Linked Open Data for Environment Protection in Smart Regions
– The SmartOpenData Project Approach –

[1]Phil Archer, [2]Karel Charvat, [3]Mariano Navarro De La Cruz,
[4]Carlos Ángel Iglesias, [5]John O'Flaherty, [4]Tomás Robles,
[7]Dumitru Roman*

[1] ERCIM, France, [2] HSRS, Czech Republic, [3] TRAGSA, Spain,
[4] Universidad Politecnica de Madrid, Spain, [5] MAC, Ireland,
[7] SINTEF, Oslo, Norway
Dumitru.Roman@sintef.no

Abstract. Many different open information sources currently exist for protecting the environment in Europe, mainly focused on Natura 2000 network, and areas where environmental protection and activities like tourism need to be balanced. Managing these data and integrating them for supporting decision makers and for novel uses is a challenging task. The SmartOpenData project (2013-1015) 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 use. SmartOpenData will be based on previous experiences of Habitats project, which defined models and tools for managing spatial data in environmental protection areas. This paper provides an introduction to the SmartOpenData with a specific focus on the motivation, goals, and technical focus of the project, and outlines the architecture of the approach taken by SmartOpenData.

1 SmartOpenData Overview

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[1], and areas where environmental protection and activities like agriculture, forestry or tourism need to be balanced with the Habitats Directive[2] and the European Charter for Sustainable Tourism in Protected Areas[3]. Better understanding and managing these data not

* Authors listed in alphabetical order. Contact author: Dumitru Roman (dumitru.roman@sintef.no).

1 http://www.natura.org/

2 http://ec.europa.eu/environment/nature/legislation/habitatsdirective/index_en.htm

3 http://www.european-charter.org/home/

only can provide economic value of 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[4] project has set its goals to:

- Create a sustainable Linked Open Data[5] infrastructure in order to promote environmental protection data sharing among public bodies in the European Union;
- 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;
- 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;
- 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.

2 Building upon Previous Results: The HABITAS Project

The Habitats project[6] was built as an environment that enables to share and combine data from various sources. The project results were validated through the Habitats Reference Laboratory and pilot applications. On the basis of different pilots, Habitats defined and tested harmonisation rules for spatial environmental data and designed the concept of Reference Laboratory as a tool for testing the interoperability and supporting unification of outputs across different pilots.

The challenges faced by Habitats were mainly due to data availability, integration and usage ability for decision-making and, in particular, in terms of its focus on Metatada, Data Specifications, Network Services, Data and Service Sharing and Monitoring and Reporting. Habitats is to support the EU INSPIRE Directive.

The specific usage scenarios, including the state of the art baseline and user requirements coming from them represent the key input for the planned data and meta-data modeling activities and the SDI services that were developed in the Habitats project. Generally speaking, a positive correlation in all the pilots was detected, between service development and user satisfaction, while on the other hand, it cannot be taken for granted that the new services provided are also INSPIRE compliant. This can be due to several reasons, two of which seem more prominent than others:

- On the “supply side”, the cost of increasing the compliance, in terms of time, resources, etc., from the perspective of the SDI “owner”;
- On the “demand side”, lack of interest or simply ignorance of the advantages of compliance, from the perspective of the end users.

---

4 SmartOpenData: "Linked Open Data for environment protection in Smart Regions” under the call FP7-ENV-2013-two-stage concerning the Seventh Framework Programme (FP7) (2013-2015).

5 http://linkeddata.org/

6 http://www.inspiredhabitats.eu/

3 The SmartOpenData Approach

Linked Open Data is emerging as a source of unprecedented visibility for environmental data that 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, GMES, INSPIRE and voluntary data (OpenStreetMap, GEP-WIKI, etc.), and how it can impact the economic and sustainability progress in European Environment research and Biodiversity Protection are open questions that need to be addressed in order to benefit from an improved understanding and management of environmental data. The SmartOpenData project (2013-2015) will address these questions by defining mechanisms for acquiring, adapting and using Open Data with a particular focus on biodiversity and environment protection in rural and European protected areas and its National Parks.

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. SmartOpenData goal is making INSPIRE/GMES/GEOSS infrastructure better available for citizens, as well as for public and private organization. On one hand, Europe and EU invest hundreds of millions of Euros in building the INSPIRE infrastructure. On the other hand, public and private organizations, as well as citizens use for their applications 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:

- Harmonise geospatial metadata (ISO19115/19119 based) with principles of Semantic Web;
- Provide spatial data fusion introducing principles of Open Linked Data;
- Improve spatial data visualisation of Geospatial Open Linked Data;
- 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 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 (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 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 relation 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 by OGC's GeoSPARQL standard[7]. 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[8] but there remains some debate as to whether it is the best approach. The NeoGeo vocabulary[9] 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[10]. A related, but separate, Community Group is also considering better interplay between Web and geospatial technologies[11]. 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. 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.

Another 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 growing of SDI. The implementation of RDF should help ease the translation of geographic names or keywords from vocabularies like GEMET or AgroVoc.

The research focus for SmartOpenData 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 tool[12] 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 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 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 site[13] shows 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 SmartOpenData.

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,

---

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

[8] http://geosparql.org, https://twitter.com/marin_dim/status/271573164268609536, http://www.strabon.di.uoa.gr

[9] http://geovocab.org/doc/neogeo

[10] http://www.w3.org/community/locadd

[11] http://www.w3.org/community/geosemweb

[12] http://openrefine.org (formerly Google Refine)

[13] http://datahub.io/dataset/geospecies/resource/47e71c4c-9565-4185-b8c0-bdef6449278e

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, we will explore resource-limited adaptation to those domains in the context of SmartOpenData.

