(Draft) Actas de las II Jornadas Ibéricas de Infraestructuras de Datos Espaciales (JIIDE'2011), Barcelona, 9-11 de noviembre de 2011. 2011

The Delft Report: Linked Data and the challenges for geographic information standardization

Francisco J. Lopez-Pellicer[1], Luis M. Vilches-Blázquez[2], F. Javier Zarazaga-Soria[1], Pedro R. Muro-Medrano[1], Oscar Corcho

[1]Sistemas de Información Avanzados (IAAA),
Dpto. Informática e Ingeniería de Sistemas
Universidad de Zaragoza

{fjlopez,javy,prmuro}@unizar.es

[2]Ontology Engineering Group (OEG), Dpto. Inteligencia Artificial,
Facultad de Informática, Universidad Politécnica de Madrid

{lmvilches, ocorcho}@fi.upm.es

Abstract

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). The combination of Linked Data and SDIs, its benefits and challenges are collected in the report on Linked Data presented at 32nd ISO/TC 211 plenary in Delft. This paper presents a brief summary of the mentioned report, where we focus on the main recommendations in the context and evaluate their potential impact in SDIs.

Keywords: Semantic Web, Linked Data, ISO/TC 211, Spatial Data Infrastructures.

1. Introduction

The resolution 532 of the 31st ISO/TC 211 plenary in Canberra mandated the creation of an ad-hoc group of experts to investigate whether to address the Web of

Data and related issues, in particular Linked Data, in the geographic information standardization, and, if that is the case, to make recommendations for action. The group included experts from Australia, Canada, Germany, Italy, Finland, France, Korea, Spain, United Kingdom and USA. The first report of this group was presented to the 32nd ISO/TC 211 plenary held in Delft, Netherlands, on May 2011, hereafter the Delft Report [1]. The most significant recommendation of that report is to proceed with the use of Linked Data in geographic information standardization.

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), such as cross domain data sharing or spatial data discovery. The Delft Report's recommendations are the first step in that direction. This paper presents those recommendations in context, and evaluates their potential impact in SDIs.

2. The Web of Data

The Web is a system of interlinked hyperlinked documents build on top of the Internet. The Web of Data is a natural extension of the Web based on the W3C standards and best practices related to Semantic Web where structured data are interlinked in the same way as documents are linked on the Web.

2.1. The Semantic Web

The Web is based in three essential technologies: a system of globally unique identifiers known as URIs, the markup language for documents HTML (Hypertext Markup Language) and the networking protocol HTTP (Hypertext Transfer Protocol). The origin of the Web backs to the work of Berners-Lee and Cailliau in the 1990 at CERN. Eleven years later, Berners-Lee and colleagues proposed an extension oriented towards machine processable data: the *Semantic Web* [2]. The Semantic Web was initially focused on developing a single data interchange model named Resource Description Framework (RDF) with multiple serializations (RDF/XML, N3, Turtle, RDFa, etc.), a RDF query language (SPARQL) and notations for defining taxonomies (RDFS), ontologies (OWL) and rules (RIF) (see Table 1 for a complete overview).

The RDF data model is based on the idea of making statements about resources. These statements take the form of subject-predicate-object expression or triples. The subject is the name of a thing about which something is asserted. The predicate is a named property that expresses the relationship between the subject and the object. The object is a value or a resource associated to the subject by the predicate. Taxonomies, ontologies and rules can define classes and properties that help the understanding of statements. In turn, taxonomies, ontologies and rules can be related each other through statements that lead to discover additional knowledge. The statements can be stored in a repository of RDF data. These repositories are known as *triplestores* and can be built on top of a database or a file system. The content of those repositories can be exposed with a Web endpoint that implements the SPARQL Protocol, which enables the use of the SPARQL query language for querying the remote *triplestore*.

