SEMANTIC ANNOTATION

MARCH 2016

DELIVERABLE

Project Acronym: **SDI4Apps**

Grant Agreement number: **621129**

Project Full Title: **Uptake of Open Geographic Information Through Innovative Services Based on Linked Data**

D5.2.1 SEMANTIC ANNOTATION 1

Revision no. 03

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

REVISION HISTORY

Revision	Date	Author	Organisation	Description
01	07/03/2016	Otaka Čerba	UWB	Initial draft
02	29/03/2016	Martin Tuchyna, Karel Charvat	SAZP/CCSS	Internal review
03	30/03/2016	Otakar Cerba	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This deliverable contains original unpublished work except where clearly indicated otherwise.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 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:

  o 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.

  o 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. It is standardized as ISO 15836. Dublin Core elements are in SPOI used for description of feature metadata such as data provider, original data resource, date of integration to SPOI or license.

- Friend of a friend (FOAF; http://www.foaf-project.org/) is a machine-readable ontology 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 or web page.

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

- ISA Programme Location Vacabulary (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. With help of LOCN SPOI data enables to split particular components of addresses such as street, post name or post code.

- Web Ontology Language (OWL; https://www.w3.org/TR/owl2-overview/) is designed for description of ontological systems. 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 or LinkedGeoData.

- Resource Description Framework Schema (RDFS; https://www.w3.org/TR/rdf-schema/) is a set of classes with certain properties using the RDF extensible knowledge representation data model, providing basic elements for the description of ontologies, otherwise called RDF vocabularies, intended to structure RDF resources. SPOI uses classes for labels, comments (longer textual descriptions) or rdfs:seeAlso for linking to external resources such as Wikipedia pages.

- 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, GeoNames.org or LinkedGeoData.

There are also other 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 INSPIRE and Linked Data (https://joinup.ec.europa.eu/asset/are3na-reuse/description#INSPIRE)

- SmOD INSPIRE Vocabularies (https://www.w3.org/2015/03/inspire/)

3 EXAMPLES IN SDI4APPS

SPOI represents the biggest open dataset of POIs using the Linked data approach. It contains more than 24 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

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 purpuses). 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 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 two classification systems: (1) classification based on GPS-based geographical navigation Waze, which is the primary, mandatory, linked as a vocabulary and used classification for visualization, and (2) classification based on OpenStreetMap (optional). 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 was selected based on its simplicity and clarity, as well as its use in numerous applications. Harmonization of the source data classification, if exists, to the target classification, here Waze or OpenStreetMap, is performed during the transformation through a predefined mapping between categories. The Waze categories are connected to data as URI's to self-standing RDF vocabulary. Since the majority of the data originate from OpenStreetMap, two classification types from Open Street Map were used. Mapping rules between the Waze nomenclature, the OpenStreetMap 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, cuisine, access to the Internet, code of airports 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 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.
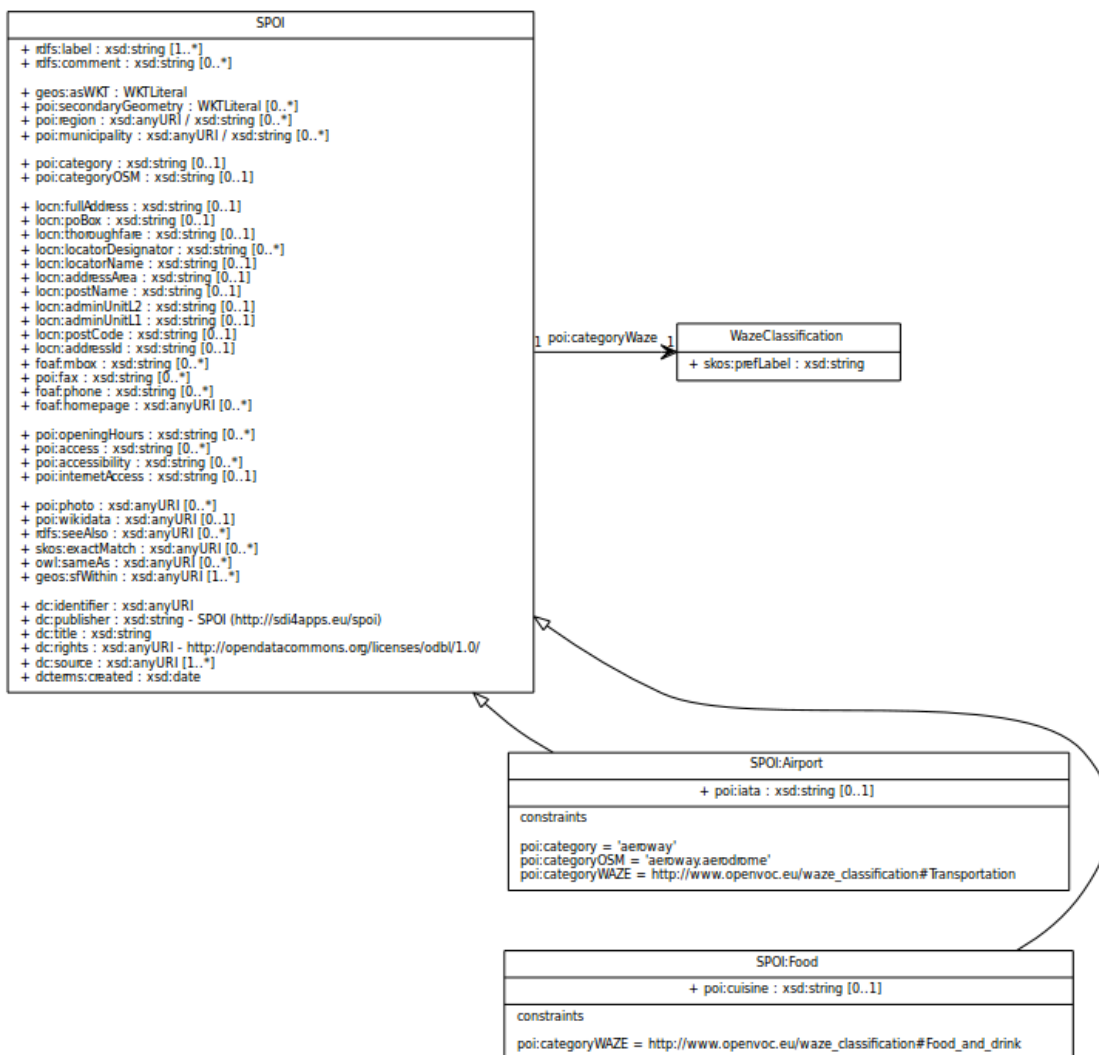
Figure 1 SPOI data model

3.3 Testing of Data and Improving the Data Model

The first version of SPOI was published in May 2015. From May there were realized many activities connected to testing and implementation of new components of data model or replacing of particular components by a existing and more fitting parts. The following list summarizes the most important changes from May 2015 to March 2016:

- The property poi:address was 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 Vacabulary.

- The complete SPOI data (24 106 299 POIs) are 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) was 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 was extended by rdf:datatype="http://www.openlinksw.com/schemas/virtrdf#Geometry" (to better implementation in Virtuoso).

- To the OpenStreetMap harmonization process was extended by the creation of links to relevantLinkedGeoData.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 are available via the SPOI web page (sdi4apps.eu/spoi) in the section links.

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 a 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. Last but not least the semantic annotation is very important for all users of spatial data, because it limits possible errors connected to wrong data interpretation.