SEMANTIC METADATA FOR HETEROGENEOUS SPATIAL PLANNING DOCUMENTS

A.Iwaniak [a], I.Kaczmarek[a,] *, J.Łukowicz [a], M.Strzelecki [a], S.Coetzee [b], W.Paluszyński[c]

[a] Wrocław University of Environmental and Life Sciences, Grunwaldzka 53, 50-357 Wroclaw, Poland- (adam.iwaniak, iwona.kaczmarek)@up.wroc.pl, (jlukowicz,marek.strzelecki)@gmail.com
[b] Centre for Geoinformation Science, Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Pretoria 0002, South Africa- serena.coetzee@up.ac.za
c Wroclaw University of Technology, Department of Cybernetics and Robotics, Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland- witold.paluszynski@pwr.edu.pl

KEY WORDS: semantic web, urban planning, metadata, land management, spatial development plan

ABSTRACT:

Spatial planning documents contain information about the principles and rights of land use in different zones of a local authority. They are the basis for administrative decision making in support of sustainable development. In Poland these documents are published on the Web according to a prescribed non-extendable XML schema, designed for optimum presentation to humans in HTML web pages. There is no document standard, and limited functionality exists for adding references to external resources. The text in these documents is discoverable and searchable by general-purpose web search engines, but the semantics of the content cannot be discovered or queried. The spatial information in these documents is geographically referenced but not machine-readable. Major manual efforts are required to integrate such heterogeneous spatial planning documents from various local authorities for analysis, scenario planning and decision support. This article presents results of an implementation using machine-readable semantic metadata to identify relationships among regulations in the text, spatial objects in the drawings and links to external resources. A spatial planning ontology was used to annotate different sections of spatial planning documents with semantic metadata in the Resource Description Framework in Attributes (RDFa). The semantic interpretation of the content, links between document elements and links to external resources were embedded in XHTML pages. An example and use case from the spatial planning domain in Poland is presented to evaluate its efficiency and applicability. The solution enables the automated integration of spatial planning documents from multiple local authorities to assist decision makers with understanding and interpreting spatial planning information. The approach is equally applicable to legal documents from other countries and domains, such as cultural heritage and environmental management.

1. INTRODUCTION

Spatial planning documents contain information about the principles and rights of land use in different zones of a local authority and form the basis for administrative decision making in support of sustainable development. Resources and tools for spatial planning in Poland are described in several types of spatial policy documents, appearing under different names such as concepts, plans and studies. One specific type of document, a spatial development plan, presented as example in this article, is an act of local law. It specifies land use and development principles through provisions that impose restrictions on the use of land. A spatial development plan consists of text and graphics (Fig.1). The textual part describes the spatial planning regulations for different zones and the graphics part presents the spatial objects (zones) to which these regulations apply. The text and graphics elements in a spatial development plan are linked to each other through textual descriptions. Currently, there are no technical standards for the content of spatial development plans. As a result, the content of these plans varies and is heterogeneous.

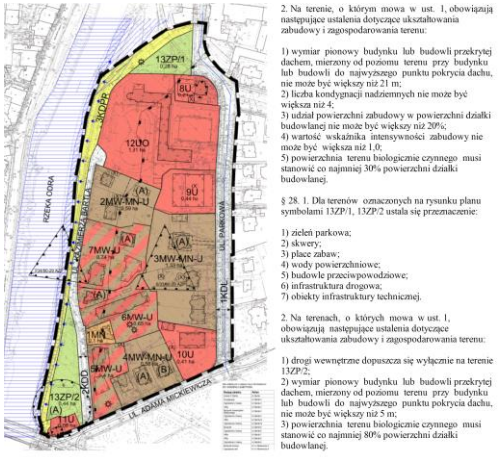

Figure 1. Example of a spatial development plan from Wroclaw municipality (source:http://uchwaly.um.wroc.pl/uchwala.aspx?numer=LXIV/1661/14)

The sharing of plans is regulated through the general principles of access to public information. Spatial development plans, as acts of local law, must be published in the regional Legal Registers (i.e. Official Journal of Voivodship) in electronic form and the text is freely available on the Web. The textual

* Corresponding author

part has to be published as an XML document, validated against a specified XML schema, which defines the editorial structure of the document, comprising elements such as headers, sections, chapters, and paragraphs. These plans do not include any description of the substantive content. There is also no standard method for linking the textual content of a spatial development plan to objects in the graphical part and vice versa. References to external resources or documents can only be made to legal documents, not to, e.g. attachments with additional information. The rigid structure of the schema prevents augmenting these documents with metadata or object classes about regulated issues. The approach described above was designed for isolated viewing of individual text documents. Because of the non-structured form of spatial planning documents this approach does not allow users to integrate, query and analyse multiple spatial planning documents in different contexts. It also hinders the querying and processing of spatial planning data in conjunction with information collected from other sources, such as, statistical offices, a variety of open public services (e.g. GeoNames, OpenStreetMap), open encyclopaedia (e.g. Wikipedia, DBpedia), dictionaries and thesauri (e.g. WordNet, GEMET).