RDF	*Resource Description Framework*: a model based on subject-predicate-object statements for data interchange.
OWL	*Web Ontology Language*: a family of languages for the encoding of ontologies (i.e. formal representation of knowledge).
SPARQL	*SPARQL Protocol And RDF Query Language*: a Web protocol for querying remote RDF stores and a RDF query language.
RDFa	*RDF Annotations*: a collection of attributes and processing rules for extending markup languages to embed RDF data.
SKOS	*Simple Knowledge Organization System*: a common data model for sharing and linking knowledge organization systems.
RDFS	*RDF Schema*: a language for the encoding of taxonomies (i.e. hierarchical classification scheme).
GRDDL	*Gleaning Resource Descriptions from Dialects of Languages*: a markup for declaring that an XML document embeds RDF data and for linking to algorithms for extracting this data.
POWDER	*Protocol for Web Description Resources*: a mechanism to discovery and describe Web resources.
RIF	*Rule Interchange Format*: a family of languages for the encoding of rules.
SAWSDL	*Semantic Annotations for WSDL*: a mechanism to enable semantic annotation of Web services descriptions in WSDL.

Table 1. The W3C Semantic Web standards

2.2. Linked Data

In 2006, Berners-Lee proposed the idea of *Linked Data* [3]: a Web of Data that has almost the same properties of the Web where a data interchange model (RDF) replaces HTML. This means that data (its Web representation) can contains links to data located elsewhere on the Web. Moreover, he proposes to adhere to the REST [4] architectural style that governs the Web: request and responses are built around the transfer of representation of resources using the existing HTTP capabilities (e.g. authentication, caching, content type negotiation) and verbs (GET, POST, PUT, DELETE). This approach differs from other approaches, such as SOAP, where developers are encouraged to design arbitrary features disregarding those offered by HTTP. As a result, Linked Data enables large-scale integration of data over the Web because is aligned with the Web architecture. Additionally, as Linked Data is based on RDF, Semantic Web languages with strong logical foundations, such as OWL, allow modelling what data mean.

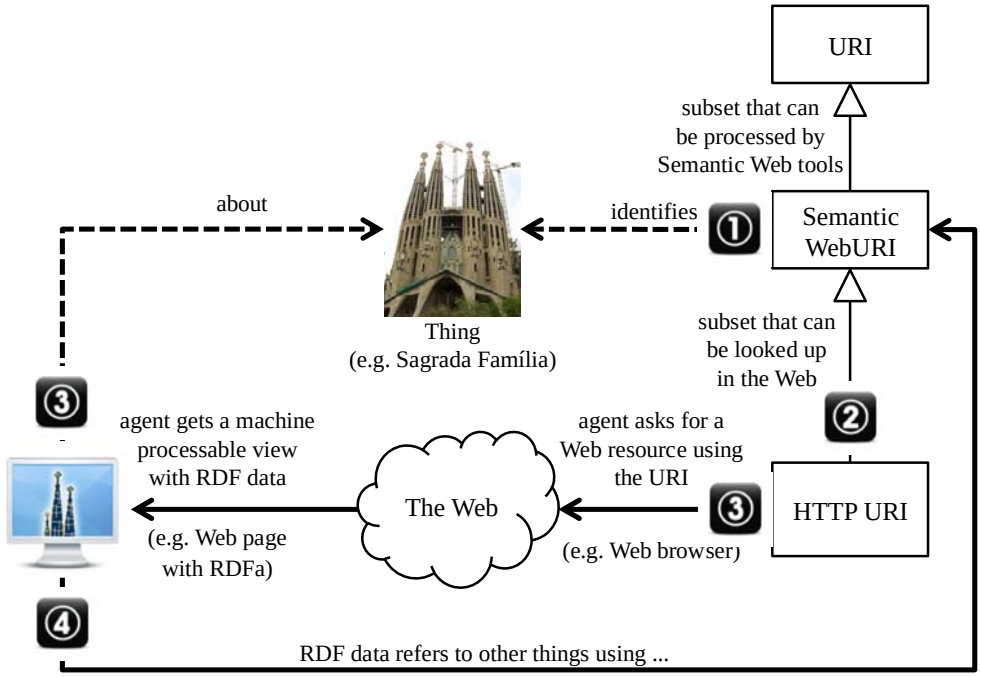

Figure 2. Linked Data rules are expectations of behaviour of agents and data publishers required to make data interconnected on the Web.

