Smart Points of Interest: Big, Linked and Harmonized Spatial Data

Otakar Cerba and Tomas Mildorf

ABSTRACT: The current trends in geomatics and geoinformatics are very often connected with three words – "big", "linked" and "harmonization". In many cases this triple is used just as an attractive label or buzzwords. But the future of the above mentioned disciplines consists in collecting, processing and visualizing of large data set, which can be created as a combination of existing data with links to other data. This is the most efficient way how to deal with spatial data in terms of data volume, speed of processing or intelligibility of data presentation and visualization.

The Smart Points of Interest (SPOI) dataset represents a real example of big and linked spatial data. Moreover, the data are created as a combination of many heterogeneous spatial data, including voluntary data. SPOI data is being developed in the international SDI4Apps project, an EU-funded project coordinated by the University of West Bohemia in Plzen, Czech Republic. SPOI data set is a crucial component of the Open Smart Tourist Data pilot of this SDI4Apps project.

The SPOI data contains global data (selected objects from OpenStreetMap, GeoNames.org and Natural Earth) and several local data sets (for example data from the Posumavi region, Czech Republic; Sicily, Italy; experimental ontologies including ski regions in Europe or historical sights in Rome, Italy or data from the Citadel On the Move project including various cities and regions over the world).

The added value of the SDI4Apps approach consists in implementation of the linked data approach and RDF format, using standardized and respected properties and development of a completely harmonized data set with a uniform data model and a common classification.
The current version of SPOI (June 2016) contains almost 24 millions POIs from the whole world. The SPOI data are transformed to the newly developed data model and published under the Open Database Licence as SPARQL endpoint (to query and download) as well as in the map client (to view).

The main objective of this paper is to to introduce the experience and knowledge gained during the development and the initial operational phase of the SPOI data set. This includes mainly the design of the data model, re-use of existing vocabularies and spatial data resources, harmonization and transformation process or visualization. The final reason why to present SPOI is to open discussion and exchange experience with other experts focused in linked spatial data.

KEYWORDS: big data, spatial data harmonization, tourism, linked data, RDF, XSLT, spatial data, data modelling.

Introduction

The current trends in geomatics and geoinformatics (as well as in geotechnologies as the application domain of both mentioned sciences) are very closely connected with big data, linked data and data harmonization. In many cases this triple is used just as attractive labels or buzzwords. But the future of the above mentioned disciplines as well as all related domains consists in collecting, processing and visualizing large data sets, which can be created as a combination of existing data with links to other data. This is the way how to deal with spatial data the most efficiently on the level of data volume, speed of processing or intelligibility of data processing, presentation and visualization. This is also related to elimination of errors, misinterpretation and misunderstanding of spatial data and information, because all items of the triple mentioned in the title of this paper is connected to semantics and standardization.

The main goal of this paper is to present the development process including the initial phase of management of the Smart Points of Interest (SPOI) data, which consist in harmonization of many heterogeneous data resources to a big data product based on linked data principals. The authors would like to mention mainly activities including data model design, re-using existing vocabularies and spatial data resources, harmonization and transformation process. The provided information could serve as a "best practice" of how to create a harmonized, big and linked spatial data set. The paper also tries to point out specifics of spatial data in the harmonization process. The final reason why to present SPOI is to open discussion and exchange experience with other experts focused in linked and big spatial data.

The SPOI data set represents a real example of big and linked spatial data. Moreover, the data are created as a combination of many heterogeneous spatial data, including voluntary data. The SPOI data is being developed in the international SDI4Apps project, an EU-funded project coordinated by the University of West Bohemia in Plzen, Czech Republic.

The SPOI data includes global data resources (selected objects from OpenStreetMap, GeoNames.org and Natural Earth) and several local data sets (for example data from the Posumavi region, Czech Republic; Sicily, Italy; experimental ontologies including ski regions in Europe or historical sights in Rome, Italy or data from the Citadel On the Move project including various cities and regions over the world).

The main added value of the SDI4Apps approach consists in implementation of linked data approach (including supporting the 5-star ranking scheme), RDF (Resource Description Format) format, using standardized and respected vocabularies and development of a completely harmonized data set with a uniform data model and a common classification.

The current version of SPOI (June 2016) contains almost 24 millions POIs from the whole world. The SPOI data are published under the Open Database Licence as SPARQL endpoint (to query and download) as well as in the map client (to view).

