Semantic Annotation 2

MARCH 2017

DELIVERABLE

Project Acronym: **SDI4Apps**
Grant Agreement number: **621129**
Project Full Title: **Uptake of Open Geographic Information Through Innovative Services Based on Linked Data**

D5.2.2 SEMANTIC ANNOTATION 2

Revision no. 03

Authors: **Otakar Čerba** (University of West Bohemia)

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	X
C	Confidential, only for members of the consortium and the Commission Services	

REVISION HISTORY

Revision	Date	Author	Organisation	Description
01	07/03/2016	Otakar Čerba	UWB	Initial draft
02	29/03/2016	Martin Tuchyna, Karel Charvát	SAZP/CCSS	Internal review
03	30/03/2016	Otakar Čerba	UWB	Final version

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Disclaimer:

Views expressed in this document are those of the individuals, partners or the consortium and do not represent the opinion of the Community.

TABLE OF CONTENTS

LIST OF FIGURES

EXECUTIVE SUMMARY

This document describes a semantic annotation realized in the SDI4Apps project. At first the general steps how to add semantic annotation to spatial data is described. Because the re-using of existing solutions (data models, properties, vocabularies etc.) is the crucial part of semantic annotation issues, the next section presents selected vocabularies and data models. Similarly, to the part Examples in SDI4Apps there are described components applied in the Pilot II (Open Smart Tourist Data), which consist in Smart Point of Interest (SPOI) dataset. Therefore, the section Examples in SDI4Apps deals with SPOI data model and semantic annotation. The last part (Conclusions) provides lessons learnt during semantic annotation activities in the Pilot II which can be implemented in other pilots in SDI4Apps project and other issues related to semantic and linked data.

1 GENERAL DESCRIPTION OF SEMANTIC ANNOTATION ISSUES

Semantic annotation issues include:

- Selection of a suitable processes for semantic data model design – this step compares and evaluates various approaches describing a semantic data structure. This covers ontologies (not only own ontological models, but also possibilities how to adapt or extent existing ontologies), database models (including noSQL databases) and mark-up languages (including different schema languages and graph data structures). There is not only discussed the most convenient approach, but also data formats (such as OWL, RDF, RELAX NG, XML Schema or newly developed format based on XML). The final selection (see following two parts of this document) has to depend not only on the user requirements but also on the opportunities for the interconnection and re-use of existing solutions.
- Design of the semantic data model consisting of two sub-steps:
    - Definition of the minimal structure (mandatory elements and attributes) of data model – this sub-step can be illustrated by mandatory properties of SPOI objects such as geometry, label, classification, metadata or topological relation connected POI with a relevant country. Similarly, to the next sub-step there are re-used existing format and vocabularies as much as possible.
    - Enrichment/extension of the minimal version of the model – incorporation of selected existing structures (as parts of model or as links). This sub-step can lead to a development of other specific models for different cases (pilot applications or existing products). This step is conducted in parallel to the previous one. It consists of the testing of structures that are based on the exploitation of the selected open data and its transformation to the last version of semantic data model(s).
- Publication of the finalised first version of the semantic data model (or models) – it includes the model(s) in the major format (selected in the step 1), its/their detailed description (metadata) and also transformations of the model(s) to other usable formats (if they will be required).

2 EXISTING POSSIBILITIES AND SOLUTIONS

The following vocabularies (in alphabetical order) based on standardized formats or query languages have been used in the SPOI solution, which is introduced in the following part of this document:

- Dublin Core (http://dublincore.org/) represents the basic set of metadata properties. There are two levels of implementation of this set of vocabulary terms in the SPOI: the original Dublin Core Metadata Element Set (with prefix "dc"), which contains 15 original metadata terms standardized as ISO 15836, and The Dublin Core Metadata Initiative (DCMI) Metadata Terms (with prefix "dcterms"). This set is an up-to-date specification of all metadata terms maintained by DCMI and includes fifteen terms of Dublin Core Metadata Element Set as well as many others. Dublin Core elements are in the SPOI used for description of feature metadata such as data provider, original data resource, date of integration to SPOI or license.