Linked Data addresses identifier syntax, data access and data model, and proposes a data integration approach. Linked Data is often summarized in four principles (see Figure 2):

1. **Use URIs as names for things**. Not just any kind of URI. Only those URIs compatible with specifications of the Semantic Web can be used.
2. **Use HTTP URIs** so people can look up those names on the Web. That is, if a publisher uses a different schema[1], it tells people those names cannot be looked up.
3. **Use dereferenceable HTTP URIs**. When someone looks up a URI, provide useful information using the Semantic Web standards. That is, return a Web resource that is either a RDF data serialization (e.g. a static file, a SPARQL response), a Web resource with embedded RDF data (e.g. XHTML+RDFa), or a Web resource with a standardized transformation to RDF data (e.g. XML+GRDDL).
4. **Include links to other URIs**, so they can discover more things. Most of the links should be RDF links, that is, *URIs that provide useful information*. Links that provide alternate representations are also advised. For example, a RDF document that describes a geographic resource can provide a link to an ISO metadata document that describes the same resource. Other links can be deduced with the help of ontologies and rules.

The links introduced in the latter point are known as *RDF links*. They enable to navigate from a data item (e.g. a metadata record) within one data source (e.g. a catalogue) to related data items (e.g. a term) within other sources (e.g. a thesaurus) using a Web agent (e.g. a Web browser in the simplest scenario). RDF links can also be followed by the crawlers of general purpose and Semantic Web search engines. Linked Data is indexable. Therefore, Linked Data datasets are searcheable by means of simple (e.g. Google, Bing) or semantic-aware (e.g. Sindice) search engines. In addition, as RDF is serialized in a structured format (e.g. XML, JSON) it can be consumed by applications.

[1] The OGC uses the *urn* schema for naming persistent resources. See http://www.opengeospatial.org/ogcna

3. Linked Data for the ISO/TC 211

3.1. The Semantic Web at the ISO/TC 211 and liaised organizations

The ISO/TC 211 tasked a group in 2006 to investigate how Semantic Web approaches can benefit in the development of interoperable geospatial information. That group recommended in 2009 the revision of the TC/211 reference model and the development of rules for the development of notifies based on Semantic Web languages, among other recommendations. New projects have been initiated as consequence of those recommendations, such as the Project 19150-2 for developing ontologies with OWL. In relation to Linked Data, the 31st ISO plenary (2010) mandated the creation of an ad-hoc group to investigate that notion and its consequences for geographic information standardization. The recommendations of that group are discussed in this paper.

The most notable ISO/TC 211 liaised organization is the international industrial consortium OGC. The OGC has developed several Web standards that have been standardized by ISO/TC 211. The OGC Working Groups Geosemantics (2007) and GeoSPARQL Standard (2010) concentrate the work related to the Semantic Web. The mission of the Geosemantics group is to establish a semantic framework for representing and mediating the geospatial knowledge. The goal of the GeoSPARQL group is to define a vocabulary for representing geospatial data in RDF, and a function library for SPARQL that enables spatial queries.

It is remarkable that ISO/TC 211 and W3C, which maintains the Semantic Web standards, have no liaison agreement although W3C is liaised with other ISO technical committees, in particular ISO/IEC JTC 1 – Information Technology Standards and ISO TC 68 – Financial Services. Identical situation happens between OGC and W3C, which are only liaised for the standardization of SVG.

