SUPPORTING SPATIAL DATA HARMONIZATION PROCESS WITH THE USE OF ONTOLOGIES AND SEMANTIC WEB TECHNOLOGIES

M. Strzelecki[a] *, A. Iwaniak[a], J. Łukowicz[a], I. Kaczmarek[a,]

[a] Wrocław University of Environmental and Life Sciences, Wrocław, Poland – (adam.iwaniak iwona.kaczmarek)@up.wroc.pl, (jlukowicz marek.strzelecki)@gmail.com

Commission II, ICWG II/IV

KEY WORDS: Semantic interoperability, spatial data harmonization, Open Data, Semantic

ABSTRACT:

Nowadays, spatial information is not only used by professionals, but also by common citizens, who uses it for their daily activities. Open Data initiative states that data should be freely and unreservedly available for all users. It also applies to spatial data. As spatial data becomes widely available it is essential to publish it in form which guarantees the possibility of integrating it with other, heterogeneous data sources. Interoperability is the possibility to combine spatial data sets from different sources in a consistent way as well as providing access to it. Providing syntactic interoperability based on well-known data formats is relatively simple, unlike providing semantic interoperability, due to the multiple possible data interpretation. One of the issues connected with the problem of achieving interoperability is data harmonization. It is a process of providing access to spatial data in a representation that allows combining it with other harmonized data in a coherent way by using a common set of data product specification. Spatial data harmonization is performed by creating definition of reclassification and transformation rules (mapping schema) for source application schema. Creation of those rules is a very demanding task which requires wide domain knowledge and a detailed look into application schemas. The paper focuses on proposing methods for supporting data harmonization process, by automated or supervised creation of mapping schemas with the use of ontologies, ontology matching methods and Semantic Web technologies.

1. INTRODUCTION

Recent development in geographic information systems and services available within spatial data infrastructures and commercial solutions, along with very sharp increase in the number of users of those systems, has proven that the need for spatial information is significant and still growing. The emergence of new IT technologies causes the appearance of new possibilities and approaches for sharing this type of information. As a result, spatial data is now widely used by both professionals and common users.

Spatial data sets maintained by government administration are usually published under spatial data infrastructures or, more recently, in Open Government Data catalogues where it can be accessed free of charge and re-used by users or developers (Open Knowledge Foundation, 2012). Due to this new way of publishing and increasing data availability, it is necessary to provide proper formats and structure for this data. Creation of a new model or application schema is only a part of the work needed to be put into process of publishing spatial data. Usually, the source data had already been collected and processed under other tasks or projects. Each of these data sets often has various characteristics and is expressed according to different model or application schema. In that case, we are dealing with heterogeneous data. Publication of open data can be done in different ways: one is to publish unprocessed source data - the “raw data” approach, and the second is “cooked data”, where data has been processed to carry some information value. In the first case, data consumer has a possibility to process the data and to extract the information himself. However, when data consumer doesn’t have essential knowledge or resources, it would be easier for him to just download the data expressed with given application schema and to limit data processing to integration.

One of the major problems associated with the publication and integration is spatial data interoperability – both syntactic and semantic. It is especially important for open data, which should be highly interoperable for the purpose of integration and usage in publications, reports and applications. Harmonization is the process of preparing interoperable spatial data sets in a way that allows the provision of access through spatial data services in a representation that allows for combining it with other harmonized data in a coherent way (INSPIRE, 2013).

Harmonization of spatial data from heterogeneous sources into one common application schema is a demanding task. The authors used their previous experience in carrying out the spatial data harmonization and began researching the possibility of supporting the process of harmonization, with semi-supervised automation of certain aspects, using ontologies and Semantic Web technologies.

The aim of this paper is to present proposed methodology and workflow for automating and supporting the creation of the mapping schema between two different application schema, which is crucial during the process of spatial data harmonization. Authors justify the need for research in this area and the use of ontologies and Semantic Web technologies. Description of methodology will be followed by indication of common problems and potential solutions. The paper will also cover the opportunities for further research of the presented issue.

This contribution has been peer-reviewed.
doi:10.5194/isprsarchives-XL-7-W2-231-2013

2. MOTIVATION

Due to the diversity of geographic information systems, data gathered at the local level may be available in different forms. Usually, it is stored in spatial databases, which structure (physical model) is based on the guidelines derived from conceptual model. The data is stored primarily for the use by administration that has relevant knowledge of its structure, characteristics and knows how to use it to extract useful information. Spatial data sets are usually published within spatial data infrastructure in the form of geospatial web services in accordance with the standards provided by Open Geospatial Consortium (Nebert, 2009). Access to those services is standardized, but the data itself can be different syntactically or semantically. Therefore services provided within SDIs are often not interoperable, the data cannot be easily exchanged, integrated and used.