  Example: <dc:source rdf:resource="https://www.openstreetmap.org"/>

- Friend of a Friend (FOAF; http://www.foaf-project.org/) is a machine-readable ontology of people-related terms describing persons, their activities and their relations to other people and objects. The specification helps to describe contact information of SPOI such as email address, telephone number or web page.

  Example: <foaf:homepage>www.refugisorteny.com</foaf:homepage>

- GeoSPARQL (http://www.opengeospatial.org/standards/geosparql) is a standard from the Open Geospatial Consortium (OGC) for representing and querying geospatial linked data on the Semantic Web. It defines a vocabulary for representing geospatial data in RDF as well as an extension to the SPARQL query language for processing geospatial data. There are two different ways for GeoSPARQL implementation in the SPOI: coding of coordinates in WGS 84 system and expressing of topological relation between POI and relevant country.

  Example: <geos:sfWithin rdf:resource="http://www.geonames.org/3041565"/>

- ISA Programme Location Core Vocabulary (LOCN; https://www.w3.org/ns/locn) is a simplified, reusable and extensible data model that captures the fundamental characteristics of a location, represented as an address, a geographic name or a geometry. The vocabulary is also designed to aid the publication of data that is interoperable with EU INSPIRE Directive. In the SPOI LOCN enables to describe not only full address of POI, but also particular components of addresses such as street, post name or post code.

  Example: <locn:postCode>2180</locn:postCode>

- Resource Description Framework Schema (RDFS; https://www.w3.org/TR/rdf-schema/) represents semantic extension of the basic RDF vocabulary. It is a set of classes, associated properties and utility properties built on the vocabulary of RDF providing basic elements for the description of ontologies intended to structure RDF resources. RDFS was the first language enabling the development of RDF vocabularies by defining the basic blocks such as class, property, domain or range. SPOI uses classes such as label or comment for human-readable name or description of the subject as well as utility property "seeAlso" for linking to external resources that might provide additional information about the subject resource (e.g. Wikipedia pages).

  Example: <rdfs:label xml:lang="fr">Kaboul</rdfs:label>

- Simple Knowledge Organization System (SKOS; https://www.w3.org/2004/02/skos/) is designed for representation of thesauri, classification schemes, taxonomies, subject-heading systems, or any other type of structured controlled vocabulary. Similarly to OWL in the SPOI only the SKOS property skos:exactMatch is used for interconnecting to equivalent objects in different data sets such as DBpedia, Wikidata, GeoNames.org or LinkedGeoData.

  Example: <skos:exactMatch rdf:resource="http://linkedgeodata.org/triplify/node26864258"/>

- Web Ontology Language (OWL; https://www.w3.org/TR/owl2-overview/) is designed for describing and sharing of ontological systems on the World Wide Web. It is based on combination of RDF

structures and Description Logic Rules. Similarly to SKOS in the SPOI only the OWL property owl:sameAs is used for interconnecting to equivalent objects in different data sets such as DBpedia, GeoNames.org, Wikidata or LinkedGeoData.

Example: <owl:sameAs rdf:resource="http://linkedgeodata.org/triplify/node26864258"/>

The following vocabularies based on standardized formats have been used for metadata files of the SPOI solution:

- Description of a Project (DOAP; https://github.com/ewilderj/doap/wiki) is an XML/RDF vocabulary to describe semantic information associated especially with open source software projects. Similarly, to VOID the DOAP terms are in the SPOI used in separate metadata file to provide information about the dataset such as name, acronym, description or homepage.

  Example: <homepage rdf:resource="http://sdi4apps.eu/spoi/"/>

- Vocabulary of Interlinked Datasets (VOID; http://www.w3.org/TR/void/) is an RDF Schema focused on metadata about RDF datasets with the aim of interconnecting the publishers to users of RDF data. VOID terms are in the SPOI used in separate metadata file to point out existence of a SPARQL endpoint and its URL or show the total number of triples contained in dataset.