The paper is structured as follows. Section Methods describes the main processes of the SPOI development – designing the data model, re-using existing vocabularies, linking to external resources and the complete data harmonization process. Section Results describes SPOI as the output of the previously mentioned processes. The conclusions, future steps and main benefits are summarized in the last section.

Method

The SPOI development can be considered as a specific data harmonization task, because it includes processing of many heterogeneous data resources and their integration to a final data set. Janecka et al. (2013) proposed a general framework how to harmonize data coming from the spatial planning domain. Nevertheless, the proposed 5-step harmonization framework is also applicable to Linked Data such as SPOI, because there are many similarities to Linked (Open) Data lifecycles published in Bauer & Kaltenböck (2011) or Villazón-Terrazas et al. (2011). The following paragraphs describe specifics of the implementation of this approach to Linked Data.

The slightly modified 5-step harmonization framework (presented in Cerba et al. 2016a) is composed of two main phases – a cognitive phase, including steps 1-4 (Harmonization theory, Input data knowledge, Output data knowledge and Design & development of particular harmonization steps), and a technical phase (step 5 – Practical realization of data harmonization).

The harmonization theory of the SPOI development (Step 1 of the harmonization framework) is based on three types of resources:

- documents focused on Linked (Open) Data publishing in general (Bizer et al., 2008; Bizer et al., 2009; Bauer & Kaltenböck, 2011),
- texts dealing with Linked Data considering spatial data (Auer et al., 2009; Kuhn et al. 2014),
- researches related to a modelling and harmonization of spatial data (Goodchild, 1992; Shekhar et al., 1997; Longley et al., 2001; Janecka et al. 2013).

Table 1 (next page) shows steps 2-5 of the harmonization framework, relevant Linked Data development components based on Bauer & Kaltenböck (2011), Villazón-Terrazas et al. (2011) and Ding et al. (2012) and SPOI development issues.

Results

The SPOI data set is being developed in the international SDI4Apps project as part of the Open Smart Tourist Data pilot. It is is an EU-funded project coordinated by the University of West Bohemia in Plzen, Czech Republic. The university is also the main developer of SPOI.

The SPOI data covers the entire world, but the biggest number of SPOI users can discover in large highly developed countries (for example in the United States or in Germany) and in countries supporting spatial data publishing and VGI activities such as Switzerland or the United Kingdom.

Table 1: Overview of 5-step harmonization framework, Linked Data development components and SPOI development issues.

Step of harmonization framework	**Linked Data development component**	**SPOI development issue**
2. Input data knowledge	Data analyses and cleaning	Input data are very heterogeneous (see Figure 2). Therefore, it is necessary to find information that is important from the view of data integration. This information includes identifier management, occurrence of the mandatory attributes of the model (Figure 1), coordinate system, updating frequency or licence (which has to be in harmony with ODbL).
3. Output data knowledge	Specification of target data	The output is determined by user requirements (which information about interesting places has to be or should be provided) as well as by demands on Linked Data (published for example in Bizer et al. 2009). Both types of requirements were taken into consideration during the data model (Figure 1) development in the next phase.
4. Design & development of particular harmonization steps	Data modelling Linking to vocabularies and other semantic resources RDF conversion	Data modelling consists in development of the uniform data model (Figure 1) and creating mapping between attributes in source and target data. There are several specific harmonization steps (Figure 2) for Linked Data such as linking to semantic resources and conversion to RDF format. The SPOI data uses the following vocabularies – Web Ontology Language (OWL), Simple Knowledge Organization System (SKOS), Friend of a friend (FOAF), GeoSPARQL, Dublin Core and ISA Programme Location Vacabulary.
5. Practical realization of data harmonization	Interlinking Licensing Linked Data generating and publishing	The SPOI data are connected to these external data resources – DBpedia, LinkedGeoData, GeoNames.org and Wikidata. SPOI is published under the Open Database License (ODbL). SPOI is available in a map client (Figure 3) and as a SPARQL endpoint (details in the next chapter).

The data uses a common classification system based on the nomenclature implemented in the Waze open navigation system. This classification splits all POIs to ten categories – Natural features (~30%), Transportation, Professional and public, Shopping and services, Food and drink, Culture and entertainment, Lodging, Car services, Outdoors and Other. This classification is very rough and is used only for visualization in large scales. For a

detailed classification, a combination of the value and the key from OpenStreetMap is used. But this information is not mandatory for all data records (only 86% of POIs contain this classification).