Creation of European Spatial Data Infrastructure is the basic concept behind INSPIRE (Infrastructure for Spatial Information in Europe) project. The regional infrastructure will be based on local infrastructure maintained by the member states. Because of the diversity of architectures and data models in local infrastructures and the need to access, exchange and process it in a consistent way, implementation rules of INSPIRE focuses on the interoperability of network services and data sets. Interoperability can be achieved by providing means for both syntactic (technical) interoperability and semantic interoperability (Toth, 2012). INSPIRE defines interoperability as the possibility to combine spatial data sets and services from different sources in a consistent way, as well as to provide access to it through network services. According to implementation rules, maintaining interoperability of both data sets and services is an important task. Providing syntactic interoperability based on well-known data formats and communication protocols is relatively simple, unlike providing semantic interoperability, which can be nontrivial task due to multiple possible data interpretation. Apart from being syntactically interoperable, data sets from two different services can have different meanings, thus they are not interoperable (Harvey, 1999).

Achieving syntactic interoperability in systems used for spatial information processing has provided many improvements in aspects of spatial data sharing. It was maintained by the use of standards in geospatial service architectures and common data models. Exchanged data has become consistent and widely available to use in different systems. As important progress in technical interoperability was made, semantic interoperability still remained as a problem. The simplest example of issue connected with semantic interoperability can be the problem of expressing height of the building (Kuhn, 2008). Common data model containing building properties can define height of the building as integer data type. In one system it means that height is expressed in meters above ground but in another it can be expressed as a number of floors. In both systems it is still defined as integer number, so technical interoperability condition is fulfilled. But in scope of semantic interoperability meaning of those two numbers is different which can lead to misinterpretation of spatial information.

Another important aspect in terms of sharing spatial data under Open Data initiative is the preparation of the interoperable data which can be integrated, processed and used. With the use of this approach, spatial data can be published in commonly accepted formats. Another approach is the publication of the data in Open Linked Data cloud, which focuses on the possibility of linking with other data sets.

Spatial data harmonization is a process aimed at achieving syntactic and semantic data interoperability, so the data can be accessed through geospatial web services and integrated with other harmonized data. First of all, to conduct this process it is necessary to review source and target application schemas. The next step is to prepare a mapping schema that defines the rules of reclassification and the transformation of entities (classes, attributes) from one application schema to another. With correctly prepared mapping schema it is possible to perform data transformation process with the use of dedicated software. This process can be performed on-line or off-line. In the first case the conversion occurs when the data is needed, usually when user sends request. In the case of off-line transformation, the process is performed for the whole data sets and then the converted data is published. There is many software solutions available, capable of transforming large amounts of spatial data (FME, ArcGIS Data Interoperability, HALE). It is also possible to perform this process with the use of XSL-Transformations.

However, the main problem associated with spatial data harmonization is the preparation of mapping schema. In practice, before the schema is expressed with the use of appropriate transformation software, it is created in other form, mainly in spreadsheets. This process requires a deep knowledge of source and destination application schemas which often requires domain knowledge. In case of large divergence between two schemas and its complexity, creation of mapping schemas requires large amount of time and resources. Objects and their properties from application schemas must be aligned in terms of both form and meaning. In case of matching multiple source schemas to one target schema, the mapping must be prepared for each source schema. Therefore, the assistance in the form of semi-automated support system will be highly welcomed.

The possibilities of expressing semantic aspects of application schemas, elastic data model, expressive ontology language and possibility of measuring compatibility and similarity of entities with the use of concept mapping techniques, make ontologies and the Semantic Web technologies the potential means to create methodology and solution for supporting spatial data harmonization process.

3. ONTOLOGIES AND SEMANTIC WEB TECHNOLOGIES

The term “ontology” emerges from philosophy. It is a study of the nature of being and existence. With the development of information technologies the term has taken on a new meaning, but is still connected to studies of being and description of reality. The new, computer science meaning was formed by T. Gruber as a “formal, explicit specification of a shared conceptualisation” (Gruber, 1993). Ontology describes specified model of reality with the use of conceptualization and hierarchization. Conceptualization can be defined as an abstract model of specified real life phenomena, which recognizes concepts from this phenomena. Description of reality is performed by answering the questions of: Which beings can exist or cannot exist? How can they be classified? Which relations can occur between them? Formal ontologies contains definition of classes, relations, properties and restrictions. Ontologies are often formalized with the use of description logic (DL), which enables a use of inference engines.

