprise-ready tools and methodologies for exposing and managing very large amounts of structured information on the Data Web,
- a testbed and bootstrap network of high-quality multi-domain, multi-lingual ontologies from sources such as Wikipedia and OpenStreetMap.
- algorithms based on machine learning for automatically interlinking and fusing data from the Web.
- standards and methods for reliably tracking provenance, ensuring privacy and data security as well as for assessing the quality of information.
- adaptive tools for searching, browsing, and authoring of Linked Data.

\(^{107}\) https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-application-profile-data-portals-europe-final#download-links
\(^{108}\) https://github.com/nvdk/lodms-core/tree/virtuoso
\(^{109}\) http://lod2.eu/Welcome.html
They are integrating and syndicating linked data with large-scale, existing applications and showcasing the benefits in the three application scenarios of media and publishing, corporate data intranets and eGovernment.

Version 3.0 of the LOD2 Stack (as shown\(^\text{110}\)), comprises a number of tools for managing the life-cycle of Linked Data, comprising the following stages:

- Extraction of RDF from text, XML and SQL
- Querying and Exploration using SPARQL
- Authoring of Linked Data using a Semantic Wiki
- Semi-automatic link discovery between Linked Data sources
- Knowledge-base Enrichment and Repair

The LOD2 Stack suite of tools and components include\(^\text{111}\):

- **D2R Server** is a tool for publishing the content of relational databases on the Semantic Web, a global information space consisting of linked data.
- **Spatial Semantic Browser** allows geographic visualization of RDF data, a new iteration, called Facete also shows a faceted view of the data.
- **CubeViz** is a faceted browser for statistical data utilizing the RDF Data Cube vocabulary which is the state-of-the-art in representing statistical data in RDF. This vocabulary is compatible with SDMX and increasingly being adopted. Based on the vocabulary and the encoded Data Cube, CubeViz is generating a faceted browsing widget that can be used to filter interactively observations to be visualized in charts. Based on the selected structure, CubeViz offer beneficiary chart types and options which can be selected by users.
- **R2R Framework** enables Linked Data applications which discover data on the Web, that is represented using unknown terms, to search the Web for mappings and apply the discovered mappings to translate Web data to the application’s target vocabulary. The R2R Framework is aimed to be used by Linked Data publishers, vocabulary maintainers and Linked Data application developers.

\(^{110}\) From [http://stack.lod2.eu/blog/](http://stack.lod2.eu/blog/)

\(^{111}\) The LOD2 Stack Components and their APIs are described at [http://wiki.lod2.eu/display/LOD2DOC/LOD2+Stack+Components](http://wiki.lod2.eu/display/LOD2DOC/LOD2+Stack+Components)
- **OntoWiki** is a tool providing support for agile, distributed knowledge engineering scenarios. It facilitates the visual presentation of a knowledge base as an information map, with different views on instance data. It enables intuitive authoring of semantic content, with an inline editing mode for editing RDF content, similar to WYSIWIG for text documents. ... ([http://lod2.eu/Project/OntoWiki](http://lod2.eu/Project/OntoWiki))

- **ORE (Ontology Repair and Enrichment)** tool allows for knowledge engineers to improve an OWL ontology by fixing inconsistencies and making suggestions for adding further axioms to it.

- **DL-learner** software learns concepts in Description Logics (DLs) from user-provided examples. Equivalently, it can be used to learn classes in OWL ontologies from selected objects. It extends Inductive Logic Programming to Descriptions Logics and the Semantic Web. The goal of DL-Learner is to provide a DL/OWL based machine learning tool to solve supervised learnings tasks and support knowledge engineers in constructing knowledge and learning about the data they created.

- **Sparqlify** is a SPARQL-SQL rewriter that enables one to define RDF views on relational databases and query them with SPARQL. It is currently in alpha state and powers the Linked-Data Interface of the LinkedGeoData Server – i.e. it provides access to billions of virtual triples from the OpenStreetMap database.

- **LIMES** is a link discovery framework for the Web of Data. It implements time-efficient approaches for large-scale link discovery based on the characteristics of metric spaces. It is easily configurable via a web interface. It can also be downloaded as standalone tool for carrying out link discovery locally.

- **SigmaEE** is a tool to explore and leverage the Web of Data. At any time, information in Sigma is likely to come from multiple, unrelated Web sites - potentially any web site that embeds information in RDF, RDFa or Microformats (standards for the Web of Data).

- **SIREn** - Semantic Information Retrieval Engine - is a Lucene plugin to efficiently index and query RDF, as well as any textual document with an arbitrary amount of metadata fields.

- **Silk Link Discovery Framework** supports data publishers in accomplishing the second task. Using the declarative Silk - Link Specification Language (Silk-LSL), developers can specify which types of RDF links should be discovered between data sources as well as which conditions data items must fulfil in order to be interlinked. These link conditions may combine various similarity metrics and can take the graph around a data item into account, which is addressed using an RDF path language. Silk accesses the data sources that should be interlinked via the SPARQL protocol and can thus be used against local as well as remote SPARQL endpoints. ... ([http://lod2.eu/Project/Silk](http://lod2.eu/Project/Silk))

- **Sieve** allows Web data to be filtered according to different data quality assessment policies and provides for fusing Web data according to different conflict resolution methods.
• **LODrefine** is OpenRefine with integrated extensions that enable you to reconcile and extend data with DBpedia, extract named entities and upload your data on CrowdFlower crowdsourcing service.

• **Virtuoso** is an innovative industry standards compliant platform for native data, information, and knowledge management. It implements and supports a broad spectrum of query languages, data access interfaces, protocols, and data representation formats that includes: SQL, SPARQL, ODBC, JDBC, HTTP, WebDAV, XML, RDF, RDFa, and many more. ... ([http://lod2.eu/Project/Virtuoso](http://lod2.eu/Project/Virtuoso))

• **PoolParty** is a thesaurus management system and a SKOS editor for the Semantic Web including text mining and linked data capabilities. The system helps to build and maintain multilingual thesauri providing an easy-to-use interface. PoolParty server provides semantic services to integrate semantic search or recommender systems into systems like CMS, DMS, CRM or Wikis. ...

• **PoolParty Extractor** (PPX) offers an API providing text mining algorithms based on semantic knowledge models. With the PoolParty Extractor you can analyse documents in an automated fashion, extracting meaningful phrases, named entities, categories or other metadata. Different data or metadata schemas can be mapped to a SKOS thesaurus that is used as a unified semantic knowledge model.

• **CKAN** is a powerful data management system that makes data accessible – by providing tools to streamline publishing, sharing, finding and using data. CKAN is aimed at data publishers (national and regional governments, companies and organizations) wanting to make their data open and available.

• **DBpedia Spotlight** is a tool for automatically annotating mentions of DBpedia resources in text, providing a solution for linking unstructured information sources to the Linked Open Data cloud through DBpedia. DBpedia Spotlight performs named entity extraction, including entity detection and Name Resolution (a.k.a. disambiguation). It can also be used for building your solution for Named Entity Recognition, amongst other information extraction tasks.

5.10 GetLOD\(^{112}\) and the Emilia Romagna Region Geoportal\(^{113}\).

GetLOD is an open and reusable solution for publishing geographic data on the Web as Linked Open Data, according to standard RDF/XML. GetLOD is a solution jointly designed and developed by Planetek Italia and Sinergis during the evolving developments of the Geoportal of Emilia Romagna Region.

GetLOD is an open and reusable solution that allows an organization to publish geospatial data on the Web in the form of Linked Open Data, in accordance with the standard RDF/XML. GetLOD thus ensures the publication of geospatial data as open data and linkable, starting from traditional geo-web services. Making geospatial data available in a Spatial Data Infrastructure as Open Data, GetLOD guarantees the use of data as Linked Open Data, and


\(^{113}\) [http://www.planetek.it/progetti/geoportale_regione_emilia_romagna](http://www.planetek.it/progetti/geoportale_regione_emilia_romagna)
therefore their indexing on open data search engines and the integration with open data portals or with the Comprehensive Knowledge Archive Network (CKAN), the catalog of datasets and free projects. Integrating it in a standard mode to Geoportals, Open Data portals and Spatial Data Infrastructures based on the interoperability standards defined by the Open Geospatial Consortium (OGC).

GetLOD allows users to publish open geospatial data both in RDF format (Linked Open Data precisely) and in other non-linkable interchange formats (e.g. shapefiles and GML). The data concerning the publication as Linked Open Data are extracted in RDF format using the OGC WFS (Web Feature Service) services, made available by the SDI. The geospatial metadata extraction in RDF format is instead carried out using the standard services catalogue OGC-CSW (Catalogue Service for the Web), with metadata that can be associated with RDF (Linked Open Data) and the shapefile data.

GetLOD arises from the need to maximize the geospatial data value that, made accessible in a compliant way with the standard formats of reference for open data and although published independently by different actors, may be crossed freely by third parties. This innovative solution has been developed by Planetek Italia and Sinergis in the context of enhancement activities for the SDI of the Emilia-Romagna Region. It automatically makes available data and metadata managed by an SDI as Open Data and Linked Open Data, thereby facilitating the combination with other data and then their elaboration and analysis. Any software application, thanks to the open and standardized format, can thus allow users the access to these data and their reuse in creative ways exploiting, for their own goals, relationships (links) between heterogeneous information.

**Network Services**
- View
- Discovery
- Download
- Transformation

**Monitoring and Reporting**

**SDI Reference Model Components**
- Data
- Standards Implemented
- OGC

**Output Types**
- Tool
- Software
- License Types: Open Source

In a scenario with an increasing availability of web mapping services accompanied open licenses, GetLOD allows to make the geospatial information available in Linked Open Data format, thus facilitating the aggregation, processing and analysis, with other data. GetLOD arises from the need to maximize the value of geographic data that, accessible in standard formats for the open date, even if published independently by different individuals, can be crossed freely by third parties. Any computer application, through open and standardized format, can then access this data and reuse them in creative ways leveraging, for their own goals, the relationships (links) between heterogeneous information.
The difference between exposing on the Web "raw" geographic data (i.e., in Shape File format or KML) and the Linked Open Data Geographic is that the latter, due to the size and structure standards, can be directly used by computer applications without manual intervention. It is therefore “five stars” data, as classified by Tim Bernes Lee, because in addition to being available on the Web, structured and non-proprietary formats, they comply with open standards defined by W3C and contain links to other data: they essentially are automatically linkable.

**Geoportal Emilia-Romagna region**

The Geoportal of the Emilia-Romagna is the reference point and hub for the dissemination of spatial information and knowledge "geo-localized" region, in support of local governments and individual citizens. Disclosure, use, and distribution of data, information, and services a geographical conform to the most recent directives regional, national (CNIPA, Intesa GIS) and international (INSPIRE, OGC) in terms of interoperability.

The Geoportal represents the reference point and hub of knowledge "geo-localized" in support of regional and institutional activities of local governments that operate at the regional level of the individual citizen.

The Geoportal of the Emilia-Romagna is built by Planetek Italy, and powered entirely Open Source (Plone, OpenLayers, Ratman) and exposes web services that are OGC and ISO standards natively such as "Web Map Service" (WMS), "Web Feature Service" (WFS), "Web Coverage service" (WCS), "Web Processing service" (WPS) and manages services data catalog "Catalog Service - Web" (CS-W) available to external clients automatically.

The decision to base the system of publication of geospatial data of the Emilia-Romagna on open standards, the Open Geospatial Consortium (OGC) and International Standardization Organization (ISO), means that the regional administration can count on an interoperable platform for sharing their data both internally and with third parties in a manner consistent with the requirements of the INSPIRE Directive.

**Front-End**

GeoER allows the consultation to external web users (anonymous or authenticated on pillowcase, the authentication system federate bodies of Emilia-Romagna) and internal users (authenticated into the system of Identity & Access Management, IAM).

---

114 [http://geoportale.regione.emilia-romagna.it](http://geoportale.regione.emilia-romagna.it)
GeoER therefore provides direct access services to regional mapping that exists in the database, provides access to raster and vector maps available and offers a set of functions to support that improve the usability for the user, such as:

- navigation maps in 2D and 3D modes,
- integration with the system for the management of urban decay,
- integration with the sales system Map & Bookshop,
- integration with the CMS -GIS Moka to access applications or allow existing cartographic depth analysis
- download data
- consultation with metadata, search through alphanumeric and geographic positioning
- consultation of the data legend themed
- querying alphanumeric information related to the features
- identify the features of a mapping via the search function and spatial queries

GeoER has already been successfully subjected to a series of usability tests conducted by numerous users. This allows iterative improvement of the functionality provided by GeoER. Thanks to usability testing and taking into account user feedback, have been introduced a number of enhancements to the user interface which led to the current version.

GeoER now provides the following services:

- research services, which allow users to search for related spatial data sets and services based on the content of the corresponding metadata and to display the content of metadata;
- consultation services, which allow users to perform at least the following: display, navigate, view scale (zoom in and zoom out), change in the portion of land framed (pan), or overlay viewable spatial data sets and visualization the information contained in legends, and any relevant content of metadata;
- services for downloading data, which allow users to download copies of the data set or a part of them and access them directly;
- conversion services that allow users to transform spatial data sets in order to achieve interoperability;
- services that allow users to invoke spatial data services.

### 5.11 iMarine e-Infrastructure

The iMarine project uses a linked open data e-infrastructre to enable cross-disciplinary data collection, harmonization and analysis necessary for the management and conservation of marine living resources. The goal of the iMarine project is to facilitate policy makers in addressing the global challenges that impact on the marine ecosystem, the fisheries sector, local economies and communities.

The project provides a highly efficient e-Infrastructure that accelerates data discovery, exchange, harmonization, and analysis, for a variety of stakeholder communities. iMarine services provide life-cycle management and an open processing environment for scientific data and documents. The iMarine e-Infrastructure

115 [http://www.i-marine.eu](http://www.i-marine.eu)
- provides a cutting-edge e-infrastructure supporting marine and biodiversity specialist communities, unlocking knowledge and supporting science policy decision making.
- fosters a collaborative approach to open data access and interoperability.
- ensures seamless access to existing resources, increasing accessibility to rich data streams through the on-demand integration of data sets maintained by geographically dispersed organisations.

The iMarine e-infrastructure is accessed through the iMarine gateway, which is an access point to a number of Virtual Research Environments deployed and operated in the context of the iMarine project to support the Ecosystem Approach to fisheries management and conservation of marine living resources. It empowers practitioners and policy makers from multiple scientific fields such as fisheries, biodiversity and ocean observation by ensuring that otherwise dispersed and heterogeneous data is available to all stakeholder communities through a shared virtual environment that brings together multidisciplinary data sources, supports cross-cutting scientific analysis, and assists communication.

Each Virtual Research Environment is a "system" per-se aiming at providing their users with a web-based working environment offering the entire spectrum of applications and facilities (including services, data and computational facilities) needed to accomplish a given task. It offer specialized functionalities for the management, processing, and visualization of scientific data and textual content.

This gateway benefits from the D4Science infrastructure\(^\text{116}\), a Hybrid Data Infrastructure specifically conceived to interface with a rich array of Information Systems, Infrastructures and technologies, and offer them "as-a-Service". Among the Infrastructures and Information Systems bridged by the D4Science infrastructure, the following deserve a special mention: the Catalogue of Life (CoL\(^\text{117}\)), the FAO Geospatial Data Catalogue (FAO GeoNetwork\(^\text{118}\)), the Global Biodiversity Information Facility (GBIF\(^\text{119}\)), the Integrated Taxonomic Information System (ITIS\(^\text{120}\)), the Interim Register of Marine and Nonmarine Genera (IRMNG\(^\text{121}\)), the Marine Species Distribution Infrastructure (AquaMaps\(^\text{122}\)), the Ocean Biogeographic Information System (OBIS\(^\text{123}\)), the World Register of Marine Species (WoRMS\(^\text{124}\)).

The gateway and the D4Science infrastructure are developed and operated by using the gCube technology\(^\text{125}\). gCube is a framework dedicated to scientists. It enables the declarative and interactive creation of transient Virtual Research Environments that aggregate and deploy on-demand content resources and application services by exploiting computational and storage resources of a grid infrastructure.

The gCube software system is a Service Oriented Architecture, that combines a number of subsystems organized as follow:

- gCube run-time environment

\(^\text{116}\) www.d4science.org
\(^\text{117}\) www.catalogueoflife.org
\(^\text{118}\) http://geonetwork.fao.org
\(^\text{119}\) www.gbif.org
\(^\text{120}\) www.itis.gov
\(^\text{121}\) www.cmar.csiro.au/datacentre/irmng/
\(^\text{122}\) www.aquamaps.org
\(^\text{123}\) www.iobis.org
\(^\text{124}\) www.marinespecies.org
\(^\text{125}\) www.gcube-system.org
gCube Infrastructure Enabling Services

The gCube Infrastructure Enabling Services\(^{126}\) include the services necessary for the operation of the e-Infrastructure.

gCube is an infrastructure enabling software system, composed of a large number of software components. Its development started following a typical software development methodology: requirements-design-implementation-verification-maintenance. Pressed by the user needs this methodology was changed in favour of an Agile software development methodology. This implies that new software components, or revised version of existing software components, have been developed in relatively short subsequent iterations, each containing the tasks needed to release the new functionality - planning, requirement analysis, detailed design, implementation, integration, testing and documentation. As a consequence, a number of updated or enhanced versions of gCube components have been released during the project lifetime.

The gCube Application Services\(^{127}\) delivers the functions needed to the life of any VRE.

The gCube run-time environment\(^{128}\) includes a number of software components that are installed on each machine of the e-Infrastructure, that are called gCube Hosting Nodes (gHNs). Each gHN runs an environment that enables the hosting and operation of any other gCube service.

5.12 Nature Alliance and TESS Project

Naturalliance\(^{129}\) aims to help everyone whose work or recreation depends on nature. It will build up the knowledge users need, in their own language, for local decisions to manage and restore land, water and wildlife, whilst recording the good work they are doing for nature across Europe.

Overall in Europe, about a third of rural adults gather natural flowers, fruits, fungi and fuel; about a third as many fish, and there are roughly three anglers for every recreational hunter. In some countries, a high proportion of rural folk watches and feeds wildlife.

\(^{126}\) http://www.gcube-system.org/index.php?option=com_content&view=category&layout=blog&id=47&Itemid=55&id_articolo=121

\(^{127}\) www.gcube-system.org/index.php?option=com_content&view=category&layout=blog&id=47&Itemid=55&id_articolo=122

\(^{128}\) http://www.gcube-system.org/index.php?option=com_content&view=category&layout=blog&id=47&Itemid=55&id_articolo=120

\(^{129}\) http://www.naturalliance.eu/the_portal_for_nature bqxcpxf_yrxceqwp.aspx
Overlaps between these large interest groups means that many who manage land in gardens, farms and forests also hunt, fish and gather. All these interests tend to suffer as Europe’s natural riches are lost. Gardens have fewer pollinators. People travel further to watch wildlife or gather wild foods. Numbers of hunters and anglers decline. Income opportunities from all these diminish for local farmers and foresters.

Ten years ago, a Science article entitled “Can we defy nature’s end?” contained the words “Paradoxically we are not limited by lack of knowledge but failure to synthesise and distribute what we know.” Organisations in Naturalliance have worked even longer to build knowledge of governance and management that can restore the riches of nature through sustainable use. They now wish to build a system that makes it easy for governments, citizens and civil society organisations to use all the knowledge they can find.