  Example: void:sparqlEndpoint <http://data.plan4all.eu/sparql>;

There are also other projects, vocabularies and related methodologies foreseen to be taken into the consideration by the SDI4Apps project, particularly those with close connection to the INSPIRE linked data related activities:

- ARE3NA Reusing INSPIRE (https://joinup.ec.europa.eu/asset/are3na-reuse/description) deals with reusing INSPIRE metadata in open data portals and creating new ways to make use of INSPIRE data models as Linked Data on the Semantic Web. For his purpose GeoDCAT-AP as an extension of DCAT-AP based on W3C's Data Catalogue vocabulary (DCAT) was developed. Its basic use case is to make spatial datasets, data series, and services searchable on general data portals and give owners of geospatial metadata the possibility to achieve more by providing an additional RDF syntax binding. In addition, area of work that ARE3NA is exploring is related not only to metadata, but also to representation of INSPIRE data as Linked Data. For this purpose, draft methodology to produce INSPIRE data in RDF was defined and vocabularies for selected INSPIRE spatial object types are being now developed. These vocabularies include namely for instance ontology contains classes and properties that have been derived from the INSPIRE "CadastralParcels" application schema.
- SmartOpenData (SmOD) INSPIRE Vocabularies (https://www.w3.org/2015/03/inspire/) is a set of very small vocabularies defining classes and properties that mirror those in INSPIRE. Its effort is not to recreate the full scope of INSPIRE data model in RDF, but to provide a Linked Data interpretation of INSPIRE and gain some benefits from that approach by re-using concepts wherever necessary or simplifying INSPIRE data model where possible. As an example, may be mentioned Land Cover Vocabulary, Administrative Units Vocabulary or Cadastral Parcels Vocabulary.

There are also other vocabularies based on standardized formats or query languages that can be used for semantic description of spatial data or as an alternative to vocabularies used in the SPOI solution.

- WGS84 Geo Positioning (GEO; https://www.w3.org/2003/01/geo/) provides namespace for describing points with latitude, longitude and altitude in the WGS84 geodetic reference datum.
- Time Ontology (TIME; http://www.w3.org/TR/owl-time/) defines temporal properties of resources such as topological temporal relations among instants and intervals, time position and time units or information about durations.
- The Geonames ontology (GN; http://www.geonames.org/ontology/documentation.html) contains elements of description such as name, coordinates or link to Wikipedia article for geographical features defined in the GeoNames.org data base.

- Ordnance Survey Ontologies (http://data.ordnancesurvey.co.uk/ontology). Ordnance Survey is Great Britain's national mapping agency, which provide up-to-date geographic data and publish also a number of its products as Linked Data. For this purpose, several ontologies were created. As an example, may be mentioned ontologies describing abstract geometries, basic spatial relations, the postcode geography in Great Britain or the administrative and voting area geography of Great Britain.
- The Data Cube Vocabulary (QB; http://www.w3.org/TR/vocab-data-cube/) allows multi-dimensional data, such as statistics, to be published in a web-friendly format (RDF) to enable it to be linked and combined with related information.
- NeoGeo Spatial Ontology (SPATIAL; http://geovocab.org/doc/neogeo/) is designed for describing topological relations between features.
- SPARQL 1.1 Service Description (SD; http://www.w3.org/TR/sparql11-service-description/) provides description by which a client or end user can discover information about the SPARQL service such as supported result formats or extension functions and details about the available dataset.
- Center of Excellence for Geospatial Information Science (CEGIS) Ontologies (https://cegis.usgs.gov/ontology.html) are ontologies especially for The National Map of U.S. Geological Survey (USGS) to express multiple semantic and spatial relations of topographic features. The vocabularies cover various themes such as terrain, water surface or ecological regimes.
- Creative Commons Rights Expression Language (CC, https://creativecommons.org/ns) describes copyright licenses in RDF.

3 EXAMPLES IN SDI4APPS

SPOI represents the biggest open dataset of POIs using the Linked data approach. It contains more than 27.5 million points over the whole world. Just because SPOI has a character of Linked data, the semantic annotation is the essential tools how to describe particular attributes of POI efficiently and to enable sharing and combination with other data.