Ontology contains definition of concepts – sets of individuals which share the same properties, and describes associations between these concepts. This information can be used to structurize domain knowledge and to limit the possible interpretations of concepts. Ontologies consist of two parts – TBox and ABox. Terminology box (TBox) contains terminology part while assertional box (ABox) contains description of facts. Ontologies can be divided in two major groups, based on domain of knowledge. Those groups are: domain ontologies and top ontologies. Domain ontologies focuses on description of narrow knowledge, generally from one domain. Top ontologies or upper ontologies describes wide aspects of reality and creates representation of general knowledge (Goczyla, 2011).

Semantic Web is a further evolution stage of the Internet and WWW, which main goal is to transform distributed, published resources into a global knowledge base which contain information that is understandable for both people and machines (Berners-Lee, 2001). The basic concept of creating Semantic Web is the semantic annotation of resources with the use of common vocabularies and ontologies. It is viewed as one of the possible means of transition from Web 2.0, based on content created by users, to Web 3.0 where content can be processed by machines. Semantic Web preserves the advantages of the WWW and creates a new, semantic layer that supplements WWW resources with information representing meaning. Semantic Web solutions and technologies include:

- RDF - elastic data model,
- OWL – language for formal representation of ontologies,
- SPARQL – RDF graph query language.

When Semantic Web concept was stated and technologies connected with this project became popular, a new possibility for representation of meaning has emerged. The use of ontologies and Semantic Web technologies was perceived as a solution to the problem of achieving semantic interoperability.

4. PROPOSED METHODOLOGY

The main objective for potential semi-automatic system supporting harmonization of spatial data is to create a mapping between application schemas. The use of ontologies and Semantic Web technologies is justified with possibilities of representing the meaning of concepts, mostly formalized as classes in each application schema, and also the possibility of enriching semantic information with other sources like domain ontologies and thesauri. Authors have elaborated a methodology and outlined the workflow of the process (Figure 1), which potentially allows to support the user in creation of mapping schemas, which will be used later in spatial data transformation process.

The main input for described process are spatial application schemas prepared with guidelines from ISO19109 (general rules) and ISO19107 (representation of geometry) standards. The important step is to transform selected application schemas into the representation expressed in OWL/RDF which is the proper ontology representation according to Semantic Web technologies. It will be used in the next steps of the process. It is essential to enrich created ontological representation with domain ontologies and thesauri. In the next step the techniques of ontology mapping will be applied, which can be used to estimate the similarity of each concept from two ontologies. The proposals of matching classes and potential reclassification rules will be made according to results of ontology mapping. These rules will be presented to the user, who will be able to evaluate and modify it. Applied changes will have an impact on subsequent executions of the process. After accepting the final reclassification strategies, mapping file will be created, which can be used to transform given data sets. In the following subsections authors describe each steps of the proposed methodology, identify potential problems and present solutions.

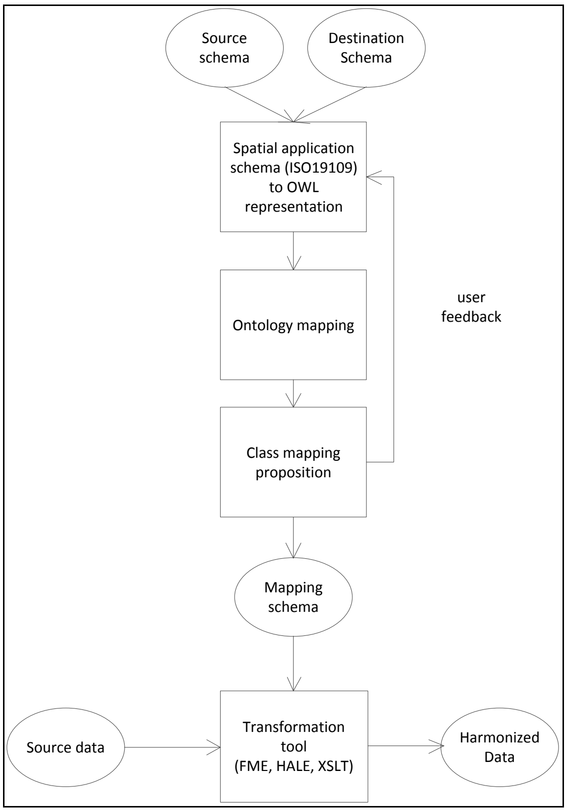

Source schema
Destination Schema
Spatial application schema (ISO19109) to OWL representation
Ontology mapping
user feedback
Class mapping proposition
Mapping schema
Source data
Transformation tool (FME, HALE, XSLT)
Harmonized Data