Naturalliance seeks to reverse the current downward spiral by enlightening, enthusing and enabling local people to maintain and restore nature’s variety across Europe. They aim to help all who manage land and species to enhance the riches of nature, in ways that improve rural recreation and employment, through:

- “Learning by doing” to build interest in wild resources and increase their availability.
- encouraging local people and organisations to be ambassadors for health and wealth from nature.
- enhancing revenue for spending locally by reducing need to travel to enjoy wild flora and fauna.
- increasing gains from natural riches above the costs of the management needed to secure them.
- enabling local people to share the knowledge that makes the guidance system better and better.

**Free habitat and species mapper software from Anatrack**[^10].

The Mapper is Anatrack freeware, provided for the TESS (Transactional Environmental Support System) research project in the 7th Framework Programme of the European Commission, but built outside TESS using proprietary concepts.

The Mapper allows anyone to map habitats and species in their local environment. Starting with a map or aerial photograph of their garden, farm, school or local park for example, users can quickly and easily draw habitat and species information in intricate detail. It is possible to use this information online to study and manage our environment.

**TESS – Transactional Environmental Support System**[^11]

TESS was an FP7 environment project that designed a decision support system that can make it easy for policy makers to integrate local knowledge into their decision making, while also guiding and encouraging local activities that restore and maintain biodiversity and


ecosystem services. Their vision is to enlighten, encourage and empower local communities to support biodiversity restoration across Europe, through an internet system that unifies all available knowledge to guide decisions for the benefit of biodiversity and livelihoods.

TESS assists policy makers to integrate knowledge from the EU, national, regional and local level into the decision making process while also encouraging local people to maintain and restore biodiversity ecosystem services. To achieve this, a transactional environmental decision support system was designed, linking central policy planning to local livelihoods. To develop this system, TESS first researched the needs and capacities of central policy makers and local actors, identified paths and trajectories of cooperation, and modelled required transactions between the central and the local in relation to each one’s needs. A set of representative case studies from the whole EU (including the New Member States and pre-accession countries) was used to test the validity of the models and consolidate the project’s results into the design for a transactional environmental decision support system, named TESS. TESS also includes base-line information and predictive models for Strategic Environmental Assessment (SEA), Sustainability (Impact) Assessment (SIA) and Environmental Impact Assessment (EIA). TESS is supplemented by a set of brief and memorable policy guidelines to ensure its usefulness and enable its application in a European context. The process of developing TESS was facilitated by a large interdisciplinary consortium, in which participants include European associations with a strong network of support and influence not only in the Brussels milieu, but also at the grassroots.

The central result is the socio-economic and technical design for a Transactional Environmental Support System (TESS) to support exchange of environmental information between central and local levels, as well as meeting commitments in many areas of the Convention of Biological Diversity. The design is being tested by implementing socio-economic design in a knowledge portal to continue beyond TESS (www.naturalliance.eu) towards the intelligent GIS that could exchange decision support for fine-scale mapping of decision outcomes. This approach enables integration and delivery of formal environmental assessment systems with local knowledge and practices, through information and communication technologies, including GPS, remote/local sensing, and internet/mobile services. The aims of TESS are more ambitious than supporting central policy. The ultimate aim is to aid restoration and maintenance of biodiversity and natural resources by reversing the processes that caused so much degradation. Results also include recommendations and policy guidelines based on how biodiversity trends relate to different practices across Europe, addressed to those involved in the formulation, implementation, monitoring and evaluation of policies - at European, national, regional, and local levels.

Overall TESS has been a thoroughly Pan-European collaboration. Although much more research about information needs and technical development of decision-support mechanisms is required. They remain convinced that environmental information needs to be gathered and used freely by ordinary citizens, within a common EU-wide framework and subject to safeguards about what is sensitive at an individual level. They believe that such an approach will demonstrate that land-managers are not the problem but the solution to conserving and restoring Europe’s biodiversity.
5.13 DataLift Platform

Datalift brings raw structured data coming from various formats (relational databases, CSV, XML, ...) to semantic data interlinked on the Web of Data. Datalift is an experimental research project funded by the French national research agency. Its goal is to develop a platform to publish and interlink datasets on the Web of data. Datalift will both publish datasets coming from a network of partners and data providers and propose a set of tools for easing the datasets publication process.

A few steps to data heaven

The project aims to provide tools allowing to facilitate each step of the publication process:

- selecting ontologies for publishing data
- converting data to the appropriate format (RDF using the selected ontology)
- publishing the linked data
- interlinking data with other data sources

Open Data raises problems of heterogeneity due to the various adopted data formats and metadata schema descriptions. These problems may be overcome by using Semantic Web technologies in order to move from raw data to semantic data interlinked in the Web of Data.

However, lifting Open Data to Linked Open Data is far from being straightforward. Many challenges were faced in developing the DataLift platform, and the difficulties encountered in dealing with Open Data towards the publication of semantic interlinked data.

The ambition of DataLift is to act as a catalyst for the emergence of the Web of Data by providing a complete path from raw data to fully interlinked, identified, and qualified linked datasets. The Datalift platform supports the following stages in lifting the data:

- Selection of ontologies for publishing data;
- Conversion of data to the appropriate format (e.g., from CSV to RDF);
- Interlinking of data with other data sources;
- Publication of linked data;
- Access control and licence management.

The architecture of DataLift is modular. Several levels of abstraction allow decoupling between the different stages from raw data to semantic data. The dataset selection allows us to identify the data to be published and migrate them to a first RDF version. The ontologies selection step asks the user to input a set of vocabularies’ terms that will be used to describe the lifted data.

Once the terms are selected, they can be mapped to the raw RDF and then converted to properly formatted RDF. The data is then published on the DataLift SPARQL endpoint. Finally, the process aims at providing links from the newly published data to other datasets already published as Linked Data on the Web.

---

132 www.datalift.org
133 http://linkeddata.org/
1. Dataset Selection
   The first step of the data lifting process is to identify and access the datasets to be processed. A dataset is either a file or the result of a query to retrieve data from a datastore. The kinds of files currently considered are CSV, RDF, XML, GML and Shape files. Queries are SQL queries sent to an RDBMS or SPARQL queries on a triple store.

2. Ontologies Selection
   The publisher of a dataset should be able to select the vocabularies that are the most suitable to describe the data, and the least possible terms should be created specifically for a dataset publication task. The Linked Open Vocabularies (LOV) developed in Datalift provides easy access methods to this ecosystem of vocabularies, and in particular by making explicit the ways they link to each other and providing metrics on how they are used in the linked data cloud. LOV targets both vocabulary users and vocabulary managers:
   f) vocabulary users are provided with a global view of available vocabularies, complete with metadata enabling them to select the best available vocabularies for describing their data, and assess the reliability of their publishers and publication process,
   g) vocabulary managers are provided with feedback on the usability of what they maintain and publish, and tools to show the dependencies and history of the vocabularies. LOV is integrated as module in the DataLift platform to assist the ontology selection.

3. Data Conversion
   Once URIs are created and a set of vocabulary terms able to represent the data is selected, it is time to convert the source dataset into a more precise RDF representation. Many tools exist to convert various structured data sources to RDF. The major source of structured data on the Web comes from spreadsheets, relational databases and XML files. Datalift proposes a two steps approach.
   1. First, a conversion from the source format to raw RDF is performed.
   2. Second, a conversion of the raw RDF into “well-formed” RDF using selected vocabularies is performed using SPARQL Construct queries.
   Most tools provide spreadsheet conversion to CSV, and CSV to RDF is straightforward, each line becoming a resource, and columns becoming RDF properties. The W3C RDB2RDF WG proposes the Direct Mapping to automatically generate RDF from the tables but without using any vocabulary, and R2RML to assign vocabulary terms to the database schema. In the case of XML, a generic XSLT transformation is performed to produce RDF from a wide range of XML documents. The DataLift platform provides a graphical interface to help mapping the data to selected vocabulary terms.

4. Data Protection
   This module is linked to Apache Shiro for obtaining the information, i.e., username and password, about the user who is accessing the platform. The module checks which are the data targeted by the user’s query and then verifies whether the user can access the requested data. This verification leads to three kinds of possible answers, depending on the access privileges of the user: some of the requested data is returned, all the requested data is returned, or no data is returned. This means that the user’s query is

---

135 http://lov.okfn.org/dataset/lov
136 http://www.w3.org/wiki/ConverterToRdf
137 http://www.w3.org/2001/sw/rdb2rdf/
138 http://wimmics.inria.fr/projects/shi3Id
filtered in such a way that she is allowed to access only the data she is granted access to. The access policies are expressed using RDF and SPARQL 1.1 Semantic Web languages thus providing a completely standard way of expressing and enforcing access control rules.

5. Data Interlinking

The interlinking step provides means to link datasets published through the Datalift platform with other datasets available on the Web of Data. Technically, the module helps to find equivalence links in the form of “owl:sameAs” relations. An analysis of the vocabulary terms used by the published data set and a potential data set to be interlinked is performed. When the vocabulary terms are different, the module checks if alignments between the terms used by the two data sets are available. They use the alignment server provided with the Alignment API\textsuperscript{139} for that purpose. They translate the correspondences found into SPARQL graph patterns and transformation functions are combined into a SILK\textsuperscript{140} script.

6. Data Publication

This module aims at publishing the data obtained from the previous steps to a triple store, either public or private. The providers can restrict which graphs can be accessible, they could decide whether to provide just a “Linked Data” or a “Linked Open Data”. Datalift comes by default with Sesame, but provides an API for connecting to Allegrograph and Virtuoso triple stores as well.

**Lessons Learnt from a Datalift Camp**

Datalift as a “ready-to-use” platform was tested in the release 0.67 over two days by providers of datasets, public authorities and enterprises willing to know how Linked Data can help them in their “day-to-day” business. It was a challenging task because for most of the 75 participants, it was the first time they dealt with terms such as “vocabulary”, “interlinking”, “Linked Data”, “OWL”, “SPARQL”, etc. However, most of them were committed to follow the Open Data movement in France lead by data.gouv.fr and other initiatives from regions or cities.

The following lessons were learnt:

- Going through the LD lifecycle with Datalift is not straightforward if we consider users that are not familiar to semantic slang and technologies. Providers had to face recurrent issues such as:

- Choice of the suitable vocabulary that best covers the original dataset to lift: Here it requires time and efforts to figure out which are the vocabularies in LOV where the terms can be reused without the need of creating new vocabularies.

- Automatic detection of datasets to link to, or how to go beyond the “default DBpedia” dataset for interlinking. Publishers may want to have a list of possible candidates of datasets to interlink to w.r.t. their own datasets.

- Complexity of CONSTRUCT queries that serve as an alternative to make RDF2RDF transformation in the actual version of the platform.

\textsuperscript{139} http://alignapi.gforge.inria.fr/

\textsuperscript{140} https://www.assembla.com/wiki/show/silk/
● **Time required for pre-processing tasks such as data cleaning or normalizing all the attributes of a column field before using the first module of converting data to RDF.**

Few datasets were finally published during the two days of the camp, although a big step was achieved in letting know to the different providers what is possible to achieve with Datalift for publishing their raw data as Linked Data. The camp led to publishing the catalogue of the city of Montpellier in RDF using DCAT\(^1\), and Open Food Facts data in RDF\(^2\) and the food vocabulary\(^3\). The following are some recommendations that could help improving tools (such as SmartOpenData) similar to Datalift:

1. Hide the complexity of SPARQL with natural language QA systems like QAKIS\(^4\).
2. Integrate a categorised list of candidate datasets worth considering for linkage.
3. Need of vocabularies integrating multilingualism to ease search using terms that are not English.
4. Need of tools for transforming Shape files to RDF according to any given geographic vocabulary and/or requirement.

### 5.14 Other related projects and initiatives

The following are further related projects and initiatives on which SmartOpenData will build and reuse their available open tools, components and shared experience.

- The Open Knowledge Foundation, [http://okfn.org/](http://okfn.org/)
- W3C Semantic Web, [http://www.w3.org/standards/semanticweb/](http://www.w3.org/standards/semanticweb/)
- W3C LOGD WG, [http://www.w3.org/2011/gld/wiki/Main_Page](http://www.w3.org/2011/gld/wiki/Main_Page)
- LOD Around The Clock FP7 project, [http://latc-project.eu/](http://latc-project.eu/)
- Data.gov.uk, [http://data.gov.uk/linked-data](http://data.gov.uk/linked-data)

---


\(^3\) [http://data.lirmm.fr/ontologies/food#](http://data.lirmm.fr/ontologies/food#)

\(^4\) [http://dbpedia-test.inria.fr/qakis/](http://dbpedia-test.inria.fr/qakis/)
6 Relevant legal and political fundamentals

The debate on Open Data is still evolving\textsuperscript{145}. The best open government applications seek to empower citizens, to help small businesses, or to create value in some other positive, constructive way. Open government data is only a way-point on the road to improving education, improving government, and building tools to solve other real world problems. While many arguments have been made categorically, the following discussion of arguments for and against open data highlights that these arguments often depend highly on the type of data and its potential uses.

Arguments made on behalf of Open Data include the following:

- "Data belong to the human race". Typical examples are genomes, data on organisms, medical science, environmental data following the Aarhus Convention\textsuperscript{146}
- Public money was used to fund the work and so it should be universally available.
- It was created by or at a government institution (this is common in US National Laboratories and government agencies)
- Facts cannot legally be copyrighted.
- Sponsors of research do not get full value unless the resulting data are freely available.
- Restrictions on data re-use create an anti-commons.
- Data are required for the smooth process of running communal human activities and are an important enabler of socio-economic development (health care, education, economic productivity, etc.)
- In scientific research, the rate of discovery is accelerated by better access to data.

It is generally held that factual data cannot be copyrighted. However, publishers frequently add copyright statements (often forbidding re-use) to scientific data accompanying publications. It may be unclear whether the factual data embedded in full text are part of the copyright.

While the human abstraction of facts from paper publications is normally accepted as legal there is often an implied restriction on the machine extraction by robots.

Unlike Open Access, where groups of publishers have stated their concerns, Open Data is normally challenged by individual institutions. Their arguments have been discussed less in public discourse and there are fewer quotes to rely on at this time.

Arguments against making all data available as Open Data include the following:

- Government funding may not be used to duplicate or challenge the activities of the private sector (e.g. PubChem\textsuperscript{147}).
- Governments have to be accountable for the efficient use of taxpayer's money: If public funds are used to aggregate the data and if the data will bring commercial

\textsuperscript{145} http://en.wikipedia.org/wiki/Open_data
\textsuperscript{146} http://en.wikipedia.org/wiki/Aarhus_Convention
\textsuperscript{147} http://en.wikipedia.org/wiki/PubChem
(private) benefits to only a small number of users, the users should reimburse governments for the cost of providing the data.

- The revenue earned by publishing data permits non-profit organisations to fund other activities (e.g. learned society publishing supports the society).
- The government gives specific legitimacy for certain organisations to recover costs (NIST\textsuperscript{148} in US, Ordnance Survey in UK\textsuperscript{149}).
- Privacy concerns may require that access to data is limited to specific users or to subsets of the data.
- Collecting, 'cleaning', managing and disseminating data are typically labour and/or cost-intensive processes, whoever provides these services should receive fair remuneration for providing those services.
- Sponsors do not get full value unless their data is used appropriately - sometimes this requires quality management, dissemination and branding efforts that can best be achieved by charging fees to users.
- Often, targeted end-users cannot use the data without additional processing (analysis, apps etc.) - if anyone has access to the data, none may have an incentive to invest in the processing required to make data useful (Typical examples include biological, medical, and environmental data).

6.1 Legal barriers to linked open data\textsuperscript{150}

One of the biggest debates in the use of linked open data concerns the protection of personal rights of real persons and legal entities. The LinkedUp FP7 support action\textsuperscript{151} investigated the legal constraints connected to LOD applications, and legal compliance is one of the evaluation criteria for their data challenge.

LOD legal issues mainly relate to

1. Data protection,
2. IPR and copyright,
3. Privacy.

\textsuperscript{148} http://en.wikipedia.org/wiki/National_Institute_of_Standards_and_Technology
\textsuperscript{149} http://en.wikipedia.org/wiki/Ordnance_Survey
\textsuperscript{150} Adapted from http://linkedup-project.eu/2013/05/23/legal-barriers-to-linked-open-data/
\textsuperscript{151} http://linkedup-project.eu

The LinkedUp project is a FP7 Support Action which pushes forward the exploitation and adoption of public, open data available on the Web, in particular by educational organisations and institutions. It is undertaking:

1. Open Web Data Success Stories: gathering innovative and robust scenarios of deployed tools integrating and analysing large scale, open Web data in the education sector.
2. Web Data Curation: collecting, annotating and profiling of Web Data of educational relevance to enable take-up by third parties.
3. Evaluation Framework for Open Web Data Applications: providing a complete framework for the evaluation of large-scale open Web data applications, taking into account educational aspects.
4. Technology Transfer in the Education Sector: demonstrating and promoting the benefit of open Web data technologies in education, and provide a reusable testbed in this domain.

The overall aim is to facilitate the development of innovative applications produced by the LinkedUp community and challenge participants and their deployment in real-world use case scenarios. A collection of suitable use cases is being collected by the LinkedUp consortium and associated organisations, including representatives of renowned industrial, academic and higher education institutions such as Elsevier, the BBC, or the Commonwealth of Learning (at http://linkedup-challenge.org/usecases.html)
While items 1 and 3 are relatively well established in European law, item 3 is a bit more fuzzy and connected to ethical understanding that shows a wide variety of interpretations.

Data protection laws typically demand that individuals must not be identifiable through distributed datasets. That is to say that not only direct personal information is protected from distribution (names, address, birthday, etc.), but also inferable information that may lead to the identification of an individual (e.g. personal attributes, context they live and work in).

All of the SmartOpenData content and sources will need to be checked that they comply with this demand by effectively anonymising user data.

IPR and copyright is a legal entitlement that comes automatically into force when something is created. In the world of big data, however, it is important to note that small content items with little or no originality are not falling under copyright protection laws. This means that the typical content of data fields will not enjoy legal barriers for being exposed as open data.

On the other hand, database structures are protected by copyright laws and this may lead to concerns regarding the use of datasets in the SmartOpenData context. One possibility to evade this problem is to apply an open license such as the creative commons sets of usage licenses to datasets used by the SmartOpenData system.

As this legal debate continues to evolve SmartOpenData will collaborate with LinkedUp (possibly as an Associated Partner\(^{152}\)) to monitor the latest developments and use them in forthcoming future activities. Awareness of application designers and programmers, however, is equally essential.

### 6.2 Creative Commons\(^{153}\)

Creative Commons (CC) is a non-profit organization headquartered in the United States, devoted to expanding the range of creative works available for others to build upon legally and to share. The organization has released several copyright-licenses known as Creative Commons licenses free of charge to the public\(^{154}\). These licenses allow creators to communicate which rights they reserve, and which rights they waive for the benefit of recipients or other creators. An easy-to-understand one-page explanation of rights, with associated visual symbols, explains the specifics of each Creative Commons license. Creative Commons licenses do not replace copyright, but are based upon it. They replace individual negotiations for specific rights between copyright owner (licensor) and licensee, which are necessary under an "all rights reserved" copyright management with a "some rights reserved" management employing standardized licenses for re-use cases where no commercial compensation is sought by the copyright owner. The result is an agile, low-overhead and low-cost copyright-management regime, profiting both copyright owners and licensees. For instance, Wikipedia uses one of these licenses.