3.2. Linked Data in geographic information standards: benefits and potential pitfalls

Linked Data and geographic information standards share the same basic principle: to make data easier to discover and use. The Delft Report examines several potential benefits of Linked Data that have been acknowledged by other authors [5-8]:

- **Domain independence**. ISO/TC 211 models are built on domain pillars, such as the General Feature Model [9]. Liked Data is domain agnostic and provides a common approach for the description of things. Moreover, Linked Data allows applying different perspectives and uses on the same data.
- **Data reuse**. ISO/TC 211 standards do not address the integration of different kinds of data. Linked Data integrates data in a uniform manner delegating technical issues to the Internet and the HTTP protocol. Hence, spatial data and metadata can be discovered and reused by users of other domains.
- **Commons tools**. ISO/TC 211 encodings (e.g. ISO 19139) require domain tools for processing and understanding. Linked Data uses common tools, where the treatment of domain data is delegated to specialized modules.

However, the report fails to examine potential pitfalls. For example, the report does not take into account that the hardware and software infrastructure required to process geo Linked Data is not mature yet.

3.3. Areas of concern

The Delft Report suggests some key areas of concern for ISO/TC 211:

- **Standards management**. All or part of standards and specification content should be available as Linked Data. Conformance classes, tests, code lists, etc. should have their own dereferenceable HTTP URI.
- **Geographic data and metadata**. Structured geographic data (e.g. features, sensors, coverages, metadata) should be identified with HTTP URIs and be available as RDF documents containing RDF links to related data.
- **XML encoding reuse.** RDF provides for XML content as possible literal value. Hence, Linked Data community can reuse XML documents defined by XML schemas for encoding geographic information as XML literals. This feature opens new and unsuspected possibilities.
- **Liaisons.** As it was said above, W3C and OGC are developing industrial standards related to Linked Data technologies and geographic information. Standardization work should be liaised with W3C and OGC.
- **Quality**. RDF links raises concerns about data quality. The RDF model allows merging easily datasets of different provenance and characteristics. If this process is done without care, the quality may suffer.

3.4. Challenges

There are several challenges related to the adoption of Linked Data. The most relevant are described below. Some of them have been already tackled by the Project team 19150 Geographic Information - Ontology.

- **Identification**. ISO/TC 211 standards should be revised to consider the use of HTTP URIs to identify things, such as concepts, features or metadata records. Moreover, as those URIs require stability and durability, the governance of URIs should be also considered. Well-defined rules shall govern the creation of new URIs.
- **Dereferencing identifications**. HTTP URIs shall be dereferenceable at least to a RDF document, and it should be clear what an agent should expect to retrieve. For example, ISO/TC 211 shall maintain a Linked Data front end that dereferences identifications (e.g. GM_Point URI, MD_Metadata URI, FC_FeatureType) to the corresponding specification containing their definitions. Content negotiation may be used for selecting the format (e.g. RDF document, HTML page, PDF document).
- **Model transformation**. ISO/TC 211 shall define vocabularies that support the description in RDF of geographic data and metadata, and a methodology for the transformation of existing models to RDF data, and vice versa. Ontologies (OWL) and rules (RIF) provide the necessary constructs for the transformations.
- **Data types**. Data types provided by RDF or OWL are used for the representation of primitive values, such as integers, floats and strings. The support of user defined data types, such as geometries, is an open issue with different approaches. ISO/TC 211 shall define domain data types, such as point, polygon or geometry, and schemes of integration with the formal semantics of different semantic languages.

4. The Delft Report's recommendations

The Delft Report on Linked Data was presented to the 32nd ISO/TC 211 plenary held in Delft, Netherlands, on May 2011. The report addressed many of the issues discussed in previous sections and contained a number of recommendations. The recommendations to the ISO/TC 211 can be summarized in the following list:

1. Proceed to use of Linked Data in the standardization of geographic information.

2. Develop RDF vocabularies for different aspects of geographic information (e.g. terminology register, harmonized model).
3. Set up a well-defined Linked Data frontend for ISO/TC 211 concepts.
4. Review XML encodings to support their use as XML literals in RDF datasets.
5. Establish guidelines for the support of the use Linked Data liaised with W3C and OGC.