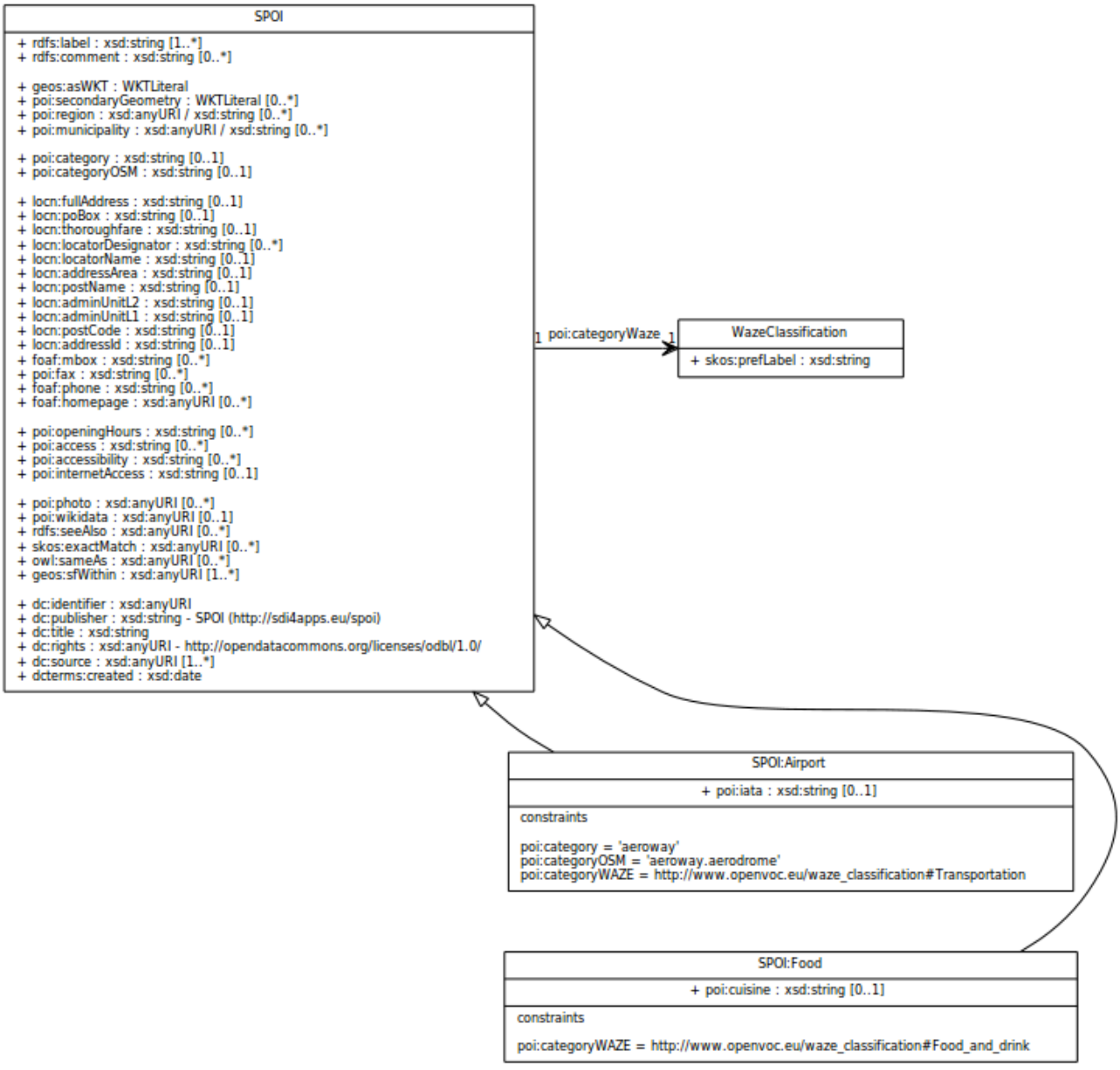

Figure 1: SPOI data model.

In the near future, the developers will introduce an important change in data classification – to use a system of keywords based on selected vocabularies or thesauri. This approach will enable 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.