According to the process mentioned in the General Description of Semantic Annotation Issues section there were realized three steps:

1. Selection of fitting approach for data modelling and harmonization related to semantic annotation.
2. Development of the data model emphasized semantic issues.
3. Testing of data and searching for new components/links to third party linked data.

3.1 Selection of Fitting Approach for Data Modelling and Harmonization Related to Semantic Annotation

Based on research of existing solutions (e.g. Open POIs by Open Geospatial Consortium) and studies there was chosen the RDF standard (https://www.w3.org/RDF/) as the fundamental component of the SPOI. The first phase was based on a development of the initial version of data model, which contained only basic properties (such as label, geometry or classification). This model used only new properties from the SPOI namespace (http://www.openvoc.eu/poi#). During further incremental update (combination of steps 2 a 3) the specific properties of data model were replaced by standards or existing relations as well as the model was extended on basis of user requirements and information provided by input data.

3.2 Development of the Data Model Emphasized Semantic Issues (SPOI Ontology)

The semantically annotated data model was designed based on a literature review and a thorough analysis of available POI data, existing standards and data models. It is designed in a way that it can be easily extended based on the user requirements (for example properties such as poi:openningHours or poi:access were added, because these properties exist in input data provided by users as well as they are important for tourist purposes). The current version of the SPOI data model has eight basic components, and is shown in Figure 1:

1. Identification – every POI is identified by a unique ID expressed as URI in harmony with Linked data requirements.
2. Labels & description – every POI is described by a label (name). There can be more labels attached to a single POI. In such case, the labels are differentiated by the xml:lang attribute. POIs can also contain a longer textual description in various languages (with use rdfs:comment property), if they are available
3. Geometry / Localization – every POI is localized by two coordinates (latitude and longitude) of the World Geodetic System (WGS) 84. WGS84 represents the most used, accepted and universal system, which is usually transformable to local systems and cartographic projections. Coordinates were originally published according to Basic Geo (WGS84 lat/long) Vocabulary. However, due to a better compatibility with the Virtuoso engine, all coordinates were transformed to GeoSPARQL standard. Latitude and longitude are written as WKT (Well-Known Text).
4. Classification – categorization of POIs is realized through **SPOI Classification Ontology** written in Web Ontology Language (OWL) and based on two classification systems: (1) classification based on GPS-based geographical navigation Waze, which represents the main categories in the ontology and is also used as a classification for visualization, and (2) classification based on OpenStreetMap, which represents subcategories in the SPOI ontology and provides more specific categorization. The

classification was originally realized only through these two classifications systems, but ontology enables to assign to one POI more categories (for example a POI can be a bank as well as ATM). Moreover, this approach improves semantics of the data and makes the integration of the SPOI data easier. The ontology includes also links to an equivalent object in Wikidata (using rdfs:seeAlso element) and links to other thesauri such as EuroVoc or GEMET, if these information are available. Every POI must contain at least one poi:class element.

The classification system used in Waze is quite short, clear and simple to visualize as well as differentiate, because it contains 10 well-defined categories: Natural features, Other, Transportation, Professional and public, Shopping and services, Food and drink, Culture & entertainment, Lodging, Car services, Outdoor. There are many different classification systems used in different domains. The Waze classification used as the main categories in the ontology was selected based on its simplicity and clarity, as well as its use in numerous applications. Since most of the data originate from OpenStreetMap, the subcategories in the ontology were inspired by classification types from OpenStreetMap. Harmonization of the source data classification, if exists, to the target classification, here SPOI Classification Ontology, is performed during the transformation through a predefined mapping between categories. The ontology categories are connected to data as URI's to self-standing RDF vocabulary. Mapping rules between ontology classification and categories used in other source data are kept in the transformation XSLT file.

5. Contact information – several POIs contain contact information such as address, e-mail, homepage, fax or phone number. Authors of the target model did not want to create new properties and decrease interoperability with other data. Therefore, existing vocabularies, for example FOAF (Friend of a Friend) or LOCN (ISA Programme Location Core Vocabulary), were used.
6. Common (tourist) information – information such as opening hours, access to the Internet or accessibility.
7. Links – all POIs include one or more of three types of links to external data – (1) links to external non-linked data resources such as photos, Wikipedia or Wolfram|Alpha; (2) links to an equivalent object in DBpedia, Wikidata and GeoNames.org; (3) links to relevant countries (in DBpedia and GeoNames.org) containing the POI. The last type of links is mandatory for each object.
8. Feature metadata - basic information on data, for example origin of data, identifier, rights or date of embedding into SPOI dataset.