In 2008, there were an estimated 130 million works licensed under the various Creative Commons licenses. As of October 2011, Flickr alone hosts over 200 million Creative Commons licensed photos. Creative Commons is governed by a board of directors and a

\(^{152}\) [http://linkedup-project.eu/about/associated-partners/](http://linkedup-project.eu/about/associated-partners/)


\(^{154}\) [http://en.wikipedia.org/wiki/Creative_Commons_licenses](http://en.wikipedia.org/wiki/Creative_Commons_licenses)
technical advisory board. Their licenses have been embraced by many as a way for creators to take control of how they choose to share their copyrighted works.

Creative Commons has been described as being at the forefront of the copyleft movement, which seeks to support the building of a richer public domain by providing an alternative to the automatic "all rights reserved" copyright, and has been dubbed "some rights reserved."

Creative Commons attempts to counter what Lawrence Lessig, founder of Creative Commons, considers to be a dominant and increasingly restrictive permission culture. Lessig describes this as "a culture in which creators get to create only with the permission of the powerful, or of creators from the past". Lessig maintains that modern culture is dominated by traditional content distributors in order to maintain and strengthen their monopolies on cultural products such as popular music and popular cinema, and that Creative Commons can provide alternatives to these restrictions.

Creative Commons licenses consist of four major condition modules:

1. Attribution (BY), requiring attribution to the original author;
2. Share Alike (SA), allowing derivative works under the same or a similar license (later or jurisdiction version);
3. Non-Commercial (NC), requiring the work is not used for commercial purposes; and
4. No Derivative Works (ND), allowing only the original work, without derivatives.

These modules are combined to currently form the six major licenses of the Creative Commons.

i. Attribution (CC BY),
ii. Attribution Share Alike (CC BY-SA),
iii. Attribution No Derivatives (CC BY-ND),
iv. Attribution Non-Commercial (CC BY-NC),
v. Attribution Non-Commercial Share Alike (CC BY-NC-SA) and
vi. Attribution Non-Commercial No Derivatives (CC BY-NC-ND)

As of the current versions, all Creative Commons licenses allow the "core right" to redistribute a work for non-commercial purposes without modification. The NC and ND options will make a work non-free according to the Definition of Free Cultural Works.\(^{155}\)

An additional special license-like contract is the CCO option, or "No Rights Reserved." This license dedicates a work to the public domain (or an equivalent status in jurisdictions where a dedication to public domain is not possible). Compared with a "public domain" statement added to the work, a CCO statement is less ambiguous and achieves the desired effect on a global scale, rather than limited to some jurisdictions.

For software, Creative Commons endorses three free licenses created by other institutions: the BSD License\(^{156}\), the CC GNU LGPL license\(^{157}\), and the CC GNU GPL\(^{158}\).

All of these will be acceptable and used in the SmartOpenData system and content.

---

6.3 Open GI Data and the PSI Directive\textsuperscript{159}

In recent years the Open Data philosophy has gained a considerable momentum. In the public realm the free release of PSI datasets, besides enabling novel and promising forms of governmental accountability, paves the way to third-party developed products and services. Nevertheless, PSI re-use performed by private sector entrepreneurs is struggling to take-off due to the presence of numerous inherent roadblocks which are coupled with a certain vagueness surrounding the rationale underlying business endeavours. Taking stock of this evidence, SmartOpenData will investigate and promote the mechanisms allowing profit-oriented value creation based on public datasets. Delving into the intricacy of PSI re-use, the following eight business models have been identified as being employed by enterprises present in the world-wide PSI-centric ecosystem\textsuperscript{160}.

2. Freemium Product / Service.
3. Open Source Like.
4. Infrastructural Razor & Blades.
5. Demand-Oriented Platform.
7. Free as Branded Advertising.
8. White-Label Development.

These will be investigated further in the exploitation plans of WP7.


7 Quality Constraints

Once open spatial data is on the web then SmartOpenData must address:

- How to consolidate this movement and improve its quality in order to achieve linked open data useful to all?
- What is the toolbox needed by data producers to guide their publication efforts?
- How data quality can be addressed and fill the needs expressed by developers, citizens and public or private producers?

Data is of high quality "if it is fit for its intended uses in operations, decision making and planning.,” or more specifically: “High quality data is accurate, available, complete, conformant, consistent, credible, processable, relevant and timely.”

The Open Data Support initiative recommends 161:

- The quality of data is determined by its fitness for (re-)use by data consumers.
- Metadata is “data about data”, i.e. metadata is a type of data.
- The same quality considerations apply to data and metadata alike.
- Data quality has multiple dimensions and is about more than the correctness of data.
  It includes accuracy, availability, completeness, conformance, consistency, credibility, processability, relevance, timeliness.

The main dimensions to be taken into account for delivering good quality (meta)data are 162:

1. **Accuracy**: is the data correctly representing the real-world entity or event?
   - The accuracy of data is the extent to which it correctly represents the characteristics of the real-world object, situation or event.
   - For example:
     - Correct measurement of weather conditions (temperature, precipitation).
     - Correct indication of re-use conditions of the dataset.
   - SmartOpenData requirements:
     - Balance the accuracy of the data against the cost in the context of the application; it needs to be good enough for the intended use.
     - Make sure that there is organisational commitment and investment in procedures and tools to maintain accuracy.

2. **Availability**: Can the data be accessed now and over time?
   - The availability of data is the extent to which it can be accessed; this also includes the long-term persistence of data.
   - For example:
     - A Dataset that is identified by a http: URI that resolves persistently to the right resource (and does not give back 404 Not found).
     - A description of the dataset that is included in the search engine of a data portal.
   - SmartOpenData requirements:

---

161 See [www.slideshare.net/OpenDataSupport/open-data-quality](http://www.slideshare.net/OpenDataSupport/open-data-quality)
162 Adapted from [www.slideshare.net/OpenDataSupport/open-data-quality](http://www.slideshare.net/OpenDataSupport/open-data-quality)
• Follow best practices for the assignment and maintenance of URIs.
• Make sure that responsibility for the maintenance of data is clearly assigned in the organisation.

3. Completeness: Does the data include all data items representing the entity or event?
• The completeness of data is the extent to which it includes the data items or data points that are necessary to support the application for which it is intended.
• For example:
  • A Dataset that includes spending data for all ministries enables a complete overview of government spending.
  • A description of data that is generated in real time that includes the date and time of last modification.
• SmartOpenData requirements:
  • Design the capture and publication process to include all necessary data points.
  • Monitor the update mechanisms on a continuous basis.

4. Conformance: Is the data following accepted standards?
• The conformance of data is the extent to which it follows a set of explicit rules or standards for capture, publication and description
• For example:
  • A dataset that expresses coordinates in WGS84\(^{163}\) and statistics in SDMX\(^{164}\).
  • A description of a dataset according to the DCAT Application Profile.
• SmartOpenData requirements:
  • Apply the most used standards in the domain that is most relevant for the data or metadata.
  • Define local vocabularies if no standard is available, but publish vocabularies according to best practice (e.g. dereferenceable URIs).

5. Consistency: Is the data not containing contradictions?
• The consistency of data is the extent to which it does not contain contradictions that would make its use difficult or impossible.
• For example:
  • A dataset that combines data from different sources that has been processed to detect conflicting statements which have been resolved.
  • A description of a dataset that does not contain multiple licence statements or where the data of last modification is not before the creation date.
• SmartOpenData requirements:
  • Process all data before publication to detect conflicting statements and other errors (in particular if data is aggregated from different sources).

6. Credibility: Is the data based on trustworthy sources?
• The credibility of data is the extent to which it is based on trustworthy sources or delivered by trusted organisations.
• For example:
  • A dataset that contains data from processes that can be independently verified, e.g. election results or parliamentary proceedings.

\(^{163}\) World Geodetic System (WGS), see http://en.wikipedia.org/wiki/World_Geodetic_System
\(^{164}\) Statistical Data and Metadata eXchange, http://sdmx.org/
• A description of a dataset that is published by a government agency.
  
  • SmartOpenData requirements:
    • Base data on sources that can be trusted or on explicit Service Level Agreements where possible and appropriate.
    • Make appropriate attributions so that re-users can determine whether or not they can trust the data.

7. **Processability**: Is the data machine-readable?
   - The processability of data is the extent to which it can be understood and handled by automated processes.
   - For example:
     • A dataset that contains coded information based on publicly available controlled vocabularies and code lists.
     • A description of a dataset that expresses dates in W3C Date and Time Format (e.g. 2013-06-01) rather than as text (e.g. 1 June 2013).
   
   • SmartOpenData requirements:
     • Identify the source of terminology and codes used in the data in machine-readable manner.
     • Apply recommendations for syntax of data given in common standards and application profiles.

8. **Relevance**: Does the data include an appropriate amount of data?
   - The relevance of data is the extent to which it contains the necessary information to support the application.
   - For example:
     • A Dataset that contains temperature measurements rounded to degrees Celsius for climate calculations; a dataset with precision of a thousandth of a degree for chemical reactions.
     • A description of a dataset that only contains temporal coverage data if necessary for its processing.
   
   • SmartOpenData requirements:
     • Match coverage and granularity of data to its intended use within constraints of available time and money.
     • However, also consider potential future usages of the data.

9. **Timeliness**: Is the data representing the actual situation and is it published soon enough?
   - The timeliness of data is the extent to which it correctly reflects the current state of the entity or event and the extent to which the data (in its latest version) is made available without unnecessary delay.
   - For example:
     • A dataset that contains real-time traffic data that is refreshed every few minutes.
     • A description of a dataset containing annual crime statistics that is made available within days of publication of the dataset.
   
   • SmartOpenData requirements:
     • Adapt the update frequency of data to the nature of the data and its intended use.
     • Make sure that processes and tools are in place to support the updating.
The common elements that OpenDataSupport have identified in best practices that SmartOpenData will adopt are:

- Provide appropriate descriptions of data (i.e. metadata).
- Use standard vocabularies for metadata and data whenever such vocabularies exist.
- Specify the license under which the data may be re-used.
- Adhere to legal requirements concerning protection of personal and other sensitive data.
- Represent metadata and data according to the Linked Data principles using persistent URIs for identifying things.
- Provide information about the source of the data.

OPQUAST (Open Quality Standards) provides a checklist of good practices for open data quality based on 72 criteria\(^{165}\). This checklist is published under a CC-BY-SA license and is shown in Annex C. It is designed to be used by open data producers in order to improve the process of data publication on the Web. It addresses directly three targets:

1. the producer himself who can control better its project and the quality level he wants to reach,
2. the developers and their needs for standards, and
3. the final users and their need of usability.

Designed to guide open data publishers, this checklist can also help them to find the way to take the necessary steps to linked data.

Linked data is designed to help machines understand the semantic meaning held by data. But the goal of the linked data approach is to create data usable not only by the machines but also by human users. Then it's probably possible and useful to design a more complete checklist with special criteria for linked data approaches. Those quality criteria need now to be identified. They can help data producers to comprehend the added value of the linked data approach. Furthermore, both automatic and manual testing can also help them to improve their data production processes.\(^{166}\)

### 7.1 Metadata Quality in a Linked Data Context\(^{167}\)

There is little attention being paid to the practical tradeoffs involved in publishing high quality metadata at low cost. To date there has been a tendency to design metadata schemas that said absolutely everything that could be said about an object, often at the expense of obscuring what needed to be said about it while running up unacceptable costs. SmartOpenData is deeply interested in least-case-driven approaches to the design of metadata models, and defining what “good” metadata might be\(^{168}\). This ALA Metadata Quality definition created a framework for talking about (and evaluating) metadata quality, in a format that would be both domain-independent and useful to those who manage and evaluate metadata projects and how it relates to the quality standards that the Linked Open


\(^{166}\) “From open data to linked data: a quality checklist” Pascal Romain & Elie Sloïm, Conseil général de la Gironde/Temesis, [http://www.w3.org/2013/04/odw/papers#al38](http://www.w3.org/2013/04/odw/papers#al38)


Data (LOD) community is discovering for itself, and how it and other standards should affect library and publisher practices and policies, as follows:

<table>
<thead>
<tr>
<th>Quality Measure</th>
<th>Quality Criteria</th>
</tr>
</thead>
</table>
| **Completeness** | Does the element set completely describe the objects?  
Are all relevant elements used for each object?  
Does the data contain everything you expect?  
Does the data contain only what you expect? |
| **Provenance**[^169] | Who is responsible for creating, extracting, or transforming the metadata?  
How was the metadata created or extracted?  
What transformations have been done on the data since its creation?  
Has a dedicated provenance vocabulary[^170] been used?  
Are there authenticity measures (eg. digital signatures) in place? |
| **Accuracy** | Have accepted methods been used for creation or extraction?  
What has been done to ensure valid values and structure?  
Are default values appropriate, and have they been appropriately used?  
Are all properties and values valid/defined? |
| **Conformance to expectations** | Does metadata describe what it claims to?  
Does the data model describe what it claims to?  
Are controlled vocabularies aligned with audience characteristics & understanding of objects?  
Are compromises documented and in line with community expectations? |
| **Logical consistency and coherence** | Is data in elements consistent throughout?  
How does it compare with other data within the community?  
Is the data model technically correct and well structured?  
Is the data model aligned with other models in the same domain?  
Is the model consistent in the direction of relations? |
| **Timeliness** | Is metadata regularly updated as the resources change?  
Are controlled vocabularies updated when relevant? |
| **Accessibility** | Is an appropriate element set for audience and community being used?  
Is the data and its access methods well-documented, with exemplary queries and URIs?  
Do things have human-readable labels?  
Is it affordable to use and maintain?  
Does it permit further value-adds?  
Does it permit republication?  
Is attribution required if the data is redistributed?  
Are human- and machine-readable licenses available? |
| **Accessibility — technical** | Are reliable, performant endpoints available?  
Will the provider guarantee service (eg. via a service level agreement)?  
Is the data available in bulk?  
Are URIs stable? |

---

[^169]: See for instance DCMI Metadata Provenance Task Group, at [http://dublincore.org/groups/provenance/](http://dublincore.org/groups/provenance/)

[^170]: See W3C, “Provenance Vocabulary Mappings”.  
[www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings](http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings).
This list is not a checklist or a set of must-haves, nor does it contain all the questions that might be asked. Rather, it is intended as a list of representative questions that might be asked when a new SmartOpenData Linked Data source is under consideration. They are also questions that should inform policy discussion around the uses of Linked Data by consuming libraries and publishers.

That is work that can be formalized and taken further. One intriguing recent development is work toward a Data Quality Management Vocabulary. Its stated aims are to

- support the expression of quality requirements in the same language, at web scale;
- support the creation of consensual agreements about quality requirements
- increase transparency around quality requirements and measures
- enable checking for consistency among quality requirements, and
- generally reduce the effort needed for data quality management activities

The apparatus to be used is a formal representation of “quality-relevant” information. We imagine that the researchers in this area are looking forward to something like automated e-commerce in Linked Data, or at least a greater ability to do corpus-level quality assessment at a distance. Of course, “fitness-for-use” and other criteria that can really only be seen from the perspective of the user will remain important, and there will be interplay between standardized quality and performance measures (on the one hand) and audience-relevant features on the other.

The Linked Open Data (LOD) community has begun to consider quality issues. In general, the material leans toward the traditional data-quality concerns of the MIS community. LOD practitioners seem to have started out by putting far more emphasis on criteria that are essentially audience-dependent, and on operational concerns having to do with the reliability of publishing and consumption apparatus. As it has evolved, the discussion features an intellectual move away from those audience-dependent criteria, which are usually expressed as “fitness for use”, “relevance”, or something of the sort (such as “community expectations”). Instead, most realize that both audience and usage are likely to be (at best) partially unknown to the publisher, at least at system design time. In other words, the larger community has begun to grapple with the reality that future uses and the extent of dissemination are impossible to predict. There is a creative tension here that is not likely to go away. On the one hand, data developed for a particular community is likely to be much more useful to that community; thus the recognition of the role of “community expectations”. On the other, dissemination of the data may reach far past the boundaries of the community that develops and publishes it. The hope is that this tension can be resolved by integrating large data pools from diverse sources, or by taking other approaches that result in data models sufficiently large and diverse that “community expectations” can be implemented, essentially, by filtering.

For the LOD community, the path that began with “fitness-for-use” criteria led quickly to the idea of maintaining a “neutral perspective”. The idea that “Data quality is the degree to which data meets quality requirements no matter who is making the requirements”.

---


important to remember that the larger LOD community is concerned with quality in data publishing in general, and not solely with descriptive metadata, for which objectivity may no longer be of much value. For that reason, it would be natural to expect the larger community to place greater weight on objectivity in their quality criteria, with a strong preference for quantitative assessment wherever possible. Librarians and others concerned with data that involves human judgment are theoretically more likely to be concerned with issues of provenance, particularly as they concern who has created and handled the data. And indeed that is the case.

The following compares quality criteria with three views taken from the LOD community:

<table>
<thead>
<tr>
<th>Bruce &amp; Hillmann 173</th>
<th>Dodds 174, McDonald 175</th>
<th>Flemming 176</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completeness</td>
<td>Completeness</td>
<td>Amount of data</td>
</tr>
<tr>
<td></td>
<td>Boundedness</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Typing</td>
<td></td>
</tr>
<tr>
<td>Provenance</td>
<td>History</td>
<td>Verifiability</td>
</tr>
<tr>
<td></td>
<td>Attribution</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Authoritative</td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>Accuracy</td>
<td>Validity of documents</td>
</tr>
<tr>
<td></td>
<td>Typing</td>
<td></td>
</tr>
<tr>
<td>Conformance to expectations</td>
<td>Modelling correctness</td>
<td>Uniformity</td>
</tr>
<tr>
<td></td>
<td>Modelling granularity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Isomorphism</td>
<td></td>
</tr>
<tr>
<td>Logical consistency and coherence</td>
<td>Directionality</td>
<td>Consistency</td>
</tr>
<tr>
<td></td>
<td>Modelling correctness</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Internal consistency</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Referential correspondence</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Connectedness</td>
<td></td>
</tr>
<tr>
<td>Timeliness</td>
<td>Currency</td>
<td>Timeliness</td>
</tr>
<tr>
<td>Accessibility</td>
<td>Intelligibility</td>
<td>Comprehensibility</td>
</tr>
<tr>
<td></td>
<td>Licensing</td>
<td>Versatility</td>
</tr>
<tr>
<td></td>
<td>Sustainable</td>
<td>Licensing</td>
</tr>
<tr>
<td>Accessibility (technical)</td>
<td>Performance (technical)</td>
<td></td>
</tr>
</tbody>
</table>

Table 6 Views on LOD Quality Criteria


Some comments on this table:

- **Boundedness** has roughly the same relationship to **completeness** that precision does to recall in information-retrieval metrics. The data is complete when we have everything we want; its boundedness shows high quality when we have only what we want.

- Flemming’s **amount of data** criterion talks about numbers of triples and links, and about the interconnectedness and granularity of the data. These seem to be largely completeness criteria, though things to do with linkage would more likely fall under “Logical coherence”. Note, again, a certain preoccupation with things that are easy to count. In this case it is somewhat unsatisfying; it’s not clear what the number of triples in a triplestore says about quality, or how it might be related to completeness if indeed that is what is intended.

- Everyone lists criteria that fit well with our notions about **provenance**. In that connection, the most significant development has been a great deal of work on formalizing the ways in which provenance is expressed. This is still an active level of research, with a lot to be decided. In particular, attempts at true domain independence are not fully successful, and will probably never be so. It appears that those working on the problem at DCMI are monitoring the other efforts and incorporating the most worthwhile features.

- Dodds’ **typing** criterion — which basically says that dereferenceable URIs should be preferred to string literals — participates equally in completeness and accuracy categories. While URIs are preferred in current models, the presence of string literals may not always be a sign of low quality. Under some circumstances, for example, they might simply indicate an early stage of vocabulary evolution.