In this article an alternative approach, based on Semantic Web technology, is presented. Semantic metadata are used to link the content of a spatial development plan with concepts in relevant ontologies. The document (in HTML or XHTML) is enhanced with semantic annotations in RDFa, making use of a spatial planning ontology describing the semantic content of the spatial development plan. This provides a precise description of the content and its meaning, enabling automated interpretation of entities defined in the XHMTL, their associated attributes and values, as well as relationships between those entities. A queryable knowledge base is prepared by extracting RDF triples from the XHTML pages. A use case for this alternative approach is described in the article to evaluate its efficiency and applicability. Results of the evaluation are presented and discussed, and future research directions described.

The remainder of the article is structured as follows: Section 2 provides a brief summary of related work on the Semantic Web and metadata for spatial planning; in Section 3 we explain how semantic metadata is embedded into a spatial development plan and provide an example; in Section 4 we present an experiment in which a spatial development plan is processed with the automated learning approach; in Section 5 a use case is presented and evaluated, followed by a discussion in Section 6 and finally, conclusions are drawn in Section 7.

2. RELATED WORK

The desire to increase the capabilities of the World Wide Web by publishing structured data and processing them by machines has led to the emergence and growing popularity of the Semantic Web, promoted by Tim Berners-Lee (Berners-Lee et al., 2001), and more recently - the Linked Data project[1]. The Semantic Web provides a set of technologies suitable for the creation and publication of metadata. Linked Data outlines basic principles for publishing and integrating distributed structured information with the use of those technologies. In the Linked Open Data (LOD) approach, links between data are based on objects, which have meaning defined in vocabularies (Heath and Bizer, 2011). RDF[2] is a specification of a data model which can be used for describing data and metadata of any resource published on the Web. It facilitates the integration and interoperability of heterogeneous data sets. The RDFa[3], loosely based on RDF, is a W3C recommendation for semantic annotation of documents. It constitutes a set of attributes, which are used for the semantic enrichment of HMTL and XHTML pages. RDFa is used in the implementation presented in this article.

In the context of spatial planning, semantic web technologies have been used to build intelligent geoportals in support of decision making in land management (Iwaniak et al., 2011). Related work on describing spatial planning documents was done in the Plan4all project. Results include a data model (Plan4all D4.2, 2010) and a metadata profile (Plan4all D3.2.2, 2010) for plans. In the Plan4all project, metadata for spatial plans is provided according to national legislation for the datasets included in the digital spatial plans and for spatial web services providing access to digital spatial plans. The metadata record contains a link to the document containing the spatial plan which it describes. The Plan4all metadata can be searched and discovered through 'traditional' metadata catalogue services, such as those available through spatial data infrastructures.

In this article, a different approach for providing metadata is followed. Instead of creating a (separate) descriptive metadata record for each spatial planning document, semantic metadata is embedded within the spatial planning document. The approach presented in this article is an implementation of the proposed bridge between the INSPIRE metadata infrastructure and distributed semantic resources through machine-readable metadata in the form of RDF graphs (Kubik et al., 2010). Linked Open Data removes the barriers to wide data interchange, which are often encountered when trying to publish documents with a complicated hermetic structure or documents prepared in natural language. Semantic annotations in spatial planning documents provide them with a well-described structure where the meanings of concepts are specified in accessible vocabularies. In this way, we achieve the capability to process them, like data retrieved from a structured database. Description of the content of spatial plans may involve links to features provided by OGC services, such as WFS. This makes it possible to dereference such digital objects in order to further process and visualize them.

In related work, Vilches-Blázquez et al. (2014) developed a process to generate, integrate and publish geospatial linked data from several national data sets. Yu and Liu (2015) present another example applied to heterogeneous observation data in the Sensor Web. Such resources published in the form of Linked Open Data can be considered to be a common, global knowledge base engine (Janowicz K., and Hitzler 2012).