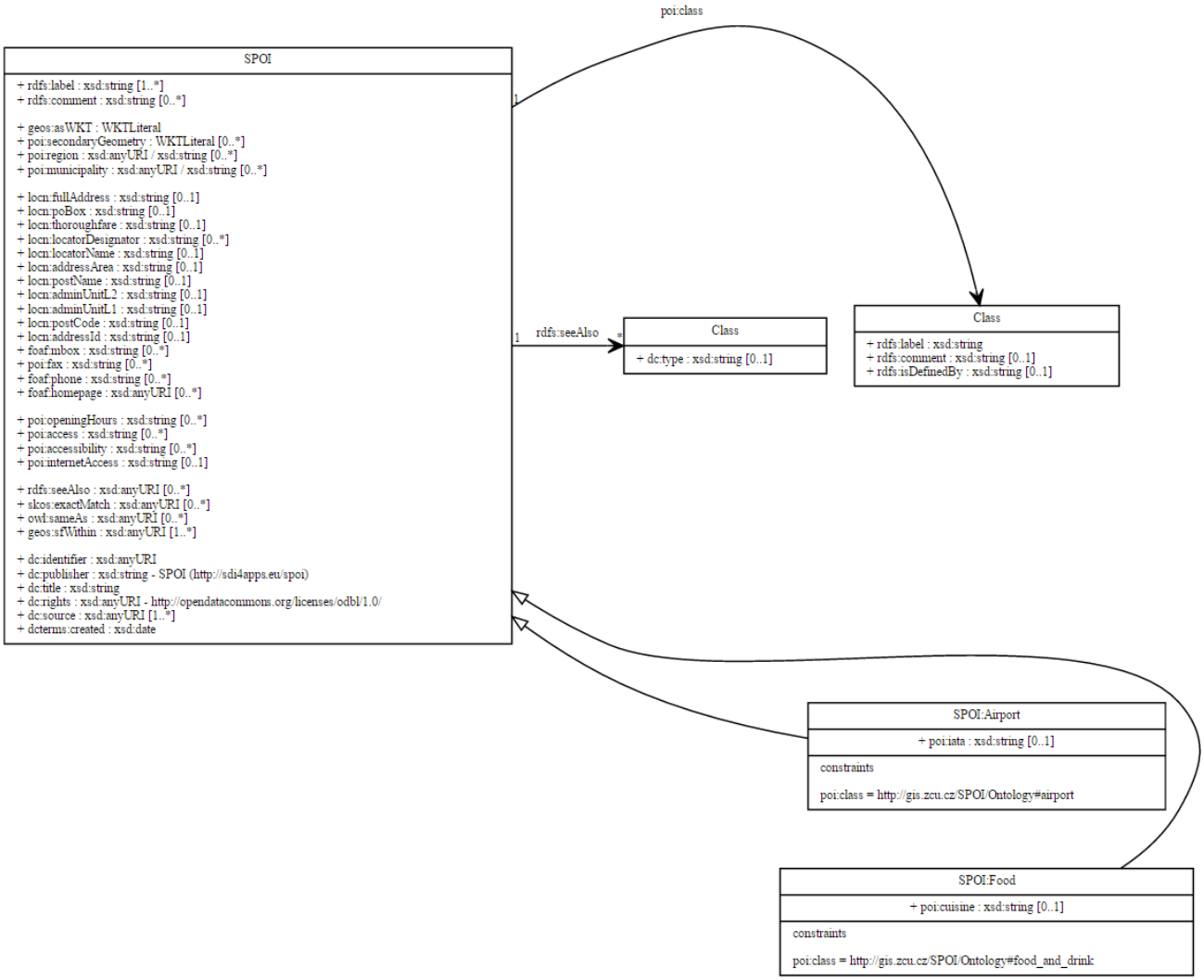

Figure 1 SPOI data model

3.3 Testing of Data and Improving the Data Model

The first version of SPOI was published in May 2015. Since then there were realized many activities connected to testing and implementation of new components of data model or replacing of particular components by an existing and more fitting parts. The following list summarizes the most important changes (from newest to oldest) from May 2015 to March 2017:

- 2017-01-18 New construction in data model - links to photo (<rdfs:seeAlso> element) contain metadata information (<dc:type> element) on type of the target document (value "image").
- 2016-10-07: The attribute poi:wikidata is replaced by standardized properties owl:sameAs and skos:exactMatch.
- 2016-08-30: Categorization of POIs realized through two classification systems based on Waze and OpenStreetMap is changed to SPOI Classification Ontology enabling to assign to one POI more categories.
- 2016-03-21: The property poi:address is replaced by address properties (locn:fullAddress, locn:poBox, locn:thoroughfare, locn:locatorDesignator, locn:locatorName, locn:addressArea, locn:postName, locn:adminUnitL2, locn:adminUnitL1, locn:postCode, locn:addressId) adopted from ISA Programme Location Core Vocabulary.

- 2016-03-21: The complete SPOI data set is generated with the new ID. The current ID was developed on the basis of discussion with experts in Open Transport Net (OTN) project (http://opentnet.eu). It is composed of prefix (for example OSM or GN), underscore character and ID adopted from original data or ID generated during data harmonization (in case of data resources not containing original ID). This change was realized in March 2016, but changes of ID were quite often (see following points from September 2016 and June 2016):
- 2015-09-02: With respect to need of keeping a persistent URI (which is not kept with xsl:generate-id usage during data updating = new data transformation) the new ID is generated as the combination of ISO 3166-1 alpha-2 country code, acronym of category of POI according Waze navigation data and both coordinates.
- 2015-06-18: The original ID (URI + code) is changed to the more understandable and readable form composed of URI (http://www.sdi4apps.eu/poi), ISO 3166-1 alpha-2 country code, category of POI according Waze navigation data and unique code (generated randomly by the XSLT script).
- Data model extended: new properties poi:region and poi:municipality describing optional link(s) to corresponding region (district, federal country...) and municipality.
- The asWKT element is extended by rdf:datatype="http://www.openlinksw.com/schemas/ /virtrdf#Geometry" (to better implementation in Virtuoso).
- The OpenStreetMap harmonization process is extended by the creation of links to relevant LinkedGeoData.org objects. Links are realized through owl:sameAs and skos:exactMatch relations.
- Changed from poi:email to standardized relation foaf:mbox.
- Changed from poi:phone to standardized relation foaf:phone.
- Changed from poi:www to standardized relation foaf:homepage.
- Adding new properties required by data providers and users, for example opening hours, IATA code of airport or metadata based on Dublin Core standard.

Latest version of the data model and more information on SPOI data, including data harmonization scheme, list of changes and metadata are available via the SPOI web page (http://sdi4apps.eu/spoi/) in the “Links” section.

4 CONCLUSIONS

The changes of the form of identifier illustrate the never-ending “fight” between semantics and readability on the one side and better implementation on the second side. All three proposed and used version of ID are not unambiguous – there might be several POIs of the same category at the same place. Also, a management of persistent ID is and will be complicated in case of data updating. The last discussion with experts from W3C, SDI4Apps and OpenTransportNetwork project led to the last change – the ID will be not readable for humans (the amount of information and semantics were decreased), but current solution using existing ID from original resources will be better from the view of persistent ID (the responsibility for persistence was returned to original data providers).

The situation described in the previous paragraph shows that semantic annotation is really never-ending process with an unsure result, because there is usually not one right solution. It does not mean that the semantic annotation of data is not helpful. It is necessary to mention that semantics plays key role in data interoperability (data sharing and combining) as well as in activities such as Linked data or Open data. Finally, the semantic annotation is very important for all users of spatial data, because it limits possible errors connected to wrong data interpretation.