- Flemming’s **verifiability** and **validity** criteria need a little explanation, because the terms used are easily confused with formal usages and so are a little misleading. Verifiability bundles a set of concerns we think of as provenance. Validity of documents is about accuracy as it is found in things like class and property usage. Curiously, none of Flemming’s criteria have anything to do with whether the information being expressed by the data is correct in what it says about the real world; they are all designed to convey technical criteria. The concern is not with what the data says, but with how it says it.

- Dodds’ **modeling correctness** criterion seems to be about two things: whether or not the model is correctly constructed in formal terms, and whether or not it covers the subject domain in an expected way. Thus, it is assigned to both “Community expectations” and “Logical coherence” categories.

- **Isomorphism** has to do with the ability to join datasets together, when they describe the same things. In effect, it is a more formal statement of the idea that a given community will expect different models to treat similar things similarly. But there are also some very tricky (and often abused) concepts of equivalence involved; these are just beginning to receive some attention from Semantic Web researchers.

---


178 E.g. [http://www.w3.org/2009/12/rdf-ws/papers ws21](http://www.w3.org/2009/12/rdf-ws/papers ws21)
• **Licensing** has become more important to everyone. That is in part because Linked Data as published in the private sector may exhibit some of the proprietary characteristics which were access barriers, and also because even public-sector data publishers are worried about cost recovery and appropriate-use issues.

• A number of criteria listed under **Accessibility** have to do with the reliability of data publishing and consumption apparatus as used in production. Linked Data consumers want to know that the endpoints and triple stores they rely on for data are going to be up and running when they are needed. That brings a whole set of accessibility and technical performance issues into play. At least one website exists for the sole purpose of monitoring endpoint reliability\(^\text{179}\), an obvious concern of those who build services that rely on Linked Data sources. Also mirror sites may be required to eliminate problems with uptime and to guarantee low latency; performance and accessibility had become major concerns. For consumers, due diligence is important.

But the biggest difference in the “new quality” is that it puts a great deal of emphasis on technical quality in the construction of the data model, and much less on how well the data that populates the model describes real things in the real world. There are three reasons for that.

1. The first has to do with the nature of the discussion itself. All quality discussions, simply as discussions, seem to neglect notions of factual accuracy because factual accuracy seems self-evidently a Good Thing; there’s not much to talk about.
2. Second, the people discussing quality in the LOD world are modelers first, and so quality is seen as adhering primarily to the model itself.
3. Finally, the world of the Semantic Web rests on the assumption that “anyone can say anything about anything”, For some, the egalitarian interpretation of that statement reaches the level of religion, making it very difficult to measure quality by judging whether something is factual or not; from a purist’s perspective, it’s opinions all the way down. There is, then, a tendency to rely on formalisms and modeling technique to hold back the tide.

Libraries have always been concerned with quality criteria in their work as creators of descriptive metadata. So traditional quality criteria will evolve as libraries become publishers of Linked Data, as they will with SmartOpenData. That much seems fairly straightforward, and there are many processes and methods by which quality criteria can be embedded in the process of metadata creation and management.

More difficult, perhaps, is deciding how these criteria can be used to construct policies for Linked Data consumption using SmartOpenData. There are significant advantages and efficiencies that can be realized by using SmartOpenData to link to data and descriptions created by others, notably in connecting up information about the people and places that are mentioned in legislative information with outside information pools. That will require care and judgement, and quality criteria such as these will be the basis for those discussions. Not all of these criteria have matured — or ever will mature — to the point where hard-and-fast metrics exist. We are unlikely to ever see rigid checklists or contractual clauses with bullet-pointed performance targets, at least for many of the factors discussed here. Some of the new accessibility criteria might be the subject of service-level agreements or other

\(^{179}\) [http://labs.mondeca.com/sparqlEndpointsStatus/]
mechanisms used in electronic publishing or database-access contracts. But the real use of these criteria is in assessments that will be made long before contracts are negotiated and signed. In that setting, these criteria are simply the lenses that help users of SmartOpenData to know quality when they see it.

7.2 Service Level Agreements (SLA)

A service-level agreement (SLA) is a part of a service contract where a service is formally defined. In practice, the term SLA is sometimes used to refer to the contracted delivery time (of the service or performance). As an example, Internet service providers and telcos will commonly include service level agreements within the terms of their contracts with customers to define the level(s) of service being sold in plain language terms. In this case the SLA will typically have a technical definition in terms of mean time between failures (MTBF), mean time to repair or mean time to recovery (MTTR); various data rates; throughput; jitter; or similar measurable details.\(^\text{180}\)

A service-level agreement is a negotiated agreement between two or more parties, where one is the customer and the others are service providers. This can be a legally binding formal or an informal "contract" (for example, internal department relationships). Contracts between the service provider and other third parties are often (incorrectly) called SLAs – because the level of service has been set by the (principal) customer, there can be no "agreement" between third parties; these agreements are simply "contracts." Operational-level agreements or OLAs, however, may be used by internal groups to support SLAs.

The SLA records a common understanding about services, priorities, responsibilities, guarantees, and warranties. Each area of service scope should have the "level of service" defined. The SLA may specify the levels of availability, serviceability, performance, operation, or other attributes of the service, such as billing. The "level of service" can also be specified as "expected" and "minimum," which allows customers to be informed what to expect (the minimum), while providing a measurable (average) target value that shows the level of organization performance. In some contracts, penalties may be agreed upon in the case of non-compliance of the SLA (but see "internal" customers below). It is important to note that the "agreement" relates to the services the customer receives, and not how the service provider delivers that service.

SLAs commonly include segments to address: a definition of services, performance measurement, problem management, customer duties, warranties, disaster recovery, termination of agreement. In order to ensure that SLAs are consistently met, these agreements are often designed with specific lines of demarcation and the parties involved are required to meet regularly to create an open forum for communication. Contract enforcement (rewards and penalties) should be rigidly enforced, but most SLAs also leave room for annual revisitation so that it is possible to make changes based on new information.

Service level agreements are, by their nature, "output" based – the result of the service as received by the customer is the subject of the "agreement." The (expert) service provider can demonstrate their value by organizing themselves with ingenuity, capability, and knowledge to deliver the service required, perhaps in an innovative way. Organizations can

also specify the way the service is to be delivered, through a specification (a service level specification) and using subordinate "objectives" other than those related to the level of service. This type of agreement is known as an "input" SLA. This latter type of requirement is becoming obsolete as organizations become more demanding and shift the delivery methodology risk on to the service provider.

Service level agreements are also defined at different levels:

**Customer-based SLA**: An agreement with an individual customer group, covering all the services they use. For example, an SLA between a supplier (IT service provider) and the finance department of a large organization for the services such as finance system, payroll system, billing system, procurement/purchase system, etc.

**Service-based SLA**: An agreement for all customers using the services being delivered by the service provider. For example:

- A car service station offers a routine service to all the customers and offers certain maintenance as a part of offer with the universal charging.
- A mobile service provider offers a routine service to all the customers and offers certain maintenance as a part of offer with the universal charging.
- An email system for the entire organization. There are chances of difficulties arising in this type of SLA as level of the services being offered may vary for different customers (for example, head office staff may use high-speed LAN connections while local offices may have to use a lower speed leased line).

**Multilevel SLA**: The SLA is split into the different levels, each addressing different set of customers for the same services, in the same SLA.

**Corporate-level SLA**: Covering all the generic service level management (often abbreviated as SLM) issues appropriate to every customer throughout the organization. These issues are likely to be less volatile and so updates (SLA reviews) are less frequently required.

**Customer-level SLA**: covering all SLM issues relevant to the particular customer group, regardless of the services being used.

**Service-level SLA**: covering all SLM issue relevant to the specific services, in relation to this specific customer group.

### 7.2.1 Common metrics

Service level agreements can contain numerous service performance metrics with corresponding service level objectives

Uptime is a common metric, often used for data services such as shared hosting, virtual private servers and dedicated servers such as the SmartOpenData infrastructure. Common agreements include percentage of network uptime, power uptime, number of scheduled maintenance windows, etc.

### 7.2.2 Web and Cloud SLAs

As SmartOpenData will consist of web services in a cloud environment, then SLAs on these will be particularly relevant.
A web service level agreement (WSLA) is a standard for service level agreement compliance monitoring of web services. It allows authors to specify the performance metrics associated with a web service application, desired performance targets, and actions that should be performed when performance is not met. For instance, the WSLA Language Specification, version 1.0 which was published by IBM on January 28, 2001.\(^{181}\)

The underlying benefit of cloud computing is shared resources, which is supported by the underlying nature of a shared infrastructure environment. Thus, service level agreements span across the cloud and are offered by service providers as a service based agreement rather than a customer based agreement. Measuring, monitoring and reporting on cloud performance is based upon an end user experience or the end users ability to consume resources. The downside of cloud computing, relative to SLAs, is the difficulty in determining root cause for service interruptions due to the complex nature of the environment.

As applications are moved from dedicated hardware into the cloud, these applications need to achieve the same or even more demanding levels of service as classical installations. SLAs for cloud services focus on characteristics of the data center and more recently include characteristics of the network to support end-to-end SLAs.

### 7.2.3 SLA Management

Any SLA management strategy considers two well-differentiated phases:

1. the negotiation of the contract and
2. the monitoring of its fulfilment in real-time.

Thus, SLA Management encompasses the SLA contract definition: basic schema with the QoS (quality of service) parameters; SLA negotiation; SLA monitoring; and SLA enforcement—according to defined policies.

The main point is to build SmartOpenData as a new layer upon the cloud, or SOA middleware able to create a negotiation mechanism between providers and consumers of services. An example is the FP7 research project, SLA@SOI, which researched aspects of multi-level, multi-provider SLAs within service-oriented infrastructure and cloud computing; while another FP7 project, VISION Cloud has provided results with respect to content-oriented SLAs.

SLA@SOI\(^{182}\) was a consortium of leading Industrial, Academic and Research Institutes from around Europe committed to research, engineer and demonstrate technologies that can embed SLA-aware infrastructures into the service economy. Driven by four industry-led use cases in the areas of ERP Hosting, Enterprise IT, Service Aggregation and e-Government, the consortium:

1. documented an SLA-enabling reference architecture suitable for both new and existing service-oriented, cloud infrastructures
2. defined powerful reusable models such as SLA*, allowing arbitrary SLAs to be described

---


\(^{182}\) [http://sla-at-soi.eu/](http://sla-at-soi.eu/)
3. released a suite of open-source software, including the SLA@SOI framework of tools and components to help implement SLA-aware solutions, supported by extensive developer documentation
4. contributed to several open standards initiatives including W3C’s USDL\(^{183}\) and OGF’s OCCI\(^{184}\) and WS-Agreement\(^{185}\)
5. published extensively at numerous conferences, in numerous books, and online.
6. Documented all of these results in a series of formal deliverable documents, including their final report\(^{186}\) and an adoption guide to adopt the technologies involved\(^{187}\).

The VISION Cloud (VISION Cloud: Virtualised Storage Services Foundation for the Future Internet) FP7 project\(^{188}\) aimed to introduce a powerful ICT infrastructure for reliable and effective delivery of data-intensive storage services, facilitating the convergence of ICT, media and telecommunications. This infrastructure supports the setup and deployment of data and storage services on demand, at competitive costs, across disparate administrative domains, while providing QoS (Quality of Service) and security guarantees. The project is focusing on dealing with the strong proliferation of data-intensive services and the digital convergence of telco, media and ICT.

The project will deliver a high level architecture and reference implementation of a cloud-based infrastructure, built on open standards and new technologies, to provide a scalable, flexible and dependable framework for optimised delivery of data-intensive storage services. The project is demonstrating how this VISION Cloud infrastructure -- where an object data model, computational storage, content-centric access, comprehensive data interoperability, and QoS and security guarantees play a central role -- supports new emerging telco/media services. In doing so, it aims to achieve significant and quantifiable improvements in service delivery productivity, quality, availability, reliability and cost. All of which is directly relevant to SmartOpenData.

The VISION SLA Management system is shown in the following figure

\(^{183}\) http://www.internet-of-services.com/index.php?id=288\&L=0

\(^{184}\) http://occi-wg.org/

\(^{185}\) http://www.ogf.org/gf/group_info/areagroups.php?area_id=3


\(^{188}\) www.visioncloud.eu
Figure 12 VISION Cloud SLA Management system

SmartOpenData will build on the outputs of both SLA@SOI and VISION Cloud outputs and experience.

7.2.4 Five Star Open Data

SmartOpenData will provide publishing of Linked Data at a 5-star rating for all open data, as advocated by Tim Berners-Lee, inventor of the World Wide Web and founder of the Open Data Institute. The following figures show examples for each step of the stars and explains the costs and benefits that come along with it.

The following rules apply to each level:

* Put your data on the Web (any format)
* Make it available as structured data
* Use open, standard formats
* Use URLs to identify things
* Link your data to other people’s data

---

189 http://theodi.org/
190 http://www.w3.org/DesignIssues/LinkedData.html

Figure 13 5 Star Open Data
Table 7 Rules for the 5 Stars of Open Data

- The costs & benefits of ★ Web data?
  - As a consumer ...
    - ✔ You can look at it.
    - ✔ You can print it.
    - ✔ You can store it locally (on your hard drive or on an USB stick).
    - ✔ You can enter the data into any other system.
    - ✔ You can change the data as you wish.
    - ✔ You can share the data with anyone you like.
  - As a publisher ...
    - ✔ It’s simple to publish.
    - ✔ You do not have to explain repeatedly to others that they can use your data.

“It's great to have the data accessible on the Web under an open license (such as PDDL\textsuperscript{191}, ODC-by\textsuperscript{192} or CC0\textsuperscript{193}), however, the data is locked-up in a document. Other than writing a custom scraper, it’s hard to get the data out of the document.”

- The costs & benefits of ★★ Web data?
  - As a consumer, you can do all what you can do with ★ Web data and additionally:
    - ✔ You can directly process it with proprietary software to aggregate it, perform calculations, visualise it, etc.
    - ✔ You can export it into another (structured) format.
  - As a publisher ...
    - ✔ It’s still simple to publish.

“Splendid! The data is accessible on the Web in a structured way (that is, machine-readable), however, the data is still locked-up in a document. To get the data out of the document you depend on proprietary software.”

- The costs & benefits of ★★★ Web data?
  - As a consumer, you can do all what you can do with ★★ Web data and additionally:

\textsuperscript{191} http://www.opendatacommons.org/licenses/pddl/
\textsuperscript{192} http://www.opendatacommons.org/licenses/by/
\textsuperscript{193} http://creativecommons.org/publicdomain/zero/1.0/
You can manipulate the data in any way you like, without being confined by the capabilities of any particular software.

As a publisher...

- You might need converters or plug-ins to export the data from the proprietary format.
- It's still rather simple to publish.

"Excellent! The data is not only available via the Web but now everyone can use the data easily. On the other hand, it's still data on the Web and not data in the Web."

What are the costs & benefits of ★★★★ Web data?

As a consumer, you can do all what you can do with ★★★ Web data and additionally:

- You can link to it from any other place (on the Web or locally).
- You can bookmark it.
- You can reuse parts of the data.
- You may be able to reuse existing tools and libraries, even if they only understand parts of the pattern the publisher used.
- Understanding the structure of an RDF "Graph" of data can be more effort than tabular (Excel/CSV) or tree (XML/JSON) data.
- You can combine the data safely with other data. URIs are a global scheme so if two things have the same URI then it's intentional, and if so that's well on it's way to being 5 star data!

As a publisher...

- You have fine-granular control over the data items and can optimise their access (load balancing, caching, etc.)
- Other data publishers can now link into your data, promoting it to 5 star!
- You typically invest some time slicing and dicing your data.
- You'll need to assign URIs to data items and think about how to represent the data.
- You need to either find existing patterns to reuse or create your own.

"Wonderful! Now it's data in the Web. The (most important) data items have a URI and can be shared on the Web. A native way to represent the data is using RDF, however other formats such as Atom can be converted/mapped, if required."

What are the costs & benefits of ★★★★★ Web data?

As a consumer, you can do all what you can do with ★★★★ Web data and additionally:

- You can discover more (related) data while consuming the data.
- You can directly learn about the data schema.
- You now have to deal with broken data links, just like 404 errors in web pages.
- Presenting data from an arbitrary link as fact is as risky as letting people include content from any website in your pages. Caution, trust and common sense are all still necessary.

As a publisher...

- You make your data discoverable.
- You increase the value of your data.
- ✔ You own organisation will gain the same benefits from the links as the consumers.
- 💡 You'll need to invest resources to link your data to other data on the Web.
- 💡 You may need to repair broken or incorrect links.

7.2.5 SmartOpenData Service Level Agreement

While the SmartOpenData infrastructure will adopt a dynamic SLA Management approach based on the SLA@SOI and VISION framework of open tools and components to deliver a reliable operational service, all of its sources and services will adhere to good practice with lightweight Service Level Agreements, such as used, for instance, by bag42.nl\(^\text{194}\)

- Service level agreement - Best effort with no guarantees (although 99% uptime shouldn’t be a problem).
- Fair Use Policy - Access to BAG42 is unlimited, but we may have to apply a fair use policy during extreme usage (based on Dutch Database Law Article 4\(^\text{195}\))

Another example is the Swirrl PublishMyData\(^\text{196}\) Service Level Agreement:

- We target 100% availability of the system, but if it drops below 99.5% in any calendar month then partial refunds become applicable, on a sliding scale.
- You can monitor system performance via the administration tools and, although we make no guarantees against performance metrics, an average of 50-100ms per request can be expected due to our scalable architecture and caching techniques.
- Off-site backups are taken every time data is changed via the administration tools.

The SmartOpenData infrastructure will offer such an Uptime SLA, take regular backups and use load-balanced databases.

\(^{194}\) See [http://calendar42.com/en/bag42/](http://calendar42.com/en/bag42/)

\(^{195}\) [http://www.ivir.nl/wetten/nl/databankenwet.html](http://www.ivir.nl/wetten/nl/databankenwet.html)

\(^{196}\) [www.swirrl.com/publishmydata#features](http://www.swirrl.com/publishmydata#features)
8 Conclusions and Requirements

This first deliverable of WP2, describes the basic environmental requirements for the SmartOpenData infrastructure. In the previous sections it has described and is a repository on (a) the relevant components in the scenarios proposed (b) the characteristics of these components (c) how these components can be classified or generalised (d) what legacy architectures exist (e) what the relevant legal and political fundamentals are, and (f) what quality constraints exist and how they can be defined in the context of service level agreements (SLA). This section draws the conclusions from those discussions, by identifying, clustering and listing the requirements of the SmartOpenData infrastructure in the following subsections.

8.1 Generic Requirements

From a user perspective the SmartOpenData system will need to meet the following high level requirements197:

1. The system will be web based initially, but its architecture must be flexible enough that alternative front-ends may be developed (applets, cloud, etc).
2. The system must be able to contain socio-environmental data (spatial and non-spatial data, map images) and models in various formats, for various locations and with varying degrees of confidentiality.
3. All data and models used in the system will be tagged by origin, as public or private and with other appropriate meta-data and will be held secure from unauthorized access.
4. The system will also support standardized data-bases on private computers, on which the user can change data, mark it public or private, and use it with appropriate models in personal computers or on the system.
5. Public data will be acquired by the system, but may be changed by the system or originator (while keeping a transaction history and version control).
6. There must be an appropriate backup and restoration system.
7. Models may be acquired by the system for its use on a public or commercial basis, after appropriate validation.
8. The user and the system must be able to make requests for data and models of third-party databases, providing payment for access where necessary.
9. The user must be able to compare data and models from different sources and otherwise check for validity.
10. The system must be able to verify and check data and models for integrity; format conversions will be treated similarly.
11. The system must be able to accept donations, subscriptions and payments on account for models and data.
12. The system must be able to present itself and interact with the user in many languages.