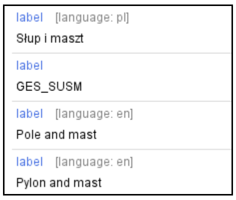
Figure 1. Proposed methodology workflow

4.1 Ontological representation of application schemas

The aim of this step is to perform automatic conversion of application schemas into OWL/RDF representation (and more specifically TBox part), which will be used in ontology mapping process.

Application schemas for geographic information systems are usually available as UML class diagrams or XML-Schemas (GML application schemas), which are converted from UML diagrams created in accordance with ISO19109 standard. This conversion can be performed with the use of software like Shapechange or Fullmoon. It is also possible that only database schema is available, but in this state of research authors limits the input to UML class diagrams, because of the problems connected with conversion from relational model into object model. If the diagram is available in XMI or EAP format, there is a possibility for generating OWL classes from UML classes. This conversion should be based on assumption that given diagrams were prepared in accordance with ISO19109 standard and contain stereotyped elements like: feature type, data type, enumeration and code lists. It is also important to remember that the reason for conversion is not to create exact and proper OWL representation of UML class diagram, but to extract classes and their characteristics, which can define their meaning in formal way and to serialize it as linked OWL classes. The tool which will be created to perform this conversion should use strategies that will maximize amount and quality of semantic information extracted from UML diagrams.

The main information stored in class diagrams about the meaning of the class is lexical information, which consists of elements such as the name of the class (primary and alternatives) and additional description, which can be often found in diagrams. Lexical information should be present in ontological representation as annotations (Figure 2). The name of the class can be expressed as "rdfs:label" property, and additional notes as "rdfs:comment".


label [language: pl]
Słup i maszt

label
GES_SUSM

label [language: en]
Pole and mast

label [language: en]
Pylon and mast

Figure 2. OWL Class annotations

Beside lexical information, the aspect that should be converted is hierarchy between classes, because structural information is widely used by ontology mapping techniques. Hierarchy of classes in UML diagram consists mostly of inheritance hierarchy, which can be easily converted to ontological representation.

Classes often have attributes which values are restricted to values defined in vocabularies and code lists. These attributes narrow the meaning of the class, and because of that, it is possible to extend structural information by redefining code list values as separate subclasses for class containing this attribute (Figure 3). In the process of defining mapping schema, these values are often reclassification conditions, and should therefore be considered at the stage of aligning concepts from ontologies.

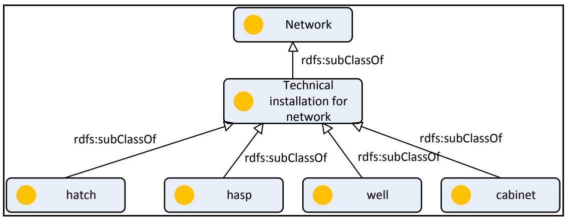


Figure 3. Class hierarchy

Class attributes should also be converted into OWL properties. According to the attribute type (from logical model) it can be converted to Object Property linking to another object or Data Property which links directly to values (literals). Type of the attribute can be expressed as property range, and attribute occurrence in class can be expressed as property domain. Properties can also contain lexical information as labels and comments, which can be extracted from diagram.

The essential element of spatial features is the possibility of containing geometry attributes, which representation in application schemas is defined by ISO19109 and ISO19107 standards. Geometric attributes can be mapped into object properties, which range should link to corresponding geometry type class from ISO19107 ontology (Figure 4).

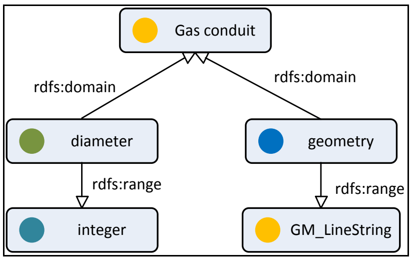


Figure 4. Class properties

Tools used for creation of logical models with the use of UML class diagrams often allows to insert constraints expressed in Object Constraint Language. It is possible to define conditions, impose restrictions and precise the meaning of diagram elements. The example of condition can be to assign more specific type of geometry, depending on some attributes value. It should be considered how OCL expressions can be mapped into ontological representation. The most natural possibility is the use of SWRL rule language, but it cannot be used during the process of concept matching. The other approach is to analyze OCL rule and to generate classes according to the possible results.