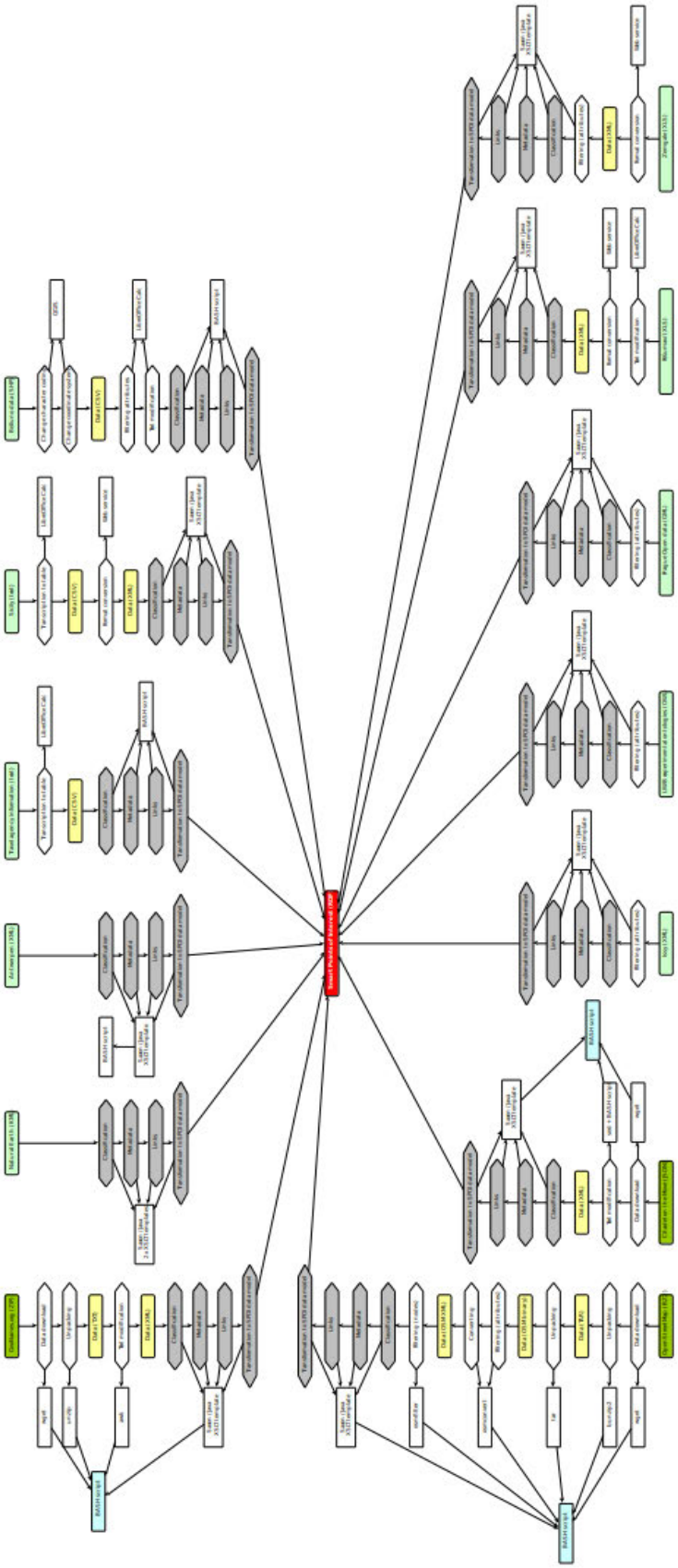

Figure 2: SPOI data harmonization scheme.

As it was mentioned in the previous section, the SPOI data originates from many heterogeneous open data resources. They could be divided into two main groups:

1. Global resources containing seamless data from almost all countries of the world. These data resources are provided by international organizations and in several cases use a VGI approach. This group includes DBpedia, GeoNames.org and partially Natural Earth (from this data two components are imported and harmonized – world's airports and natural parks in the United States).

2. Local resources covering limited area (a region or a city). These data sets can come from their producers (usually partners of the SDI4Apps consortium or the OpenTransportNet consortium) or are available on the Web. The first sub-group includes data from Antwerp (Belgium), Belluno (Italy), Issy (France), Prague (Czechia), Pošumaví region (Czechia), Zemgale region (Latvia), Sicily (Italy), experimental ontologies including ski

regions in Europe or historical sights in Rome or data from a Czech travel agency. The second group contains Natural Earth (partially) or the Citadel on the Move project (about 30 cities or regions of the world).