13. The user must be able to create a user account so that the system remembers the user’s details (name, address, subscription and account details) at login; the system will maintain a list of accounts in its central database.

14. The user must be able to search for data by various search methods - location, type, keyword, date and so on – and then view the results.

15. The user and system must be able to apply appropriate data conversions, models and uncertainty analysis in data and produce scenarios.

16. It must be possible for the user to provide feedback on the data and models and there must be a complaints mechanism.

17. There must be scope for documentation, Help and tutorials.

18. The system must be able to interact with large external databases (e.g. CORINE\textsuperscript{198}).

19. The system will be scalable for increasing number of users.

\section{Service Requirements}

Based on the previous sections and the experience of the Plan4Business and HABITATS projects in particular, the SmartOpenData Infrastructure will need to provide the following services\textsuperscript{199}:

\begin{itemize}
  \item **Discovery**: provides access to the SmartOpenData and external metadata to users or to other system components. It implements search/discovery services, thus exposes catalogue services.
  
  \item **View**: view services perform the rendering of “generic data” (catalogue entry, map image,...) into an output format delivered to the user through the “horizontal service” and then through the application services.
  
  \item **Data services**: implements view and download services, thus exposes map/feature services.
    \begin{itemize}
      \item View: view services allow display, navigate, zoom in and out, pan or overlay viewable spatial data sets and display legend information and any relevant content of metadata.
      
      \item Download: download services allow extracting copies of spatial data sets, or parts of such sets, to be downloaded and, where practicable, accessed directly\textsuperscript{200}.
    \end{itemize}
  
  \item **Transformation**: is designed to carry out the mapping between the application schemas of SmartOpenData and the application schemas of the data provided by the partners.
  
  \item **Analysis**: implements transformation services\textsuperscript{201}. thus exposes map/feature transform services.
  
  \item **Monitoring**: as basic tools will be established monitoring based on logins of users. A full log analysis will be provided:
\end{itemize}

\textsuperscript{198} \url{http://www.eea.europa.eu/data-and-maps/find#c1=Data&c1=Graph&c1=Indicator&c1=Interactive+data&c1=Interactive+map&c1=Map&c6=&c9=all&c0=12&b_start=0}

\textsuperscript{199} Adapted from HABITATS D6.3 Business Models and Exploitation Plan, January 2013

\textsuperscript{200} From D3.5 – INSPIRE Network Services Architecture v2.0

\textsuperscript{201} From D3.5 – INSPIRE Network Services Architecture v2.0
External services

Discovery, view, data, transformation, analysis, authorization and authentication services can be implemented as internal, or can be used as external services coming from remote servers. Interfaces are the same as for internal services.

Applications

The SmartOpenData geo-portal is not tools for standard users. For most users there are important applications. The idea of the SmartOpenData architecture is to ensure that Applications are not developed as independent proprietary solutions, but are composed from existing services, using the SmartOpenData portal infrastructure.

WMC

Web Map Context (WMC) describes how to save a map view comprised of many different layers from different Web Map Servers.

RSS/GeoRSS

RSS (Rich Site Summary)\(^2\) is a family of web feed formats used to publish frequently updated works—such as blog entries, news headlines, audio, and video—in a standardized format. An RSS document (which is called a "feed", "web feed", or "channel") includes full or summarized text, plus metadata such as publishing dates and authorship. Web feeds benefit publishers by letting them syndicate content automatically. GeoRSS is an emerging standard for encoding location in a Web feed\(^3\)

KML/KMZ

Keyhole Markup Language (KML)\(^4\) is an XML schema for expressing geographic annotation and visualization within Internet-based, two-dimensional maps and three-dimensional Earth browsers. A KMZ file consists of a main KML file and zero or more supporting files that are packaged using a Zip utility into one unit, called an archive.\(^5\)

CMS

A content management system (CMS)\(^6\) is the collection of procedures used to manage work flow in a collaborative environment. These procedures can be manual or computer-based. The procedures are designed to do the following:

- Allow for large numbers of people to contribute to/share stored data
- Control access to data, based on user roles (defining which information users or user groups can view, edit, publish, etc.)
- Aid in easy storage and retrieval of data
- Reduce repetitive duplicate input
- Improve the ease of report writing
- Improve communication between users

Social Networks and Media

Social media are media for social interaction, using highly accessible and scalable publishing techniques. Social media use web-based technologies to turn communication into interactive dialog. A common thread running through all definitions of social media is a blending of technology and social interaction for the co-creation of content and

---


\(^3\) [http://en.wikipedia.org/wiki/RSS](http://en.wikipedia.org/wiki/RSS)


\(^6\) [https://developers.google.com/kml/documentation/kmzarchives](https://developers.google.com/kml/documentation/kmzarchives)

\(^6\) [http://en.wikipedia.org/wiki/Content_management_system](http://en.wikipedia.org/wiki/Content_management_system)
value add.\textsuperscript{207}

**Workflow management**

A workflow consists of a sequence of connected steps\textsuperscript{208}. It is a depiction of a sequence of operations, declared as the work of a person, a group of persons, an organization of staff, or one or more simple or complex mechanisms. Workflow may be seen as any abstraction of real work.

**GeoSPARQL tools**

A suite of GeoSPARQL GeoSpatial LOD tools translating queries back and forth between RDF and geospatial engines.

### 8.3 Specific Requirements of the SmartOpenData infrastructure

From the previous sections a list of specific SmartOpenData operational system requirements can now be identified to make spatial data easier to discover and use, according to the Linked Open Data Strategy for the purpose of environmental information. To achieve the following SmartOpenData primary goals:

1. Make environmental and geospatial data concerning rural and protected areas more readily available and re-usable, better linked with data without direct geospatial reference so different distributed data sources could be easily combined together.

2. Make existing “INSPIRE based” relevant spatial data sets, services and appropriate metadata within the environmental research domain available through a new Linked Data structure.

3. Harmonise geospatial metadata (ISO19115/19119 based) using the principles of the Semantic Web, provide spatial data fusion introducing principles of Open Linked Data, improve spatial data visualisation of Geospatial Open Linked Data and publish the resulting information according to user requirements and LOD principles. Allow geospatial and linked data specialists to communicate easily: RDF to describe a location or point of interest, GI to define where it is on the Earth’s surface.

4. Enhance Linked Open Data with semantic support by integrating semantic technologies built on connected Linked Open Data catalogues aiming at building sustainable, profitable and standardised environment protection and climate change surveillance services, and contribute to standards, particularly through the W3C Community Groups mechanism and collaboration with the OGC.

5. Enable incremental sharing of quality Open Data to allow for offer and demand to meet, truly allowing pilot providers to exploit their results and prompting the commercial sector to create value added products and services, thus creating a new global environmental protection market with final user involvement.

6. Achieve real user engagement and trust to empower the further adoption of Open Data sharing and semantic services in GI, by enabling the implementation, meeting all functional requirements and validating the semantic technologies developed to be fully interoperable at data, metadata, semantic and legal levels in the 5 pilots described in Annex B, driven by strategic partners in Spain, Italy, Ireland, Czech Republic and Slovakia. Thus to demonstrate the impact of sharing and exploiting data

\textsuperscript{207} http://en.wikipedia.org/wiki/Social_media

\textsuperscript{208} http://en.wikipedia.org/wiki/Workflow
and information from many varied resources, in rural and European protected areas by providing public access to the data and developing demonstrators to show how services can provide high quality results in regional development working with semantically integrated resources.

7. Satisfy the requirements of four classes of target users: (i) Public bodies, (ii) Researchers, (iii) Companies and (iv) Citizens, in using public data resources, existing sources for biodiversity and environment protection in European protected areas.

8. Adhere to the GEOSS Data Sharing Principles\(^{209}\) and their implementation, adopt the OPQUAST OpenData best practices (as in Annex C) whenever possible, and support the UNSDI (United Nation Spatial Data Infrastructure) vision\(^{210}\):

9. Enable the use of models, innovative environmental tools and information products, based on accepted standards. Reuse of existing European Spatial Data Infrastructures (SDI), based on INSPIRE, GMES and GEOSS (Free Pan European Data Sets such as CLC, Natura 2000, HABITATS, Plan4all, Plan4business, EnviroGRIDS, Briseide\(^{211}\), GEOSS registries, national INSPIRE portals, thematic portals like National Forestry portals\(^{212}\) together with local and regional data) and extend it using Linked Open Data

10. Provide access to much data that will be available through SPARQL queries. Build an infrastructure of objects and relationships with the added value of further links. By associating existing geospatial data with URIs used elsewhere, recording semantic relationships and linking across different data sets, the objects will have greater context and therefore usefulness.

11. Work with and use the recommendations of the European Commission’s Open Data Support initiative\(^{213}\), to improve the visibility and facilitate access to datasets published on local and national open data portals in order to increase their re-use within and across borders.

To achieve these goals, the SmartOpenData Linked Open Data infrastructure will need to support the SmartOpenData use case/domains from each pilot’s end-users’ point of view\(^{214}\), focused on

- Agroforestry management,
- Environmental research and Biodiversity,
- Water monitoring and
- Forest sustainability
- Environmental data reuse.

By providing

1. An user-friendly application interface for querying data (as a simple form to be able query data without standards experience) in each Pilot, as both
   (i) Spatial queries (processed in traditional spatial database) and
   (ii) semantic queries based on SPARQL.


\(^{210}\) [http://www.unngwg.org/content/united-nations-spatial-data-infrastructure-unsdi](http://www.unngwg.org/content/united-nations-spatial-data-infrastructure-unsdi)

\(^{211}\) [www.briseide.eu](http://www.briseide.eu)

\(^{212}\) See [www.efi.int/portal/](http://www.efi.int/portal/) and [http://forestportal.efi.int/](http://forestportal.efi.int/)


\(^{214}\) See Annex B.
2. Visualisation of the results of queries, as both:
   (i) a list of objects in some form
   (ii) cartographic in relation to the original data, including mechanisms, like Filter Encoding etc. to visualise results based on queries.

3. Easy to use tools for non-experts in cartography or Linked Data to exploit the data and create web based geospatial enabled applications.

4. Mechanisms for acquiring, adapting and using Open Data provided by existing sources directly involved in the project for biodiversity and environment protection in rural and European protected areas and its National Parks.

5. Discoverability, Federation and Interoperability based on strong referenceable catalogues of metadata from numerous sources, an ecosystem of distributed data catalogues that can be propagated for discovery and compatible data catalogues. Seamless interoperability of data catalogues.

6. The following distributed service system functionalities to be accessed transparently between external data sources and consumers in the scenarios:
   - Harmonisation of data sources, as an open data source layer that exposes the external data sources fully adapted to open data standards supported by the project.
   - Adaptation specifically tuned for each external data source that does not provide the information according to the required standard.
   - An open data source layer providing both semantic information of the data and metadata:
     a) distributed semantic indexing, providing a service for searching and locating data based on semantic information collected from all available Data Sources;
     b) distributed data access, providing data collected from external data sources, as an extra data source for easier and uniform data gathering from the users at the identified scenarios;
     c) administration and notification, providing facilities for managing users, workflows and data to data providers.

7. Full support of OGC's GeoSPARQL standard, CKAN and use of the DCAT Application Profile for its data portals.

8. Ability to handle large volumes of real time and stored data in a reasonable response time.

9. Automatic search engines that can crawl available geospatial resources (OGC & RDF structures) across the deep and surface web.

10. Multilingual access, by overcoming the monolingual nature of typical datasets, using the commonality of data structure and query language to make them available in multiple languages.

11. Integration of MT modules with GI (data and metadata services). To include language resource components that support MT to translate geographical data and metadata.

215 As specified by OpenDataSupport, see https://joinup.ec.europa.eu/asset/dcat_application_profile/asset_release/dcat-application-profile-data-portals-europe-final#download-links
using the GEMET\textsuperscript{216} and AgroVoc\textsuperscript{217} thesauri and reuse existing Open Source software packages, such as Moses\textsuperscript{218}

12. Algorithms that expose the wealth of accessible environmental data as linked data and require minimal manual intervention, to be repeatable and scalable.

13. Generic open data component interfaces similar to those defined in the SemaGrow, GeoKnow and LOD2 projects.

14. Similar functionality, build on and seamlessly interface and interwork with de facto industry-standard open data platforms such as CKAN, Junar and Socrata.

15. Reuse of and build on the open infrastructures and suites of tools of the Plan4Business Open Data Repository, HABITATS Reference Laboratory and projects such as SemaGrow, GeoKnow, EnviroGRIDS, LOD2 and GetLOD.

16. Semantic Linked Data structures to integrate heterogeneous data based on different domains using common geographic concepts (e.g. river, road...) and their relations (e.g. a road crosses a river). Seamlessly integrate the definition of points, lines and polygons with the relationships between features, the names of places and, in particular, the identifiers, and encode them as linked data at scale.

17. Appropriate descriptions of data (i.e. metadata). Will use standard vocabularies whenever such exist. Specify the license under which the data may be re-used. Adhere to legal requirements concerning protection of personal and other sensitive data. Represent metadata and data according to the Linked Data principles using persistent URIs for identifying things. Provide information on the source of the data.

18. A community Governance Model offering contributors and participants guidance in deploying innovative services and environment-friendly business provided in a safe and secure manner with a common legal framework.

19. Mechanisms to address data protection, IPR/copyright, privacy and ethical issues. All user data will be effectively anonymized. All datasets used by the SmartOpenData system will use Creative Commons licenses, with the CC\textsuperscript{0}\textsuperscript{219} option, or "No Rights Reserved", being preferred, particularly for the metadata, (as recommended by OpenDataSupport).

20. High quality data and metadata that is accurate, available, complete, conformant, consistent, credible, processable, relevant and timely. Using Linked Data sources with high quality metadata with regard to Completeness, Provenance, Accuracy, Conformance to expectations, Logical consistency and coherence, Timeliness and Accessibility.

21. 5 Star Linked Open Data and preserving data quality through best-practice guidelines, QA procedures, IPR issues, Certification and Reputation for Open Data modelling.

22. Production level/quality tools (recommended by OpenDataSupport) for:
   a) Data cleansing and curation.
   b) Managing the data lifecycle with mechanisms for handling updates and deletions in the data.

---
\textsuperscript{216} \url{www.eionet.europa.eu/gemet/}
\textsuperscript{217} \url{http://aims.fao.org/standards/agrovoc/}
\textsuperscript{218} \url{www.statmt.org/moses/}
\textsuperscript{219} “No Rights Reserved” Creative Common Licence, \url{http://creativecommons.org/about/cc0}
23. A dynamic SLA Management approach based on the SLA@SOI and VISION framework of open tools and components to deliver a reliable operational service. With all of its sources and services adhering to good practice with light weight Service Level Agreements

In addition the following secondary user requirements can be identified. These are considered to be more generic and “good to have” rather than being essential:

The SmartOpenData Linked Open Data infrastructure could ...

1. Create a sustainable Linked Open Data infrastructure in order to promote environmental protection data sharing among public bodies in the EU.

2. Make European Spatial Data easily re-usable not only by GIS experts but also by various organizations and individuals at a larger scale with Google maps and other widely used sources.

3. Use Linked Open Data for modelling, acquiring, harmonising and using data provided by sources from existing catalogues and open public data portals: GEOSS Data-CORE, GMES, INSPIRE, EC Open Data Portal, data from EC projects in FP7, eContent+ and CIP, Global Forest Observation Initiative and the ISA Programme

4. Make available geographical and environment data to further application and service domains through the integration of semantic results among public open data sources available by partners and other available public data including INSPIRE, GEOSS and GMES and external semantic services such as DBPedia, Freebase, GeoLinkedData and the recently released Open Data Portal by the Commission including valuable datasets by the EEA and Eurostat.

5. Enrich the INSPIRE spatial content with LOD techniques to enable improved related services to be offered and to increase the number, performance and functionality of applications.

6. Foster integration among sources and Linked Data with regards to knowledge representation, data modelling and interfaces for publishing the resulting information, as well as contributing to the adoption of GMES (to be operational from 2014), and the GEOSS 2012-2015 Work Plan.

7. Provide a common query language that gives access to related datasets available in the linked open data cloud.

8. Be an open reference implementation using the Apache-2 licence.

9. Avoid duplicating information. Reuse existing URIs where ever possible, especially the ones coming from reference data sources, such as company registers.

---

220 http://linkeddata.org/
221 http://www.earthobservations.org/art_011_003.shtml
222 http://dbpedia.org/About
223 www.freebase.com
224 http://datahub.io/dataset/geolinkeddata
225 http://open-data.europa.eu/
226 See www.earthobservations.org/geoss_imp.php
227 Defined at www.apache.org/licenses/LICENSE-2.0.html
10. Enable the definition of business models specially focused on SMEs and based on innovative services as new opportunities to align research results, previous work and projects, tackling active involvement of the whole value chain in Smart Regions at policy, industry and society levels. Allow alternative business model for publishing linked data to be explored.

11. Promote the adoption of new standards regarding
   (i) alignment and compatibility of semantic technologies,
   (ii) geospatial data interoperability,
   (iii) new RDF-based vocabularies for integrated location description and
   (iv) Open Data Quality, Certification and Reputation.

12. Define a standardised linkage between Provenance and SPARQL Update.

13. Consider a Data Quality Management Vocabulary, to support the expression of quality requirements in the same language, at web scale; support the creation of consensual agreements about quality requirements, to increase transparency around quality requirements and measures, to enable checking for consistency among quality requirements, and generally reduce the effort needed for data quality management activities.

14. Build new collaborations among a variety of projects, initiatives and research results focused on environmental protection through the establishment of semantic services, to turn individual environmental innovation initiatives into a collective set of meaningful seamless policies with exponential impact, that eventually will bring more open data shared by new engaged public bodies (i.e. ESA, FAO, UNEP, etc.).

In conclusion, we believe that based on these requirements, SmartOpenData will be a great system, we know how we’ve come to that conclusion and now we must design, implement and validate that system, sustain it into the future and exploit it across Europe.
Annex A: Linked Open Data

This is a short introduction to Linked Data\footnote{228 Taken from \url{http://3roundstones.com/linked-data-101/}}

What is the point of using Linked Data?

The point of Linked Data is for webmasters to publish data in open, standardized formats that facilitate reuse by others. Others may include customers, suppliers and partners. If you’re a government authority, publishing Linked Data allows your data to be more readily used by other government agencies, the research community, and the general public.

Organizations employing Linked Data are improving data quality, shortening development cycles, and significantly reduce maintenance costs. Enterprises are realizing a return on investment on Linked Data projects typically within 6-12 months.

With traditional 3 tier data architectures, 60% of the full cost of an application is in application and data maintenance. Linked Data based solutions cost a small fraction of traditional applications due to efficiencies in data re-use, data exchange standards and cloud computing becoming more common. Unlike proprietary approaches, there is no vendor lock-in nor vendor dominance.

Where can I see a demonstration of Linked Data?

Round Stones operates an online Linked Data demonstration site at \url{demo.3roundstones.net}. This demonstration provides information regarding nuclear power plants located in the United States. It contains data gleaned from DBpedia, Open Street Maps, SEC Info, the U.S. Environmental Protection Agency’s Facilities Registry System, Substance Registry System and Toxic Release Inventory and Abt Associates report on corporate ownership.