The result of this step is a simple application ontology, consisted mostly of TBox part and generated from UML diagram provided in XMI or EAP format. It is not a formal and proper conversion of application schema into ontology. It should be used to improve results of ontology mapping techniques. Additional results of this step will be a definition of linkages between OWL entities (their URI) and application schemas elements (identifiers).

4.2 Ontology and concept mapping

Ontology mapping is a process aimed at finding correspondences between concepts from two ontologies. Such concepts can be class or property. Ontology mapping (or ontology matching) analyzes the ontologies and indicates the similarities between each concept. Basically, tools that carry out the automatic mapping process use the following information about concepts included in the ontology (Euzenat, 2007):

- lexical information, indicating similarity of class names and properties; dictionaries, such as WordNet, are often used.
- structural information, indicating similarity of structures, relations to other classes (hierarchy) or properties.

There are many tools performing the automatic process of concepts mapping such as Optima, FOAM, RiMOM, PROMPT and UFOme (Pirro, 2010). Due to the wide feature range, the openness of the code and the availability as a library, the authors conducted a trial mapping process of generated application ontologies using the Optima tool (Kolli, 2008). Attempts were made to perform simple reclassifications of application schema fragments and to compare them to the mappings performed earlier by experts (Figures 5 and 6). The results are provided as pairs of linked URIs of classes and properties.

Due to the characteristics of application schemas for systems that use spatial data, it may be necessary to modify some elements of the tool. Moreover, the dictionaries of general purpose (WordNet) may not be sufficient for the more specialized spatial application schemas. Thus it is possible to use domain thesauri containing hierarchical concepts from some domain. It will be necessary to perform a semantic enrichment process of generated ontologies using dictionaries or domain ontologies.

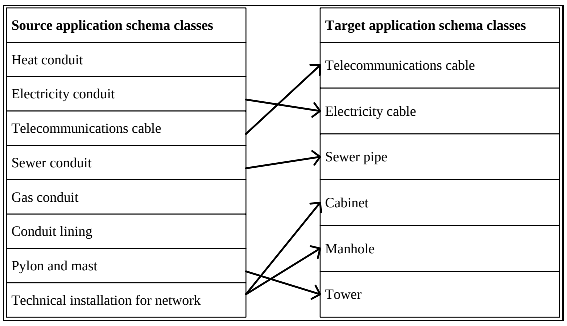


Figure 5. Simple reclassification prepared by expert

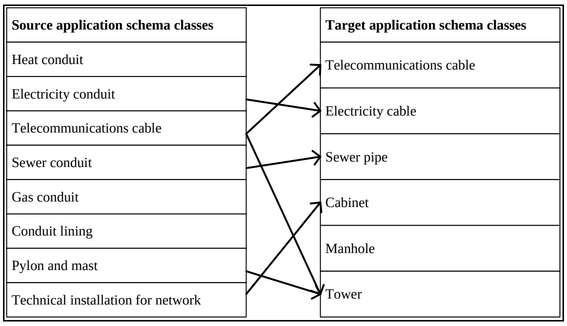


Figure 6. Simple reclassification prepared with the use of ontology matching

4.3 Mapping schema proposition

The results of ontology mapping process and the list of links between concepts from domain ontology and UML diagram elements can serve to link the classes and the attributes from the logical model. Thanks to the use of the ontological representation and the ontology mapping techniques, the linking occurred at the semantic level. Besides that, a syntactic compatibility should be also ensured, especially with regard to the attributes. It is necessary to match the multiplicity and, above all, the type of attributes. In case of incompatibility of the types, the possibility of conversion should be checked. This incompatibility can occur for specific values as early as during the process of transformation, so it should be minimized by a proper preparation and standardization of the data.

An important issue is to match elements typical for spatial application data that are compatible with ISO19107 and ISO19109. One such example is the geometry, which can often require processing. A typical use case is a conversion of polygons to points (centroids) or a conversion of polygons to bounding boxes.

The proposed mapping schema will be presented to the user. He should have a possibility to modify it. The final decision consisting in accepting is made by the user, so the whole process is supervised by him – therefore a semi-supervised spatial data harmonization process occures. Changes in the proposed mapping schema should be saved and have an impact on consecutive process executions, for instance by adding synonyms to thesauri used in the mapping process.