The map client presenting SPOI is based on the HSLayers library (a clone of Open Layers). There are used various background maps (OpenCycleMap, MTB map, OpenStreetMap and specific tiles developed for cycling in the Plzen region in the Czech Republic). Users are able to annotate data, change the order of layers (primary SPOI categories and selected types of objects such as restaurants) and their properties (such as transparency) or connect an external map or data layer (for example Panoramio or OpemWeatherMap).

The data are stored in the Virutoso engine as RDF triples. Querying and downloading the data are available via a SPARQL endpoint providing data in many different formats (XML, RDF, CSV or JSON).

All information obout SPOI is available on the SPOI web page (gis.zcu.cz/spoi). This web page includes not only a textual information, but also links to graphical schemes (data model, data harmonization schema), the map client and the SPARQL endpoint. Further information about SPOI is available in many papers and articles published during the last years (Cerba et al., 2015, Cerba et al., 2016a, Cerba et al., 2016b).

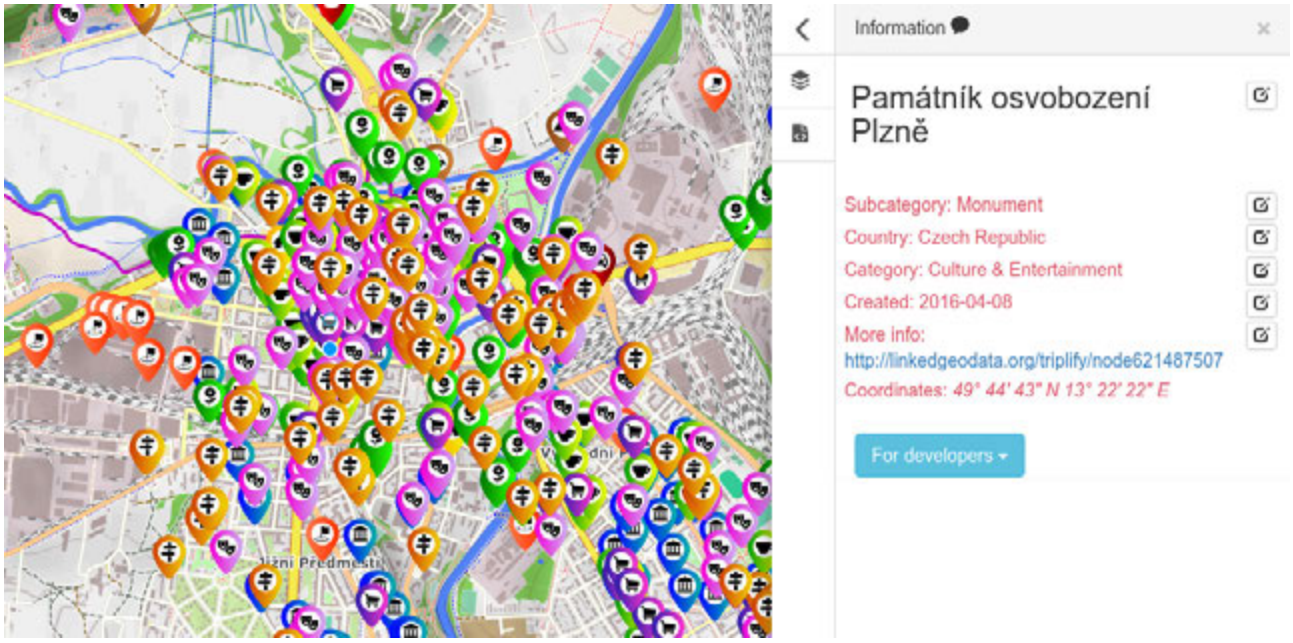

Figure 3: SPOI map client.

Conclusions

Smart Points of Interest is a very large data set containing points of interest in the whole world. SPOI covers mainly the domain of tourism (lodging, natural features or various

monuments), but it also touches transportation or protection of historical heritage and environment. The SPOI data set contains almost 24 millions of records (POIs), which are interlinked to other external data and published under the ODbL via the map client (Figure 3) and the SPARQL endpoint. Data are adopted from many heterogeneous original resources, therefore, a complicated harmonization process (Figure 2) had to be developed.