The purpose of this demonstration is to show the benefits of combining data from multiple sources and the ease and speed of creating Web applications using Callimachus.

Why should we publish Linked Data when we already publish our data in a variety of formats on the Web?

Providing data in open, standardized formats that facilitate reuse by other government agencies and/or departments, and third parties, e.g., journalists, academic, non-profit and corporate researchers and the general public. Generalized data sharing is made possible by the use of an international data exchange standard, the Resource Description Framework (RDF). Data in RDF allows for rapid combination of information from multiple data sources, including DBpedia (the RDF version of Wikipedia) and literally thousands of Linked Open Data sets available on the web.

Publishing as RDF allows people to rapidly visualize ad hoc queries on maps, in tables, bar charts and many other common business views. Publishing Linked RDF is what Tim Berners-Lee, the inventor of the Web, calls “5 star” Linked Data. Linked Data publishing and use by enterprises has resulted in cost reductions in development time, deployment cycle, and maintenance compared with traditional data sharing mechanisms.
How do I know that Linked Data is not a passing fad?

Commercial companies recognize the market opportunity and are investing millions of dollars of R&D budget to create and support production tools improved creation, discovery and visualization of data on the Web. This includes Google, Oracle, IBM, Microsoft and Facebook. The UK Parliament pioneered publishing Linked Data with the backing and support of Sir Tim Berners-Lee, the inventor of the World Wide Web. Additionally, governments such as the US, Sweden, Germany, France, Spain, New Zealand and Australia are adopting Linked Data as a data publication and consumption model for Open Government Initiatives. The BBC is using Linked Data to operate large sections of its Web site and also used it to report on the last Olympics.

While the term “Linked Data” is a relatively new term (circa 2007), it is based on International Standards and technologies that have formally and comprehensively presented, discussed and peer reviewed by literally hundreds of academic institutions, technology companies, and government agencies from around the world through the World Wide Web Consortium (W3C) for well in excess of a decade.

How long do projects typically take to implement?

A Linked Data Approach implies “cooperation without coordination.” This means that all members of an organization are not required to agree on schema, in advance or at any point in the development effort. Instead, a Linked Data approach recognizes that there is no one way to describe an organization, its products or services. Instead, a Linked Data approach embraces that individuals possess knowledge within their area of expertise and that they should be able to describe business process, rules and their data with both flexibility and standards.

Are there any gotchas?

a) There are several issues when adopting Linked Data that could become ‘gotchas’: Care must be taken to avoid biasing the value of high quality datasets by tightly coupling them to specific high profile applications. In short, do not do to your Linked Data what MDM did to your relational databases. Recognize that the core benefits of Linked Data involve the combination of data with data from other sources. Successful Linked Data projects produce generic, reusable data that may be combined with data from other sources to allow applications not yet conceived. Think reuse, not specific uses.

b) Openly publishing data, be it Linked Data or not, must be undertaken under appropriate licensing which is unambiguous, appropriate, unrestrictive, and realistic as possible. We offer more detail below under “Risks”.

c) Avoid “triplifying” data by automatic script. Triplifying data by script is not the same as creating well-structured Linked Data suitable for building applications. Proper data modeling is an essential first step. Efforts to automatically generate billions of RDF “triples” and publish them on the Web is not the same as producing high quality data sets of properly modeled data.

d) People and organizations experienced with data modeling in RDF are still relatively rare.

As one embarks on an effort to convert a dataset, what are the factors that determine conversion cost?

a) What makes a dataset complex or simple
Decisions on exposing a data set should be based on usefulness to others. Usefulness is a measure of its ability to be used by others, both intra-agency, interagency and by the public. Only data of general usefulness should generally be published as Linked Data; agency- or application-specific data is not always useful to the rest of the world.

The following are indicators of, but not hard and fast rules about, what factors impact the time required to expose data that is useful to others.

More complex relational data models (say 60 or more tables) are more time consuming and complex to model, and therefore the time required is slightly longer, (measured in weeks however, not months or years);

Data sets that require a prior knowledge of the agency organization, e.g., structure, regulations, workflow, internal vocabularies (e.g., for naming conventions) require meetings between data modelers and internal specialists;

Domain specific data sets (e.g., geography, chemistry, physics or complex regulation) may require specialized domain expertise that may be harder to find or schedule.

b) What does it cost to host RDF?

There is no one size fits all answer on pricing, however the factors are all familiar to IT managers and procurement departments. The cost of modeling Linked Data and hosting it is based on several components including:

- Time required to remodel data, typically measured in several weeks;
- Frequency and size of updates;
- Access, including query volume;
- Applications (if applicable) based on Linked Data sets.

Technology teams accustomed to managing hardware, networking, and service level agreements for traditional 3 tier applications will understand similar components to hosting a Linked Data service.

Hosting is quickly becoming commoditized in terms of pricing. The value proposition should focus more on the service level agreement, patches and upgrades, security and other features that are vital to any production data or application service.

Data consisting of millions of rows in a relational database is typically easy and inexpensive to host. There are economies of scale, hosting more data sets is not necessarily proportionally more expensive.

The value proposition in using Linked Data not on the lower cost to host the data (RDF triples), however, the ability to provide high availability production managed services using Linked Data.

There are software-as-a-service options which provide an easy scalable option in the early stages while enabling analysis of medium to long-term possibilities as the profile of the data and its use is established.

c) What is the cost of putting up a new version of the data?

This is dependent on the quantity of data, quality of modeling, frequency and size of updates, plus the ability of the chosen store to take live updates. The cost is often negligible and included in the cost of a hosting contract.
Once the data is properly modeled, scripts are run to automatically convert data to Linked Data (as RDF triples) on a routine basis (e.g., hourly, daily, weekly, etc)

**What about costs with a Linked Data approach?**

Traditional data warehouse projects require significant upfront coordination. The cost of vocabulary creation and/or schema alignment, creating data dictionaries and building applications involves teams of typically 6-12 analysts, data modelers, programmers and security specialists. By comparison, Linked Data applications are typically modeled within a 30 day sprint and applications can be created in hours. With emerging tools that are commercially supported, developers can host Linked Data applications on the cloud. Thus, within two months a reasonably complex data set can modeled, converted as Linked Data, and made available with powerful navigation and visualization features in less than two months on average.

I've heard that in the future, graph data may make relational data bases obsolete. Is that true? Please explain.

As a manager, should I consider developing in-house expertise to assist program SMEs with producing Linked Data? If I do, how can my organization assess the quality of what is being produced?

It is likely that your organization already has contractors and in-house staff familiar with the organization’s important data assets. Familiarity with Linked Data tools and techniques will come with time and should not be considered daunting.

Once data is converted to RDF, there are Web based tools and interfaces to explore and view the data. One of the important features of Linked Data is that, developers can programmatically query the data through a SPARQL endpoint, allowing them to view the content. This is similar but more flexible to a “view” in SQL. SPARQL query capability can be locked down for use by only authenticated personnel, or can be made available more widely, depending upon the use case. There techniques that are identical to the validation process performed on relational data by developers and data curators.

Their are different for Linked Data, however the concepts for data validation are similar to anyone who is a relational database professional. Your data experts will recognize the data in the RDF format. There are Linked Data tools, such as Callimachus, an Open Source Linked Data management system, to “follow your nose” and explore the data which is very powerful. Callimachus has a wiki-like interface, and a Class-based template engine that allows you to visualize and create Linked Data easily and quickly. With proper authentication, a user can update the underlying data in the graph database which is very useful.

Access to a small agile team who can guide and work within the agency on Linked Data issues is advisable. Practically speaking, not every agency will be able to have an in-house expert on Linked Data. However, if the agency has an office of information access and/or management (such as EPA’s OEI), it is logical that this team’s responsibilities would include participation with other agencies, standards groups (W3C, others), and at conferences discussing best practices with agency and other teams supporting the agency’s mission.

**What should the next steps be? What (if any) training would our staff need?**

Although Linked Data is no more complex than traditional data modeling, it does require a different way of thinking focused on expressing relationships through URIs. Just as an agency works with an in-house or contractor data modeling expert, the same would be true with Linked Data. There is both data subject domain expertise required, as well as, specialization in data modeling strategy and tactics.

Introductory training, best applied to small groups over a few days typically scheduled over 6-8 calendar weeks, is needed to discuss the differences between traditional vocabulary development and modeling approach for Linked Data. Experience has shown that these new generic techniques are then best supported and developed with situation specific workshops and/or mentoring as confidence grows.

**What hardware/software purchases are necessary?**

Many of the tools used for Linked Data are open source, including simple scripts and operating system commands, the use of which is openly shared within a community on the web. There is no cost associated with such tools. The storing and publishing of Linked Data can be handled by a simple web server. However, many more benefits flow from being able query that data, which requires it to be held in a linked data store, or RDF database. These are available as open source or proprietary services that you can host yourself or as a platform-as-a-service (PaaS) managed service.

Correct configuration of Web servers to publish Linked Data (e.g. using correct Content-Type information) is essential to reuse. Failure to understand Web standards can compromise an otherwise useful implementation. Therefore, care should be taken to have Linked Data reviewed by someone with relevant experience.

What human capital / infrastructure needs are required to support this kind of work?

This depends on the size of organization, the amount of data, and the rate of change to data. Experience has shown that having a small group/team of Linked Data aware people who can evangelize, help, support, guide and monitor a wider organization works well.

**What are the steps to creating a data-driven application?**

In practice, this approach requires speaking with one group at a time and exposing each RDBMS via either as real-time SPARQL query or periodic dump that is converted to an RDF format. Next, applications are rapidly created by Web developers using data-driven application tools and one or more Linked Data sets.

**Open data strategy will help tackle poverty and reduce corruption**

Sir Tim Berners-Lee, founder of the World Wide Web, has called on governments around the world to open up their data to the public in the same way that the UK has.

The internet pioneer believes that releasing publicly held data to the public and software developers will help fight poverty, boost innovation, empower citizens and reduce corruption.

Speaking at the Open Government Partnership Summit in London, on 4th November 2013, Berners-Lee revealed a report called the Open Data Barometer 2013, which shows that the UK has the most advanced open data policy.

---

The report, a joint project between the World Wide Web Foundation and Open Data Institute, is an investigation into how many countries are allowing open data policies.

“The Open Data Barometer is an important part of the puzzle and a good snapshot of where we are now,” claimed Berners-Lee. “What’s been brilliant in the UK is we got a project in the Cabinet Office to put a lot of data online quickly.”

However, Berners-Lee said the UK is “only 20% of the way there” when it comes to open data, adding that the nation has some serious challenges ahead. “There’s a lot to do but there’s a massive agenda,” he said, pointing out that there’s no open data that allows you to turn a postcode or address into a latitude and a longitude.

In the report, the US, Sweden, New Zealand, Denmark and Norway came immediately after the UK, out of the 77 countries surveyed.

Berners-Lee said only one in 10 of the countries that promised to open up data has actually delivered. More than half (55%) of the countries included on the report have formal open data policies in place, but many of these governments will not release certain datasets, including company registers and land registers that could provide valuable information to the public and mapping data that could be used by developers behind apps like Citymapper.

Criticisms of Open data Initiatives

While the arguments concerning the benefits of open data are well established and include contentions that open data lead to increased transparency and accountability with respect to public bodies and services; increases the efficiency and productivity of agencies and enhances their governance; promotes public participation in decision making and social innovation; and fosters economic innovation and job and wealth creation. There are potential problems affecting, and negative consequences of, open data initiatives. Four criticisms\(^\text{232}\) include:

a. Open data lacks a sustainable financial model
b. Promotes a politics of the benign and empowers the empowered
c. Lacks utility and usability
d. Facilitates the neoliberalisation and marketisation of public services.

While these critiques do not suggest abandoning the move towards opening data, they do suggest that open data initiatives need to be much more mindful of what data are being made open, how data are made available, how they are being used, and how they are being funded. For SmartOpenData these issues will be explored in the WP6 Evaluation and Assessment with User Groups, and the WP7 Exploitation sustainability plans.

**Global Open Data Initiative Declaration**

**Declaration A Citizens’ Call to Action on Open Data\(^\text{233}\)**


\(^{232}\) Taken from [http://www.nuim.ie/progcity/2013/11/four-critiques-of-open-data-initiatives](http://www.nuim.ie/progcity/2013/11/four-critiques-of-open-data-initiatives)

\(^{233}\) From [http://globalopendatainitiative.org/declaration/#sthash.81JKygZr.dpuf](http://globalopendatainitiative.org/declaration/#sthash.81JKygZr.dpuf) Comments are invited on this declaration in the current commentable version of the above. See also their Declaration announcement blog post.
Preamble

Governments exist “by and for the people”. The data they collect (or fund others to collect) in the course of carrying out their statutory duties also belongs to the people, and in the 21st century it is fast becoming one of the most valuable public goods we have – yet it often remains inaccessible or unaffordable to the vast majority. The Global Open Data Initiative aims to make Government data openly available to all – available for anyone, anywhere to download, use, re-use and redistribute without charge for any purpose.

We welcome government and multi-stakeholder efforts to advance open government data, and we seek to contribute to their success. However, to ensure that such efforts deliver real and sustained benefits for citizens, it is essential that civil society comes to the table with its own strong vision, ideals and demands. The Global Open Data Initiative seeks to engage and unite as broad a civil society constituency in a shared vision of the role of open data in accountable, inclusive and participatory governance.

In a well-functioning democratic society, citizens need to know what their government is doing. To do that, they must be able freely to access government data and information and to share that information with other citizens. Citizens’ core right to open government data arises from its increasingly critical role in enabling us to hold our governments accountable for fulfilling their obligations, and to play an informed and active role in decisions that affect us.

In addition, opening up government data creates new opportunities for SMEs and entrepreneurs, drives improved efficiency within government, and advances scientific progress. The initial costs (including any lost revenue from licenses and access charges) will be repaid many times over by the growth of knowledge and innovative data-driven businesses and services that create jobs, deliver social value and boost GDP.

We call on governments everywhere to take measurable, time-bound steps to:

1. Make data open by default:
   - Government data should be open by default, and this principle should ultimately be entrenched in law. Open means that data should be freely available for use, reuse and redistribution by anyone for any purpose and should be provided in a machine-readable form (specifically it should be open data as defined by the Open Definition and in line with the 10 Open Data Principles). Government information management (including procurement requirements and research funding, IT management, and the design of new laws, policies and procedures) should be reformed as necessary to ensure that such systems have built-in features ensuring that open data can be released without additional effort. Non-compliance, or poor data quality, should not be used as an excuse for non-publication of existing data. Governments should adopt intellectual property and copyright policies that encourage unrestricted public reuse and analysis of government data.

2. Make the process people-centered (or “put the users first”):
   - Experience shows that open data flounders without a strong user community, and the best way to build such a community is by involving users from the very start in designing and developing open data systems. Within government: The different branches of government themselves (including the legislature and
judiciary, as well as different agencies and line ministries within the executive) stand to gain important benefits from sharing and combining their data. Successful open data initiatives create buy-in and cultural change within government by establishing cross-departmental working groups or other structures that allow officials the space they need to create reliable, permanent, ambitious open data policies. Beyond government: Civil society groups and businesses should be considered equal stakeholders alongside internal government actors. Agencies leading on open data should involve and consult these stakeholders – including technologists, journalists, NGOs, legislators, other governments, academics and researchers, private industry, and independent members of the public – at every stage in the process. Stakeholders both inside and outside government should be fully involved in identifying priority datasets and designing related initiatives that can help to address key social or economic problems, foster entrepreneurship and create jobs. Government should support and facilitate the critical role of both private sector and public service intermediaries in making data useful.

3. **Provide no-cost access:**
   - One of the greatest barriers to access to ostensibly publicly-available information is the cost imposed on the public for access—even when the cost is minimal. Most government information is collected for governmental purposes, and the existence of user fees has little to no effect on whether the government gathers the data in the first place. Governments should remove fees for access, which skew the pool of who is willing (or able) to access information and preclude transformative uses of the data that in turn generates business growth and tax revenues. Governments should also minimise the indirect cost of using and re-using data by adopting commonly owned, non-proprietary (or “open”) formats that allow potential users to access the data without the need to pay for a proprietary software license. Such open formats and standards should be commonly adopted across departments and agencies to harmonise the way information is published, reducing the transaction costs of accessing, using and combining data.

4. **Put accountability at the core:**
   - Open Data needs to mean more than selective release of the datasets that are easiest or most comfortable for governments to open. It should empower citizens to hold government accountable for the performance of its core functions and obligations. At a minimum, governments should release datasets that are fundamental to citizen-state accountability and underlie key policy debates and decisions, including: (TBD list of data priorities goes here) Governments should create comprehensive indices of existing government data sets, whether published or not, as a foundation for new transparency policies, to empower public scrutiny of information management, and to enable policymakers to identify gaps in existing data creation and collection.

5. **Invest in capacity:**
   - Governments should start with initiatives and requirements that are appropriate to their own current capacity to create and release credible data, and that
complement the current capacity of key stakeholders to analyze and reuse it. At the same time, in order to unlock the full social, political and economic benefits of open data, all stakeholders should invest in rapidly broadening and deepening capacity. Governments and their development partners need to invest in making data simple to navigate and understand, available in all national languages, and accessible through appropriate channels such as mobile phone platforms where appropriate. Governments and their development partners should support training for officials, SMEs and CSOs to tackle lack of data and web skills, and should make complementary investments in improving the quality and timeliness of government statistics.

6. Improve the quality of official data:

- Poor quality, coverage and timeliness of government information – including administrative and sectoral data, geospatial data, and survey data – is a major barrier to unlocking the full value of open data. Governments should develop plans to implement the Paris21 2011 Busan Action Plan, which calls for increased resources for statistical and information systems, tackling important gaps and weaknesses (including the lack of gender disaggregation in key datasets), and fully integrating statistics into decision-making. Governments should bring their statistical efforts into line with international data standards and schemas, to facilitate reuse and analysis across various jurisdictions. Private firms and NGOs that collect data which could be used alongside government statistics to solve public problems in areas such as disease control, disaster relief, urban planning, etc. should enter into partnerships to make this data available to government agencies and the public without charge, in fully anonymized form and subject to robust privacy protections.

7. Enact legal and political reforms to create more open, transparent and participatory governance:

- Open government data cannot do its job in an environment of secrecy, fear and repression. Creating and defending open and participatory forms of governance is an ongoing challenge that requires constant work, scrutiny and engagement and there is no country that can claim to have perfected it. Governments should uphold basic rights to freedom of expression, information and association, and implement robust safeguards for personal privacy, as outlined in the UN Covenant on Civil and Political Rights. In addition, in line with their commitments in the UN Millennium Declaration (2000) and the Declaration of the Open Government Partnership (2011), they should take concrete steps to tackle gaps in participation, inclusion, integrity and transparency in governance, creating momentum and legitimacy for reform through public dialogue and consensus.
Annex B: The SmartOpenData Pilots

The activities of the SmartOpenData project’s 5 pilots, as described in the DoW, will be as follows:

1. Spain & Portugal - Agroforestry Management

This pilot will focus on building a web based collaborative spatial data infrastructure prototype with the main goal of promoting sustainable agroforestry management. It will be built as a collaborative powerful tool for environment protection and economic development of rural areas, and as a key factor for water management and drinking water protection.

The prototype will use Open Data from environmental INSPIRE compliant databases, cartographic services and remote sensing data, related with the objective of promoting monitoring and exploring agroforestry resources, and data that enable users (forest managers along with forest owners, public or private, and policy makers) to better understand agroforestry related resources management and to build consensus aligned with local and regional economic growth.