4.4 Mapping definition and data transformation

After the user accepts the proposed or modified mapping schema, it is necessary to export it into a format that can be loaded in the software allowing spatial data transformation. The authors assume the possibility of using tools such as FME or HALE. Mapping definitions for both tools are based on XML files, and in the case of HALE tool, the application source code is also available.

To ensure the portability, the possibility of batch transformation and the ability to implement the transformation within the framework of other tools, the possibility to export the mapping definition to XSL-Transformations files is also assumed. In this case however, the lack of functions that allow to process typically spatial data will constitute a problem. Nevertheless, there are XSL processors such as Apache Xalan, that allow the addition of new functions extending the processing capacities. Therefore it will be necessary to implement such extensions.

The final result of the described process will be a complete mapping definition for the spatial application schemas. The user who will supervise this process will be able to load the generated definition to a tool allowing the transformation, and thus he will receive harmonized spatial data.

5. FUTURE WORK

The authors has presented a proposition of methodology for the semi-supervised spatial data harmonization process. Its essential elements are constituted by the conversions of application schemas to a domain ontology, the ontology mapping and, on this basis, a proposition of a mapping schema for the harmonization process. This is just the beginning of the study on this issue, therefore each of these steps have to be thoroughly analyzed and tested. On the basis of these tests and analysis, it will be possible to propose improved methods to generate the most accurate applications schemas.

It is important to offer the best possible ontological representations for the applications schemas presented in the form of UML diagrams. Its goal is to obtain the best possible matches in the process of mapping concepts from ontology, so it is necessary to carry out efficiency tests of the strategy of preparing the application ontologies with reference to the applied ontology mapping techniques. Another important issue is the role of the user in the process, since he is the one that has the biggest impact on the look of the final scheme mapping. It is the authors aim that the modifications made by the user have an impact on the consecutive process activations.

The first attempts to perform the semi-automatic data harmonization on the basis of the proposed methodology has been carried out. Its results were referenced and compared to the mapping schemas prepared earlier by the specialists according to given application schemas and INSPIRE specifications. It is assumed that semi-supervised process of mapping schemas creation can result in significant reduction of the time and the resources necessary to carry out the spatial data harmonization.

6. REFERENCES

References from Journals:

Berners-Lee, T., Hendler, J., Lassila O., 2001, The Semantic Web, *Scientific American*, pp.29-37

Gruber, T., 1993, A translation approach to portable ontology specifications, *Knowledge Acquisition*, 5(2), pp. 199-220

Harvey, F., Kuhn, W., Pundt, H., Bishr, Y., Riedemann, C., 1999, Semantic interoperability: A central issue for sharing geographic information, *The Annals of Regional Science*, 33, pp.213-232

Pirro, G., Talia, D., 2010, UFOme: An ontology mapping system with strategy prediction capabilities, *Data & Knowledge Engineering*, 69(5), pp. 44-471

References from Books:

Euzenat, J., Shvaiko, P., 2007, *Ontology Matching*, Springer

Goczyla, K., 2011, *Ontologie w systemach informatycznych*, Akademicka Oficyna Wydawnicza EXIT

Kuhn, W., 2008. *Semantic interoperability. Encyclopedia of geographic information science*, SAGE Publications, pp. 383-385

Nebert, D.D., 2009, *The SDI Cookbook*, GSDI

References from Other Literature:

INSPIRE Data Specifications Drafting Team, 2013, Generic Conceptual Model

Kolli, R., Doshi, P., 2008, OPTIMA: Tool for Ontology alignment with Application to Semantic Reconciliation of Sensor Metadata for Publication in Sensor Map, Proceedings of 2008 IEEE International Conference on Semantic Computing, pp.484-485

Toth, K., Portele, C., Illert, A., Lutz M., Nunes, M., 2012, A Conceptual Model for Developing Interoperability Specificaions in Spatial Data Infrastructures, Publications Office Of the European Union

References from websites:

Open Knowledge Foundation, 2012, "Open Data Handbook Documentation", http://opendatahandbook.org/pdf/OpenDataHandbook.pdf (1 Sep. 2013)

Pollock, R., 2007, "Give us the data raw, and give it to us now", http://blog.okfn.org/2007/11/07/give-us-the-data-raw-and-give-it-to-us-now/ (1 Sep. 2013)

7. ACKNOWLEDGEMENTS

This research work has been supported through the projects granted by the National Science Centre, Poland, project number 2012/05/B/HS4/04197, 2012/05/N/HS4/00642 and 2011/03/N/HS4/03819.