5. Linked Spatial Data Infrastructures

As mentioned before, Linked Data and SDIs share the same basic principle: (1) to provide a common approach for all kinds of data, and (2) to make data easier to discover, connect and use. Linked Data adds, for example, the possibility to cross-domain barriers enabling the linkage with other domain data (e.g. health, transport). Hence, there is an important cost saving recognition by publishing once and reusing multiple times.

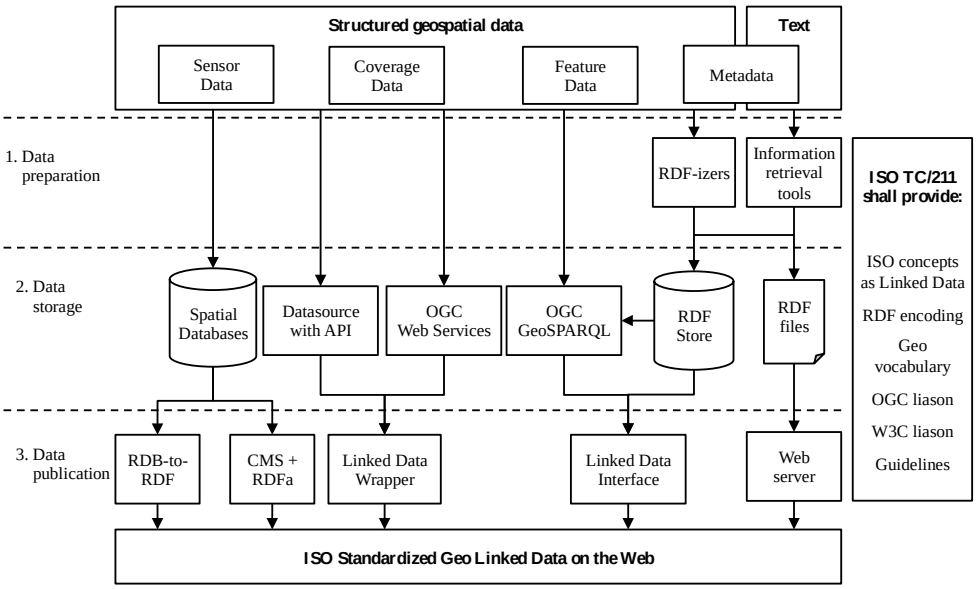

Figure 2. Technologies and standards involved in a Linked Spatial Data.

Some authors consider that Linked Data can leverage discovery and use goals of SDIs [8, 10]. The more important aspect is that Linked Data allows SDI to integrate geographic information by lowering the barrier to reuse geographic and location data by third parties. A good example is the Linking Open Data (LOD) project[2]. This project is an initiative originated in the W3C Semantic Web Education and Outreach interest group in 2007. The goal of the project is the creation of a data commons by publishing various open datasets based on the principles of Linked Data. As of September 2011, the Linking Open Data (LOD) project includes 295 data sets and consists of over 31 billion RDF triples, which are interlinked by around 504 million RDF links. The most notable of the LOD initiative is that it has attracted significant spatial data producers (e.g. Ordnance Survey - UK, National Geographic Institute – Spain, NASA). These producers have published their datasets as Linked Data enabling third parties to make RDF links to the contents of these datasets to assert the location of resources. In addition, large crowd-sourced geospatial databases (e.g. OpenStreetMap, GeoNames) have been published as Linked Data with links to the LOD datasets.