The pilot will mainly be focused on meeting the needs and requirements of the public sector regarding forest management and land use planning normative requirements. In both countries, public bodies from these areas are involved in the pilot, have defined their requirements and will provide data and assess project results as final users.

Spanish and Portuguese public bodies count on environmental and geospatial information that can be very valuable for management purposes. This information has a huge potential, but it is usually not widely accessible and not sufficiently exploited, and needs to be reorganised accordingly.

The final goal of this pilot is to provide an easy access to this information and to develop decision making tools and services available and helpful for public and private agroforestry managers. Results and outcomes will be open, standardised and of public access including web services and information management applications. The technological approach will be based on Open Data (environmental monitoring data, cartographic services and remote sensing products) linked through INSPIRE compliant semantic services. This strategy based on Open Data is necessary in order to optimise the use of public information, improve management issues and involve stakeholders in decision making processes.

The pilot will be located in a transboundary region which includes Ourense province in the NorthWest of Spain, and the region surrounding the Castelo the Bode Dam, in the Centre of Portugal. It will develop a double approach: (i) regional, focused in the province of Ourense and the region of the Castelo the Bode dam (sustainable forest management), and (ii) local, focused in the village of Maceda in Ourense-Spain (sustainable exploitation of forest resources).

Public bodies involved in the project need to (i) Optimise the use of agroforestry lands; (ii) Diversify rural economy and improve quality of life in rural communities; (iii) Enhance management practices of agroforestry lands and products; (iv) Preserve biodiversity and
environment; (v) Involve public and private stakeholders in forest management issues; (vi) Improve the efficiency of administrative processes; (vii) Improve watershed management by including land use and agroforestry related topics.

2. Ireland - Environmental Research and Biodiversity

This Pilot will focus on the use of the SmartOpenData infrastructure to provide open data and open INSPIRE-compliant geospatial sources for environmental researchers particularly focused on biodiversity and habitats, building on the participative social validation and pilots of the HABITATS project in particular. This will overlap with the Public Sector pilot (decision makers, parks environmental monitoring, cartographic services and remote sensing), the Citizens Pilot (similar but more technical sources of open data), Enterprise Pilot (exploitation open public data sources to create new value added services).

The pilot will access and aggregate sources to impact on the area of biodiversity, by allowing them to seamlessly bridge the major gap between the “worlds” of open data and INSPIRE geo-spatial sources, to validate a major value-add and impact on their work. Some of the main issues to be addressed include (i) Discovery and seamless use, and mashing together of all available sources to address immediate research issues; (ii) Overcoming the barriers (cultural, political, administrative) to opening up the data; (iii) Overcoming technical incompatibilities of datasets in terms of technical standards, semantic structuring etc. (iv) Validation of the SmartOpenData platform in the aggregation, analysis, and visualisation to support decision making of the various research and other stakeholders requirements, and the value-add/impact on their work.

This will use and complement the Biodiversity MashUp Linked Open Data Extension and Spatial Web Crawler of the Citizens pilot but will focus more on identifying and seamlessly using various sources at International, European, National and local levels by researchers. This will be complemented by various social networking and crowdsourcing mobile apps to engage stakeholders at the local level in particular.

The pilot will begin at European level by using and seamlessly integrating the SmartOpenData platform into, collaborating with, and building on various open data and geo-spatial sources and initiatives that will have a particular value for biodiversity researchers, including (i) The European Biodiversity Observation Network, EUBON project\textsuperscript{234}; (ii) European Environmental Agency (EEA), Biodiversity data centre (BDC)\textsuperscript{235}; (iii) PESI\textsuperscript{236}; (iv) FP7 EUBrazilOpenBio\textsuperscript{237}; (v) LifeWatch European research infrastructure\textsuperscript{238}; (vi) The Joinup Portal\textsuperscript{239}; (vii) EU Open Data Portal\textsuperscript{240}, (viii) The UK environmental Agency’s Datashare\textsuperscript{241}; (ix) The Global Biodiversity Information Facility (GBIF)\textsuperscript{242}. This pilot will also use SmartOpenData to mash disparate open data sources such as those listed at FreeGISdata\textsuperscript{243} and the UK

\textsuperscript{234} www.earthobservations.org/geobon.shtml
\textsuperscript{235} www.eea.europa.eu/themes/biodiversity/dc
\textsuperscript{236} www.eu-nomen.eu/portal
\textsuperscript{237} www.eubrazilopenbio.eu
\textsuperscript{238} www.lifewatch.eu
\textsuperscript{239} http://joinup.ec.europa.eu/catalogue
\textsuperscript{240} http://open-data.europa.eu/en/data/dataset
\textsuperscript{241} www.geostore.com/environment-agency/WebStore
\textsuperscript{242} www.gbif.org
\textsuperscript{243} http://freegisdata.rtwilson.com
GeoStore\textsuperscript{244}, which allows use subject to the terms and conditions of the Open Government Licence (OGL), and others, in various data formats. The implications of these for researchers will be explored in the pilot and supported using the platform.

At a national level this pilot will initially focus on the Burren National Park in Ireland\textsuperscript{245}. The Burren National Park is located in the southeastern corner of the Burren and is approximately 1500 hectares in size. The Park land was bought by the Government for nature conservation and public access. It contains examples of all the major habitats within the Burren: Limestone Pavement, Calcareous grassland, Hazel scrub, Ash/hazel woodland, Turloughs, Lakes, Petrifying springs, cliffs and Fen.

The highest point in the park is Knockanes (207 metres) which continues as a curving terraced ridge to Mullaghmór to the south. East of this ridge is an area of extensive, low lying limestone pavement containing a number of semi-permanent lakes. West of this ridge the pavement sweeps down to partially drift-covered ground which gradually rises again to reach the foot of a rocky escarpment. To the south of the park the limestone bedrock disappears under a layer of glacial till. This till area is far more intensively managed for pasture and silage.

The MWRA, with the technical support of MAC, will work with Irish public agencies, Local Authorities and community groups associated with the Burren in the implementation and use of the SmartOpenData platform in the WP5 validation trials to demonstrate its value in helping researchers, decision makers and communities to better manage, preserve, maintain and use this unique ecosystem. These will include:

- Local Authorities
  - Clare\textsuperscript{246}, Limerick\textsuperscript{247} and North Tipperary\textsuperscript{248} Councils
- National Parks & Wildlife Service\textsuperscript{249}.
- Who manage the Burren and other National Parks in Ireland.
- Burren Communities\textsuperscript{250}
  - Burrenbeo Teo Trust\textsuperscript{251}
  - BurrenLIFE project\textsuperscript{252}
  - Burren & Cliffs of Moher Geopark\textsuperscript{253}
  - Burren Ecotourism Network (BEN)\textsuperscript{254}

Their validation of SmartOpenData value-add/impact on their work will involve using the platform to access sources such as (i) The Irish National Parks & Wildlife Services (NPWS) who have extensive open online maps and datasets\textsuperscript{255}; (ii) The National Biodiversity Data Centre Ireland\textsuperscript{256}, (iii) The Irish Opendata Portal\textsuperscript{257}; (iv) The All-Island Research Observatory

\textsuperscript{244} http://www.geostore.com/environment-agency/WebStore?xml=environme
\textsuperscript{245} www.burrennationalpark.ie
\textsuperscript{246} www.clarecoco.ie
\textsuperscript{247} www.lcc.ie and www.limerickecity.ie
\textsuperscript{248} www.tipperarynorth.ie
\textsuperscript{249} www.npws.ie
\textsuperscript{250} See for instance, the Burren Community Charter at www.heritagecouncil.ie/fileadmin/user_upload/Publications/Wildlife/Draft_Burren_Community_Charter.pdf
\textsuperscript{251} www.burrenbeo.com
\textsuperscript{252} www.burrenlife.com
\textsuperscript{253} www.burrengeopark.ie
\textsuperscript{254} www.burrenecotourism.com
\textsuperscript{255} www.npws.ie/mapsandddata
\textsuperscript{256} www.biodiversityireland.ie
(AIRO)\textsuperscript{258}, (v) The Irish Spatial Data Infrastructure GeoPortal\textsuperscript{259}, (vi) The Irish Spatial Data Exchange (ISDE)\textsuperscript{260}, (vii) The Irish Heritage Council heritage maps\textsuperscript{261}, and (viii) The Marine Institute Ireland, who have extensive OGC/INSPIRE compliant geo-spatial data\textsuperscript{262}, which are all searchable using the ISDE Browser\textsuperscript{263}. In addition there are a number of interactive services available to the public, such as the vessel tracking service and Survey Planning System\textsuperscript{264}.

Even though all of these sites have open data, they are not seamless (e.g. the Irish Biodiversity Data Centre needs the latest version of Silverlight). This will be an issue to explored using the SmartOpenData platform and validate its seamless access and mashability of sources for researchers.

3. Italy - Water Monitoring

This pilot in Sicily will explore the role of aggregating information from different Open Data sources in order to support ARPA’s institutional mission of providing up to date monitoring of water quality in Sicily. Some of the main issues to address are

(i) Overcoming the barriers (cultural, political, administrative) to opening up the data;
(ii) Identifying the optimal role of the general public in crowdsourcing environmental information;
(iii) Identifying the technical means of publication;
(iv) Overcoming incompatibilities of datasets in terms of technical standards (eg JSON vs XML), semantic structuring, etc.;
(v) Optimum ways of aggregation, analysis, and visualisation to support decision-making for the different stakeholders including the general public.

The three main stakeholder groups involved are:

a) Different offices within A.R.P.A. Sicilia itself, in particular the link between the central office in Palermo and the local offices distributed throughout the region;
b) Key external stakeholders with a role in environmental monitoring, including The ASL (local health care divisions), Ex-ATO (water supply at sub-regional level) and City administrations;
c) The general public providing information through eg Geoblogging

The pilot activities will be articulated in three specific sites that represent typical situations, for example:

1. Coastal area: issues of pollution from wastewater, formation of dangerous algae.
2. Riverbed: wastewater pollution, groundwater spill off from agricultural chemicals.
3. An urban water system: water supply, water treatment, drainage sewers, etc.

\textsuperscript{257} www.opendata.ie
\textsuperscript{258} www.airo.ie
\textsuperscript{259} www.geoportal.ie
\textsuperscript{260} www.isde.ie
\textsuperscript{261} http://www.heritagecouncil.ie/heritage-maps/heritage-maps/
\textsuperscript{262} www.marine.ie/home/publicationsdata
\textsuperscript{263} http://catalogue.isde.ie
\textsuperscript{264} www.marine.ie/home/services/researchvessels
The pilot methodology will be based on a Living Lab approach engaging the key stakeholders in the co-design of the optimal solution. Elements can be integrated from the CIP Smart City “Citadel” project\(^{265}\) such as the local Open Data Governance Group and the use of an Open Data Commons of re-useable tools and APIs. SpazioDati will deploy a white-labelled version of its data market and customise it to implement the features needed by the pilot, to host and distribute all project data. This includes designing and implementing dataset-specific API that hide the details of general-purpose query languages and APIs (such as SPARQL and JSON-LD) to maximise ease of use for developers. Additionally the market will allow non-developers to easily generate data visualisations, which can be reused in reports and presentations.

The datasets that A.R.P.A. Sicilia can provide include:

1. Presence of toxic algae (Ostreopsis spp.)
2. Presence pollutants from sewage in coastal waters
3. Presence pollutants from exhaust (rivers, lakes and transitional waters)
4. Presence of residual agricultural chemicals
5. Groundwater monitoring
6. Wastewater treatment (sewage treatment plants, wetlands, etc.).

and others.

The format of the dataset is often not open and the number of monitoring stations is variable for each dataset, however, they are formed, at least, for hundreds of points.

The main parameters measured are:

- Temperature, pH, salinity, dissolved oxygen;
- Unit cell of O. ovata, Coolia monotis, Prorocentrum lima, Amphidinium spp.
- A chlorophyll, turbidity and nutrients (nitrogen ammonia (NH4), A.Nitrico (NO3), total phosphorus (Pt));
- Macrophytes, macroinvertebrates, diatoms;
- Pollutants.

The sampling is carried out in different periods and the measured parameters are processed by A.R.P.A. Sicilia in order to define the state of the quality of surface water and groundwater bodies, providing technical and scientific support for the protection, conservation and the achievement of environmental quality objectives. The excessive presence of toxic algae along the coasts cause significant effects on the benthic and, therefore, the integrity of the marine flora and fauna as well as human health.

### 4. Czech Republic - Forest Sustainability

This pilot is focused on the forest site classification, sustainable management and utilisation of forest road network using the National Forest Inventory and the Regional Plans for Development datasets. Data products and statistical outcomes will be widely open, standardised and accessible by foresters and public bodies through web services and applications\(^{266}\).

The Czech National Forest Inventory (NFI) Geoportal will present valuable information derived from NFI database. Our data infrastructure can support SME’s decision-making.

\(^{265}\) [www.citadelonthemove.eu](http://www.citadelonthemove.eu)

\(^{266}\) See [http://forestportal.efi.int/listg.php?c=CZ](http://forestportal.efi.int/listg.php?c=CZ) and [http://forestportal.efi.int/content/view/8/14/](http://forestportal.efi.int/content/view/8/14/)
process via provision of reliable, up-to-date information about forests and its wood as well as other types of resources. The geoportal will also enhance public awareness about the NFI2 project funded by Czech Government, as well as the awareness about the UHUL FMI institute itself. Data and metadata will be aggregated and standardised according to commonly used standards and SmartOpenData recommendations, so they will be accessible on a shared project platform. The geoportal will be a gate to results of the NFI1 and NFI2 projects, which are derived by statistical sound methods using a huge, high quality database fusing field survey, photogrammetric interpretation and remote sensing data. Data quantity, its quality and a thematically broad scope of the UHUL FMI activities is an asset for testing semantic search approaches among our data services and other environmental organisation sources across EU. UHUL FMI manages data covering the entire CZ territory, some of them are freely available, however free access to the rest is limited to an aggregated level. It might be useful to include also data with restricted access into semantic search tools and bring them closer to SMEs—at least on the aggregated level\(^{267}\).

HSRS will contribute to pilots with focus on discovery and geovisualisation of Open Spatial Data. HSRS will bring experience from implementation of many global, national, regional and local SDIs. HSRS will cooperate with UHUL FMI, SAZP and IMCS.

CCSS will participate on pilot activities related to environment protection, environmentally friendly tourism, environmental research and education. Here will mainly apply experiences from past projects NaturNet Redime, NaturNet plus\(^{268}\), and EnviroGrids BlackSee\(^{269}\). CCSS will also support linkage of pilots with GEOSS and UNSDI activities.

5. **Slovakia-Environmental Data reuse.**

This pilot will include the development and deployment of two conceptually different types of web applications in order to achieve reuse of environmental data and information in line with European Open Data Strategy:

1. **Biodiversity MashUp Linked Open Data Extension:**
   - Extension of existing Biodiversity MashUp web 2.0 application with web 3.0 semantic linked open data dimension. The initial version of this MashUp application was the result of effort to show the possibility to link various biodiversity related information, where selected birds species were combined with their spatial distribution and descriptive information from various sources (national spatial data providers with information from Encyclopaedia of Life\(^{270}\)). The main aim of the pilot is to make current application content available in RDF encoding allowing linking available content with other sources in a way to be read and queried by human readers as well as computers.
   - The geographical and content scope can be extended with the coverage of the countries represented in the project consortium as well as relevant information resources such as PESI (Pan-European Species Directories Infrastructure)\(^{271}\).

2. **Spatial Web Crawler:**

---


\(^{268}\) [http://www.naturnet.org/simplecms/?menuID=1&action=article&presenter=Article](http://www.naturnet.org/simplecms/?menuID=1&action=article&presenter=Article)

\(^{269}\) [http://envirogrids.net/](http://envirogrids.net/)

\(^{270}\) [http://eol.org/](http://eol.org/)

\(^{271}\) [www.eu-nomen.eu/portal/](http://www.eu-nomen.eu/portal/)
A web-based application allowing a query definition for a particular spatial and thematic domains related information resources available in the deep web retrieved from indexing processes in the context of the most used search engines (e.g., Google) and making this content available for SDI as well as linked data world. Nowadays a geospatial user within an SDI searches for geospatial information using discovery clients of a geoportal application (such as the INSPIRE Geoportal\textsuperscript{272}). If the data producer wants to promote related resources and make them available in SDI, they need to create metadata according to predefined rules (e.g., the INSPIRE metadata regulation) and publish them using CSW standard.

This approach allows either for distributed searches or harvesting. Nevertheless, there is still very many data producers, making their resources available without documenting and publishing through the CSW interface.

The pilot proposes a complex solution, which aims at the design and implementation of a framework allowing users to discover and provide an access to geospatial resources that exist in the deep web which is not directly discoverable within existing SDIs (i.e., INSPIRE, GEOSS) or by surface web search engines (i.e., Google). In addition, a semantic extension based on Linked Data principles will be taken into the consideration of content collected by this pilot framework will be published using a standard RDF format. This step should close the circle by interconnecting across different levels of the geospatial web and at the same time enhance use of current open data principles.