The SPOI data has the following essential attributes which differentiate SPOI from other similar data sources (for example OpenPOI, a short comparison of SPOI and OpenPOI is published in Cerba et al., 2015):

- Many heterogeneous input data.
- Complex data harmonization process (Figure 2).
- Uniform data model (Figure 1) based on standards, semantic description and Linked data.
- Seamless data.
- Published as a web service (Figure 3) and as a SPARQL endpoint.
- Available under the Open Database License (ODbL).

The proposed harmonization process (innovated 5-step harmonization framework) could be re-used for other activities related to harmonization and integration of various spatial data into the form of Linked Data.

SPOI could be used as data (together with a map layer) for any application focused on tourism, travelling and promoting of regions or localities. It has also a big potential for advertising purposes or analyses of trends in tourism.

In addition to tourism, SPOI can be used for example in car navigation systems. Considering the number and the heterogeneity of input data sources, the linked data approach seems as the only way how to keep such a large data set up-to-date without major efforts. However, producers of car navigation systems prefer to maintain their own copy of POI data and not relying on links to third party data sources. Despite so many advantages of SPOI, there exist some drawbacks, especially in relation with commercial activities.

The SDI4Apps project contributes to the EU Strategy of the Danube Region (EUSDR) within the initiative of the Joint Research Centre of the European Commission – the Danube Reference Data and Services Infrastructure (http://drdsi.jrc.ec.europa.eu/). The SPOI data set is one of the data sources contributing to this SDI and plays an important role especially for the Priority Area 3 Culture & Tourism.

References

Auer, S., Lehmann, J., & Hellmann, S. (2009). *Linkedgeodata: Adding a spatial dimension to the web of data* (pp. 731-746). Springer Berlin Heidelberg.

Bauer, F., & Kaltenböck, M. (2011). *Linked open data: The essentials*. Edition mono/monochrom, Vienna.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. *International journal on semantic web and information systems*, 5(3), 1-22.

Bizer, C., Heath, T., Idehen, K., & Berners-Lee, T. (2008). Linked data on the web (LDOW2008). In *Proceedings of the 17th international conference on World Wide Web* (pp. 1265-1266). ACM.

Cerba, O., Charvat, K., Mildorf, T., Berzins, R., Vlach, P., & Musilova, B. (2016). SDI4Apps Points of Interest Knowledge Base. In *Progress in Cartography* (pp. 229-237). Springer International Publishing.

Cerba, O., Berzins, R., Charvat, K. & Mildorf, T. (2016). *Smart POI: Open and linked spatial data*. In European Geosciences Union, General Assembly 2016, Vienna (Austria).

Cerba, O., Mildorf, T., & Berzins, R. (2015). Designing SDI4Apps POI Base. In *First Joint International Workshop on Semantic Sensor Networks and Terra Cognita, The 14th International Semantic Web Conference*, Bethlehem (USA).

Ding, L., Peristeras, V., & Hausenblas, M. (2012). Linked open government data [Guest editors' introduction]. *Intelligent Systems*, IEEE, 27(3), 11-15.

Goodchild, M. F. (1992). Geographical data modeling. *Computers & Geosciences*, 18(4), 401-408.

Janecka, K., Cerba, O., Jedlicka, K., & Jezek, J. (2013). Towards Interoperability Of Spatial Planning Data: 5-Steps Harmonization Framework. *International Multidisciplinary Scientific GeoConference: SGEM: Surveying Geology & mining Ecology Management*, 1, 1005.

Kuhn, W., Kauppinen, T., & Janowicz, K. (2014). Linked data-A paradigm shift for geographic information science. In *Geographic Information Science* (pp. 173-186). Springer International Publishing.

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2001). *Geographic information system and Science*. England: John Wiley & Sons, Ltd.

Shekhar, S., Coyle, M., Goyal, B., Liu, D. R., & Sarkar, S. (1997). Data models in geographic information systems. *Communications of the ACM*, 40(4), 103-111.

Villazón-Terrazas, B., Vilches-Blázquez, L. M., Corcho, O., & Gómez-Pérez, A. (2011). Methodological guidelines for publishing government linked data. In *Linking government data* (pp. 27-49). Springer New York.

Otakar Cerba, Department of Geomatics, Faculty of Applied Sciences, University of West Bohemia, Plzen, Czech Republic

Tomas Mildorf, Department of Geomatics, Faculty of Applied Sciences, University of West Bohemia, Plzen, Czech Republic