Another area of 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[14] 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[15] and SPARQL Update.

4 The SmartOpenData Architecture

The SmartOpenData infrastructure is depicted in the following figure where three main elements can be identified.

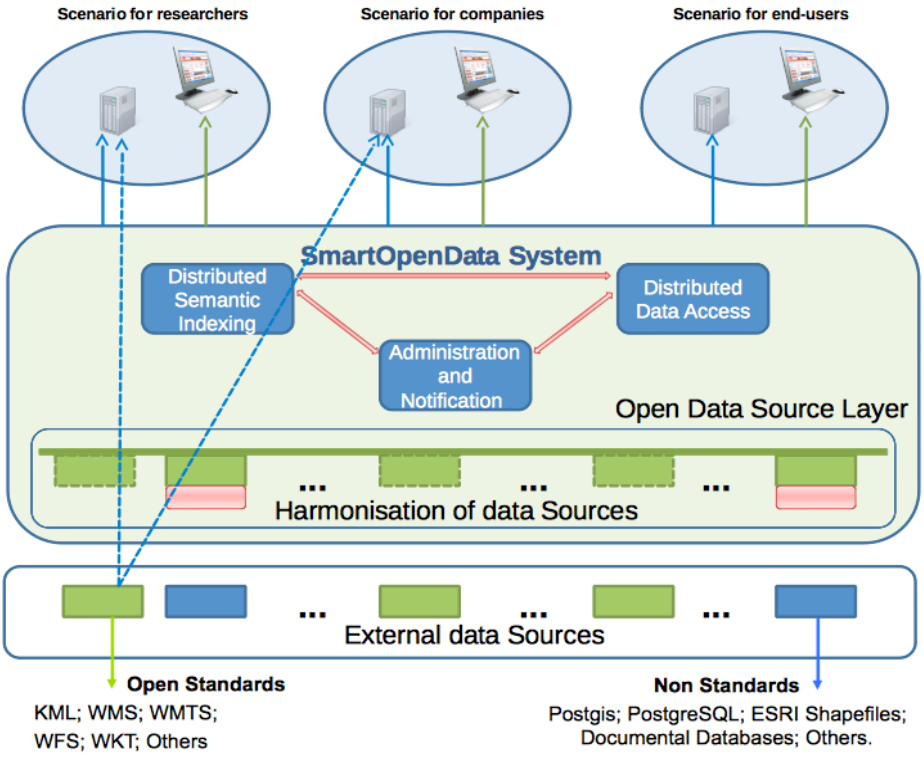

Figure 1. SmartOpenData Infrastructure

In the lower level the external data sources are depicted. Data sources can be grouped in two different sets. The first one is composed by data sources that fulfil some of the standards supported by SmartOpenData (green boxes). The second group is composed by data sources that does not fulfil those standards (blue boxes). In the upper layer, three different scenarios

[14] http://www.w3.org/standards/techs/powder#w3c_all

[15] http://www.w3.org/standards/techs/provenance#w3c_all

have being identified: scenario for researches, scenario for companies and 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 is placed providing key functionalities. 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. If an external data source does not provide the information according to the required standard, and adaptation is required, which is depicted in the figure as an extra box, which provides such adaptation specifically tuned for each external data source. 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:

- Distributed semantic indexing, which provides a service for searching and locating data based on semantic information collected from all the available Data Sources;
- 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;
- 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 accesses 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 thought the distributed semantic indexing functionality of the SmartOpenData System if they are provided using one standard as shown on the picture.

5 Outlook

SmartOpenData 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 environment protection and research in rural and European protected areas and its National Parks. SmartOpendata is going to evaluate infrastructure and tools by the development and deployment of five advanced demonstrators focused on agroforestry management, environmental research water monitoring respectively, forest sustainability and environmental data reuse. 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 the biodiversity status, maintenance and protection, including achievement of "the INSPIRE and Open Data Ready" status for their digital (not only spatial) content. 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 will be encouraged by the promotion of royalty-free open standards and best practices generated, initiated or simply highlighted by SmartOpenData. The project will also contribute and where possible benefit from ongoing and upcoming related initiatives like Open government partnership[16], INSPIRE

[16] http://www.opengovpartnership.org/

maintenance and implementation framework[17], European Union Location Framework[18], or Interoperability Solutions for European Public Administrations[19] (e.g. Working Group on Spatial Information and Services).

Acknowledgment. The SmartOpenData project is funded by the European Commission and is a collaboration between 16 European organizations (Empresa de Transformación Agraria SA, Universidad Politécnica de Madrid, The National Microelectronics Applications Centre LTD, Sindice LTD, Mid-West Regional Authority, Environment Protection Regional Agency, Fondazione Bruno Kessler, Spazio Dati, Help Service-Remote Sensing SRO, Forest Management Institute, Czech Centre for Science and Society, Stiftelsen SINTEF, Latvijas Universitates Matematikas Un Informatikas Instituts, Direção Geral do Território, Slovak Environmental Agency, European Research Consortium for Informatics and Mathematics).

[17] http://inspire.jrc.ec.europa.eu/index.cfm/pageid/5160

[18] http://ec.europa.eu/isa/actions/documents/isa-2.13_eulf-strategic-vision-lite-v0-3_final_en.pdf

[19] http://ec.europa.eu/isa/