\footnote{\text{272} http://inspire-geoportal.ec.europa.eu/}
## Annex C: OPQUAST OpenData best practices

<table>
<thead>
<tr>
<th>No</th>
<th>Type</th>
<th>Open Data Best Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Animation</td>
<td>The concept of Open Data is explained</td>
</tr>
<tr>
<td>2</td>
<td>Animation</td>
<td>The site provides a way for users to keep informed of updates to the data</td>
</tr>
<tr>
<td>3</td>
<td>Animation</td>
<td>The site provides an online channel for feedback</td>
</tr>
<tr>
<td>4</td>
<td>Animation</td>
<td>Examples of reuse (fictitious or real) are provided</td>
</tr>
<tr>
<td>5</td>
<td>Animation</td>
<td>A news feed provides information about updates to each dataset</td>
</tr>
<tr>
<td>6</td>
<td>Animation</td>
<td>Datasets are accompanied by links or resources that provide user guidance and support</td>
</tr>
<tr>
<td>7</td>
<td>Animation</td>
<td>There is a discussion forum about the datasets</td>
</tr>
<tr>
<td>8</td>
<td>Animation</td>
<td>It is possible to request or suggest the publication of new datasets</td>
</tr>
<tr>
<td>9</td>
<td>Animation</td>
<td>The site enables users to rate and/or evaluate the datasets</td>
</tr>
<tr>
<td>10</td>
<td>API</td>
<td>The mechanisms for accessing and interacting with datasets are documented</td>
</tr>
<tr>
<td>11</td>
<td>API</td>
<td>If there are conditions on access to a dataset, the method of retrieving the data is documented</td>
</tr>
<tr>
<td>12</td>
<td>API</td>
<td>A specific API command allows the user to determine a dataset’s licence</td>
</tr>
<tr>
<td>13</td>
<td>API</td>
<td>Dynamic and static versions of the datasets are synchronised</td>
</tr>
<tr>
<td>14</td>
<td>API</td>
<td>When data content exists in multiple formats, the server provides a mechanism for content negotiation</td>
</tr>
<tr>
<td>15</td>
<td>API</td>
<td>If web services for data access are available, the current version of each web service is specified</td>
</tr>
<tr>
<td>16</td>
<td>API</td>
<td>No system is set up to prevent automatic or programmatic data retrieval</td>
</tr>
<tr>
<td>17</td>
<td>API</td>
<td>Each dataset is available through at least one web service</td>
</tr>
<tr>
<td>18</td>
<td>API</td>
<td>The metadata catalogue can be retrieved using a standard protocol</td>
</tr>
<tr>
<td>19</td>
<td>Applications</td>
<td>The site enables users to add links to relevant applications</td>
</tr>
<tr>
<td>20</td>
<td>Applications</td>
<td>The site lists existing applications already developed from the datasets available</td>
</tr>
<tr>
<td>21</td>
<td>Applications</td>
<td>Users can keep informed about applications</td>
</tr>
<tr>
<td>22</td>
<td>Metadata</td>
<td>The identity and role of the person responsible for each dataset is specified</td>
</tr>
<tr>
<td>23</td>
<td>Metadata</td>
<td>Each dataset is accompanied by a descriptive record</td>
</tr>
<tr>
<td>24</td>
<td>Metadata</td>
<td>Each dataset includes at least a title and a description</td>
</tr>
<tr>
<td>25</td>
<td>Metadata</td>
<td>A creation date is given for each dataset</td>
</tr>
<tr>
<td>26</td>
<td>Metadata</td>
<td>A last-updated date is given for each dataset</td>
</tr>
<tr>
<td>27</td>
<td>Metadata</td>
<td>The datasets are categorised</td>
</tr>
<tr>
<td>28</td>
<td>Metadata</td>
<td>Metadata associated with each dataset is available in a standard format</td>
</tr>
<tr>
<td>29</td>
<td>Metadata</td>
<td>Each dataset is accompanied by a reference to the language used</td>
</tr>
<tr>
<td>30</td>
<td>Metadata</td>
<td>Metadata describing the datasets is structured in a standard way</td>
</tr>
<tr>
<td>31</td>
<td>Metadata</td>
<td>If a dataset refers to a specific geographical area, its spatial coverage is specified</td>
</tr>
<tr>
<td>32</td>
<td>Metadata</td>
<td>Topics are linked to published vocabularies and taxonomies</td>
</tr>
<tr>
<td>33</td>
<td>Format</td>
<td>Each dataset includes a reference to the charset used</td>
</tr>
<tr>
<td>34</td>
<td>Format</td>
<td>The format of downloadable files is indicated</td>
</tr>
</tbody>
</table>

273 From [http://checklists.opquast.com/en/opendata](http://checklists.opquast.com/en/opendata), Those good practices are published under the Creative Commons BY-SA license. You're free use them as you see fit, even for a commercial use, the chosen licence makes you a co-owner of its contents. On the other hand, you need to maintain it’s paternity, thank you to quote its origin, or even better add a link to Opquast’s website.
<table>
<thead>
<tr>
<th>No</th>
<th>Type</th>
<th>Open Data Best Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>Format</td>
<td>Dates are given in documented formats</td>
</tr>
<tr>
<td>36</td>
<td>Format</td>
<td>Dates are available in a standard format</td>
</tr>
<tr>
<td>37</td>
<td>Format</td>
<td>Data is provided in at least one open format</td>
</tr>
<tr>
<td>38</td>
<td>Format</td>
<td>The character encoding used in each dataset is declared</td>
</tr>
<tr>
<td>39</td>
<td>Format</td>
<td>Where data is provided in a standard format, the format is specified and adhered to</td>
</tr>
<tr>
<td>40</td>
<td>Format</td>
<td>The version or syntax (doctype) or file formats used is available</td>
</tr>
<tr>
<td>41</td>
<td>Historique</td>
<td>Each dataset comes with a changelog</td>
</tr>
<tr>
<td>42</td>
<td>Historique</td>
<td>One can access different versions of datasets</td>
</tr>
<tr>
<td>43</td>
<td>Identification</td>
<td>The descriptive record contains a direct link to the URL of the data</td>
</tr>
<tr>
<td>44</td>
<td>Identification</td>
<td>Each dataset is given a unique identifier</td>
</tr>
<tr>
<td>45</td>
<td>Identification</td>
<td>A checksum and/or signature is available to verify the validity of each file</td>
</tr>
<tr>
<td>46</td>
<td>Identification</td>
<td>Datasets are accompanied by at least one means of contact for the producer (or maintainer)</td>
</tr>
<tr>
<td>47</td>
<td>License</td>
<td>The datasets are accompanied by a licence</td>
</tr>
<tr>
<td>48</td>
<td>License</td>
<td>The licence sets out the conditions of attribution, reuse, redistribution and commercialisation</td>
</tr>
<tr>
<td>49</td>
<td>License</td>
<td>Usage rights are provided for an unlimited period</td>
</tr>
<tr>
<td>50</td>
<td>License</td>
<td>Data producers declare their policy on releasing data</td>
</tr>
<tr>
<td>51</td>
<td>License</td>
<td>The datasets are accompanied by a summary and a link to the full version of the license</td>
</tr>
<tr>
<td>52</td>
<td>Linkeddata</td>
<td>Any vocabularies used within the dataset are identified and documented</td>
</tr>
<tr>
<td>53</td>
<td>Linkeddata</td>
<td>Data adheres to the defined syntax of any specified vocabularies</td>
</tr>
<tr>
<td>54</td>
<td>Linkeddata</td>
<td>It is possible to query data and metadata in accordance with standards of the web of data (Linked Open Data)</td>
</tr>
<tr>
<td>55</td>
<td>Linkeddata</td>
<td>Data sources linked from a dataset are reported</td>
</tr>
<tr>
<td>56</td>
<td>Naming</td>
<td>The names of data files contain only alphanumeric characters or characters considered safe</td>
</tr>
<tr>
<td>57</td>
<td>Naming</td>
<td>If the naming of datasets is the subject of a convention, the convention is made available</td>
</tr>
<tr>
<td>58</td>
<td>Transparency</td>
<td>If a dataset contains an expiry date, it is declared</td>
</tr>
<tr>
<td>59</td>
<td>Transparency</td>
<td>If a limit is set on the number of accesses or the amount of data accessed via web services, this limit is specified</td>
</tr>
<tr>
<td>60</td>
<td>Transparency</td>
<td>Any restrictions on access to the data are stated</td>
</tr>
<tr>
<td>61</td>
<td>Transparency</td>
<td>If a dataset contains errors or uncertainties or is incomplete, this fact is stated</td>
</tr>
<tr>
<td>62</td>
<td>Transparency</td>
<td>An update frequency is given for each dataset</td>
</tr>
<tr>
<td>63</td>
<td>Transparency</td>
<td>The publisher of the dataset provides information on its origin</td>
</tr>
<tr>
<td>64</td>
<td>Transparency</td>
<td>It is possible to obtain information regarding the level of trust accorded to the data</td>
</tr>
<tr>
<td>65</td>
<td>Usability</td>
<td>The size of downloadable files is indicated</td>
</tr>
<tr>
<td>66</td>
<td>Usability</td>
<td>It is possible to search the metadata</td>
</tr>
<tr>
<td>67</td>
<td>Usability</td>
<td>Extracts from the different datasets are available for download</td>
</tr>
<tr>
<td>68</td>
<td>Usability</td>
<td>If the site offers several licences, the search engine allows filtering by licence</td>
</tr>
<tr>
<td>69</td>
<td>Usability</td>
<td>The metadata catalogue is available as a dataset</td>
</tr>
<tr>
<td>70</td>
<td>Usability</td>
<td>An extract of the data can be previewed</td>
</tr>
<tr>
<td>71</td>
<td>Privacy</td>
<td>If the data contains personal information, its publication is in accordance with the legislation under which the producer operates</td>
</tr>
<tr>
<td>72</td>
<td>Privacy</td>
<td>The site provides a means to request removal or anonymisation of data</td>
</tr>
</tbody>
</table>
## Annex D: SMEspire Best Practice Catalogue

All Practices Sorted By Title in Ascending Order on 01/11/2013.

<table>
<thead>
<tr>
<th>Title</th>
<th>Keywords</th>
<th>Authors</th>
<th>Classification</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D CityGML renderer web Applet</td>
<td>map, data, images, ortophoto, GIS, web, applet, cityGML</td>
<td>Fondazione Graphitec</td>
<td>Geospatial Activities: Delivering GIS/SDI software (development, customization, etc), Processing, editing, modelling, analyzing and managing geodata, Phases of an SDI: Concept and Design, Architecture Reference Model Services: Human Interaction Services, Standards Implemented: OGC, EU Initiatives: i-Scope, Output Types: Tool</td>
<td>30/08/2013</td>
</tr>
<tr>
<td>Communication tools</td>
<td>brand, web site, communication</td>
<td>Lacosa Srl</td>
<td>Output Types: Tool</td>
<td>24/10/2013</td>
</tr>
<tr>
<td>Cross-border, multi-lingual and harmonized access to the Member States’ territorial data</td>
<td>Geoportal, NSDI, Cross-border data, Multilingual data, data harmonization, Open Source</td>
<td>Planetek Italia S.r.l.</td>
<td>Geospatial Activities: Delivering GIS/SDI software (development, customisation, etc), Processing, editing, modelling, analyzing and managing geodata, Measuring, collecting and storing geodata, Presenting, producing and distributing geodata, INSPIRE and SDI Related Activities: Spatial Data Services, View, Discovery, Monitoring and Reporting, Data, Phases of an SDI: Implementation, Standards Implemented: OGC, Output Types: Tool, Software, Procedure, License Types: Open Source, Customer Types: International</td>
<td>08/08/2013</td>
</tr>
<tr>
<td>Easy management of maps, geolocated data, and images / Maps: web, tablets, and mobile</td>
<td>map, data, images, public facilities, GIS, web, mobile,</td>
<td>Geographica</td>
<td>INSPIRE and SDI Related Activities: Spatial Data Services, View, Download, Monitoring and Reporting, Output Types: Tool, Software, Data Model, License Types: Enterprise, Open Source, Customer Types: Local</td>
<td>20/06/2013</td>
</tr>
<tr>
<td>EasySDI</td>
<td>SDI</td>
<td>Depth France</td>
<td>Geospatial Activities: Delivering GIS/SDI software (development, customization, etc), Processing, editing, modelling, analyzing and managing geodata, Presenting, producing and distributing geodata, Phases of an SDI: Implementation, Architecture Reference Model Services: Human Interaction Services, Workflow/Task Services,</td>
<td>26/09/2013</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Title</th>
<th>Keywords</th>
<th>Authors</th>
<th>Classification</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeoBrowser@A22</td>
<td>3D, Geovisualisation, Geovisual Analytics, LBS</td>
<td>Fondazione Graphitec</td>
<td>Geospatial Activities: Delivering GIS/SDI software (development, customization, etc), Processing, editing, modelling, analyzing and managing geodata Presenting, producing and distributing geodata INSPIRE and SDI Related Activities: Spatial Data Services Phases of an SDI: Implementation Standards Implemented: OGC EU Initiatives: Interreg Italy-Switzerland Output Types: Tool, Software, Report, Guidelines License Types: Open Source</td>
<td>16/09/2013</td>
</tr>
<tr>
<td>High quality real time environments reconstruction App for mobile tourism support</td>
<td>map, data, images, ortophoto, GIS, web, mobile, tourism</td>
<td>Fondazione Graphitec</td>
<td>Geospatial Activities: Delivering GIS/SDI software (development, customization, etc), Processing, editing, modelling, analyzing and managing geodata Measuring, collecting and storing geodata Presenting, producing and distributing geodata INSPIRE and SDI Related Activities: View, Discovery, Download, Transformation, Monitoring and Reporting, Data Standards Implemented: OGC Output Types: Tool, Software, Procedure License Types: Open Source Customer Types: Regional</td>
<td>28/06/2013</td>
</tr>
<tr>
<td>Intelligent Transportation</td>
<td>Urban Mobility, Mobile, Multimodal</td>
<td>Fondazione Graphitec</td>
<td>Geospatial Activities: Presenting, producing and distributing geodata</td>
<td>23/10/2013</td>
</tr>
<tr>
<td>Title</td>
<td>Keywords</td>
<td>Authors</td>
<td>Classification</td>
<td>Date</td>
</tr>
<tr>
<td>-------</td>
<td>----------</td>
<td>---------</td>
<td>----------------</td>
<td>------</td>
</tr>
<tr>
<td>System for Urban Environments</td>
<td>Transportation, Recommender Systems, Routing, Access Control</td>
<td>Graphitec</td>
<td>EU Initiatives: iTour Output Types: Tool, Software License Types: Free, Trial</td>
<td></td>
</tr>
<tr>
<td>MANTRA: How to turn Big Data into Smart Data</td>
<td>altitila, mantra, smart data, smartdata, smart data management platform, big data, semantic technologies, unstructured information management, text analytics, sentiment analysis, semantic etl, semantic master data management, business intelligence, ontologies, NoSQL, data capture, data integration</td>
<td>ALTILIA S.r.l.</td>
<td>Geospatial Activities: Processing, editing, modelling, analyzing and managing geodata Output Types: Software, Data Model License Types: Enterprise</td>
<td>28/10/2013</td>
</tr>
<tr>
<td>RAE – SDI in accordance to the INSPIRE Directive</td>
<td>SDI, open source, Geospatial, data, metadata</td>
<td>GET Ltd</td>
<td>Geospatial Activities: Processing, editing, modelling, analyzing and managing geodata, INSPIRE and SDI Related Activities: Data Phases of an SDI: Implementation, Validation Implementation Types: Project Standards Implemented: EN-ISO 191XX</td>
<td>16/10/2013</td>
</tr>
<tr>
<td>Title</td>
<td>Keywords</td>
<td>Authors</td>
<td>Classification</td>
<td>Date</td>
</tr>
<tr>
<td>-------</td>
<td>----------</td>
<td>---------</td>
<td>----------------</td>
<td>------</td>
</tr>
<tr>
<td>Results assessment report</td>
<td>assessment, harmonization</td>
<td>Aerogeod ezijos institutas</td>
<td>INSPIRE and SDI Related Activities: Data</td>
<td>12/06/2013</td>
</tr>
<tr>
<td>SIGMA TER</td>
<td></td>
<td></td>
<td>Geospatial Activities: Delivering GIS/SDI software (development, customization, etc), Advising, educating, researching, communicating about processes, use of geo-information products, Processing, editing, modelling, analyzing and managing geodata</td>
<td>16/09/2013</td>
</tr>
<tr>
<td>SITR-IDT REGIONE SARDEGNA</td>
<td>INSPIRE, OGC, WMS, WFS, Geoportal, Region, Sardegna</td>
<td>SINERGIS SRL</td>
<td>INSPIRE and SDI Related Activities: View, Discovery, Download, Transformation, Data and Service Sharing, Data, Orchestration and Composition</td>
<td>16/09/2013</td>
</tr>
<tr>
<td>Spatio-temporal data management and interactive visualisation</td>
<td>Services, Spatio-temporal analysis, Web Processing Services, Civil</td>
<td>Fondazione Graphitech</td>
<td>Geospatial Activities: Delivering GIS/SDI software (development, customization, etc), Processing, editing, modelling, analyzing and managing geodata, Measuring, collecting and</td>
<td>23/10/2013</td>
</tr>
<tr>
<td>Title</td>
<td>Keywords</td>
<td>Authors</td>
<td>Classification</td>
<td>Date</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------</td>
<td>--------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>Thesaurus service using GEMET REST API</td>
<td>INSPIRE metadata, metadata resource keywords</td>
<td>InfoLogica</td>
<td>Geospatial Activities: Delivering GIS/SDI software (development, customization, etc), Processing, editing, modelling, analyzing and managing geodata, Measuring, collecting and storing geodata, Presenting, producing and distributing geodata INSPIRE and SDI Related Activities: Spatial Data Services, Monitoring and Reporting, Data License Types: Enterprise Customer Types: Local</td>
<td>28/10/2013</td>
</tr>
<tr>
<td>Validation Briefcase</td>
<td>data validation, schema validation, metadata validation</td>
<td>EPSILON ITALIA SRL</td>
<td>Architecture Reference Model Services: Human Interaction Services, System Management Services Standards Implemented: EN-ISO 191XX Output Types: Report</td>
<td>31/05/2013</td>
</tr>
<tr>
<td>Video of Mantra platform to process big-data</td>
<td>video, motion graphic, brand design, big data</td>
<td>Lacosa Srl</td>
<td>Geospatial Activities: Delivering GIS/SDI software (development, customization, etc), Processing, editing, modelling, analyzing and managing geodata INSPIRE and SDI Related Activities: Data Phases of an SDI: Validation Implementation Types: Project Standards Implemented: EN-ISO 191XX EU Initiatives: NatureSDIplus Output Types: Procedure, Guidelines License Types: Free</td>
<td>24/10/2013</td>
</tr>
<tr>
<td>WorldDroid³</td>
<td>WorldDroid³ Android Mobile Device</td>
<td>Trilogis</td>
<td>Geospatial Activities: Delivering GIS/SDI software (development, customization, etc), Advising, educating, researching, communicating about processes, use of geoinformation products, Presenting, producing and distributing geodata Architecture Reference Model Services: Human Interaction Services Output Types: Software License Types: Free, Open Source</td>
<td>25/10/2013</td>
</tr>
</tbody>
</table>
Annex E: List of Abbreviations

AJAX Asynchronous JavaScript and XML
API Application Programme Interface
CMS Content Management System
Copernicus - the European Earth Observation Programme, used to be known as GMES
CSS Cascading Style Sheets
CSW - Catalogue Service for the Web
DCAT-AP - Data Catalogue vocabulary Application Profile
DOM Document Object Model
DOW Description of Work, Annex I to the Grant Agreement.
EC – European Commission
ECMS Enterprise Content Management System
EU – European Union
FI Future Internet
GA SmartOpenData Grant Agreement.
GEOSS - Global Earth Observation System of Systems
GI Geospatial/Geographic Information
GIS Geographic/Geospatial Information Systems
GLOD - Geospatial Linked Open Data.
GMES - Global Monitoring for Environment and Security – now known as Copernicus
GML – Geography Markup Language
HTML Hypertext Markup Language
IDE Integrated Development Environment
INSPIRE – INfrastructure for SPatial InfoRmation in Europe
IOT Internet of Things
ISO – International Organisation for Standardisation
ISO 19118 – ISO 19118 Geographic Information-Encoding
JSON JavaScript Object Notation
KML – Keyhole Markup Language
LOD - Linked Open Data
LOGD – Linked Open Government Data
LR - Language Resources
MDA - Model Driven Architecture
MMS  Model Management System assistive system
MT - Machine Translation
MVC  Model View Controller
MVP  Model View Presenter
NGO  Non-Governmental Organisation
NLP – Natural Language Processing
OGC - Open Geospatial Consortium
OSM - OpenStreetMap
PM   Person Month
PPP - Public-Private-Partnership
QoS  - Quality of Service
RDF  Resource Description Framework
REST Representational State Transfer
RL   - Reference Laboratory
ROI  Return on Investment
SaaS Software as a Service
SDI – Spatial Data Infrastructure
SEIS – Shared Environmental Information System
SGML Standard Generalized Markup Language
SLA – Service Level Agreement
SME  Small to Medium Enterprise
SmOD SmartOpenData
SMT - Statistical MT
SOA – Service Oriented Architecture
SOAP Simple Object Access Protocol
SSRI - Social Spaces for Research and Innovation
SVG  Scalable Vector Graphics
UI   User Interface

275 Defined by the Commission at http://ec.europa.eu/enterprise/policies/sme/facts-figures-analysis/sme-definition/
UML – Unified Modelling Language
URI – Uniform Resource Identifier
URL – Uniform Resource Locator
URM – Uniform Resource Management
W3C – World Wide Web Consortium
WCS – Web Coverage Map
WFS – Web Feature Map
WMC – Web Map Context
WMS – Web Service Map
WPS – Web Processing Services
WWW - World Wide Web
XHTML – Extensible HyperText Markup Language
XML – eXtensible Markup Language