Figure 2 presents different strategies that can be considered for publishing data as Linked Data in a SDI. Structured data (e.g. sensor data, coverage data, feature data, metadata) can be exposed as RDF using one or several strategies depending on their amount and nature. For example, if the data is available through an API or exposed with and OGC Web service, it is possible to use a wrapper to expose the data as Linked Data. In some scenarios, where data is embedded in text (e.g. legacy metadata), entities can be retrieved with information retrieval tools and then structured and merged with existing RDF data. The ISO/TC 211 shall provide vocabularies, rules and guidelines for ensuring interoperability and consistent semantics across different SDIs.

6. Conclusions

The 32nd ISO/TC 211 plenary has requested the ad hoc group continue their work and report to the next plenary [11]. The work programme is to liaise with the ongoing work of OGC, in particular for the definition of shared vocabularies and the URI governance, to review XML encodings to determine potential conflicts with their use as XML literals in RDF, and to look about guidelines on the publication of standard geospatial information as Linked Data.

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

Linked Data will not replace existing service oriented solutions based on existing ISO/TC 211 standards. In fact, it is rather a disruptive technology for data reuse. That is, Linked Data would complement existing service oriented approaches by providing a Web-oriented infrastructure for broadcasting geospatial data and theirs semantics out of the geospatial community.

Acknowledgements. This work has been partially supported by the Spanish Government (projects "España Virtual" ref. CENIT 2008-1030 and TIN2009-10971), the Centro Nacional de Información Geográfica (CNIG) and CDTI (programme Ingenio 2010), the National Geographic Institute (IGN), the Spanish thematic network of Linked Data (TIN2010-10811-E), and GeoSpatiumLab S.L.

Bibliography

[1] Brodeur, J. (Chair): Linked Data – Report to 32nd ISO/TC 211 plenary, Delft, The Netherlands, May 26, 2011. Technical Report. ISO/TC 211 Ad hoc group on Linked Data. Delft, The Netherlands (2011)

[2] Berners-Lee, T., Hendler, J., and Lassila, O.: The Semantic Web. Scientific American 284(5): 34-43 (2001)

[3] Berners-Lee, T.: Linked Data - Design Issues, (2007). http://www.w3.org/DesignIssues/LinkedData.html

[4] Fielding, R. and Taylor, R.: Principled design of the modern Web architecture. *ACM Trans. Internet Technol.* 2(2): 115-150 (2002)

[5] Smits, P.: Linked Open Data and Spatial Data Infrastructures. EcoInformatics Linked Open Data workshop (2010). http://tinyurl.com/28poe5d

[6] Murray, K., Sheridan, J., Hart, G., Tennison, J., Goodwin, J., Davidson, P., et al.: Linked Data and INSPIRE – extending the benefits of data . INSPIRE Conference 2010 - INSPIRE as a Framework for cooperation, Krakow, Poland (2010)

[7] Lopez-Pellicer, F.J, Silva, M.J., Chaves, M., Zarazaga-Soria, F.J., and Muro-Medrano, P.R.: Geo Linked Data. Database and Expert Systems Applications, 21st International Conference, DEXA 2010, Bilbao, Spain, August/September 2010, Proceedings, Part I. Lecture Notes in Computer Science, 6261: 495-502 (2010)

[8] Hjelmager, J., Moellering, H., Delgado, T., Cooper, A., Rajabifard, A., Rapant, P., et al.: An initial formal model for a Spatial Data Infrastructure. International Journal of Geographical Information Science. 22(11/12), 1295-1309 (2008)

[9] Kresse, W., Fadaie, K.: ISO Standards for Geographic Information. Springer, Berlin (2004)

[10] Cox, S., Schade, S., and Portele, C.: Linked Data in SDI. INSPIRE Conference 2010 - INSPIRE as a Framework for cooperation, Krakow, Poland (2010)

[11] ISO/TC 211 plenary: Resolutions from the 32nd ISO/TC 211 plenary meeting in Delft, The Netherlands, 2011-05-26/27, Resolutions, ISO/TC 211, Delft, The Netherlands (2011) http://www.isotc211.org/opendoc/211n3148/