3. EMBEDDING SEMANTIC METADATA IN A SPATIAL DEVELOPMENT PLAN

3.1 Overview

The first step in making spatial development plans both machine and human readable is the transformation from XML to XHTML, which can be easily automated. XHTML allows additional markup tags to express the semantics for spatial planning documents with RDFa, microdata or microformats.

[1] http://www.w3.org/DesignIssues/LinkedData.html
[2] http://www.w3.org/RDF/
[3] http://www.w3.org/TR/2015/REC-rdfa-core-20150317/

Since every valid XHTML is also valid XML, this operation could also have been performed on XML documents, but for the purpose of human readability, annotations were added to XHTML documents. The purpose of additional semantic markup is to improve the discoverability of spatial planning documents by standard and semantic web search engines and also to extract and properly use information from the documents in computer systems.

The next step is to annotate text with concepts from the spatial planning ontology described in Section 3.2. A customized web editor, based on the open source WYMeditor[4], was used to manually add semantic annotations to the document. The editing process is described in Section 3.3. Annotations are in the form of HTML tags containing RDFa attributes that identify fragments of natural language text (the informal model) and specify their relationship to the formal model of the real world (the spatial planning ontology). The tags include a phrase to identify the relevant concept in the formal model. The value of this concept can be specified in one of two ways: explicitly in the RDFa tag; or as a property value in the content of an annotated HTML tag.

Individuals indicated through references may be local (embedded in the same HTML document), or they could be external, such as resources available in a spatial data infrastructure or semantic resources. The RDFa metadata for external objects is in the same vocabulary as the rest of the document and therefore the external objects can be related to the model described in the annotated document.

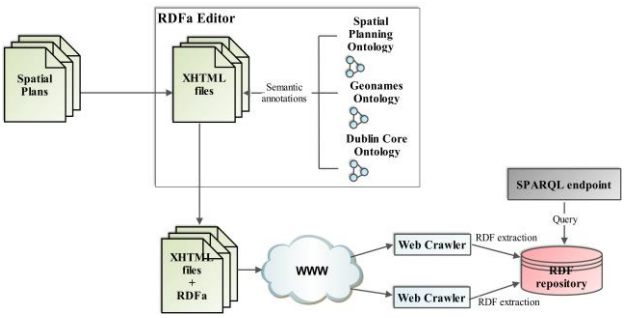

Figure 2. The concept of the solution presented in this article.

This process of document annotation and RDF graph extraction produces a formal model of the reality described in the textual content of the document. The annotated spatial planning documents can be published on the Web as standard XHTML documents on a standard HTTP web server (see Fig.2). An example of an annotated document is provided in Section 3.3.

3.2 The spatial planning ontology

In this work the formal (entities from ontology) and informal model (text in natural language) of a spatial development plan are explicitly linked by mapping objects, classes and relationships through RDFa. By embedding presentation-invisible RDFa annotations in natural language text, the meaning of these textual phrases in the content of a document can be interpreted and transferred to the formal model.
Before the annotation process can start, a set of entities, grouped into a vocabulary, has to be chosen for the semantic annotation so that applications wishing to consume RDF data may understand the vocabularies or schemas used to describe data. The spatial planning ontology presented here consists of several classes and properties used for describing the spatial plan as a document as well as some of its content. It is mainly based on a Polish Regulation (Regulation, 2003) and the Polish Spatial Planning and Land Development Act (Spatial Planning, 2003), which define the basic elements to describe the main characteristics of a spatial development plan. The examples of semantic annotations presented in the article are based on Polish spatial development plans. For this reason, the spatial planning ontology is mainly adapted to the specific regulations concerning these documents. However, the concepts in the ontology are quite general and can also be applicable for other spatial plans in Europe. The ontology is published at http://wogis2.igig.up.wroc.pl/tbox/vocabs/mpzp_voc.
The main classes of the spatial planning ontology are:

1. *SpatialObject* describes a spatial feature.
2. *DocumentOfSpatPlan* describes the spatial development plan as a legal document.
3. *SpatialPlanningObject* describes a spatial object in the plan. Its subclasses are *PlanningZone*, the basic entity in a plan, and *ElaborationArea*, the area covered by the plan.
4. *SpatPlanLandUseDesignation* represents the land use classification in use in Poland. The minimal list of land use classes is specified in a Polish Regulation (Regulation, 2003) and the user can annotate the spatial planning document either with the property of the primary land use designation or the supplementary one.

The ontology was aligned with GeoSPARQL vocabulary. The *SpatialObject* class in the spatial planning ontology is equivalent to a *Feature* class in GeoSPARQL.

The characteristics of the spatial development plan are described by relevant properties of *DocumentOfSpatPlan* in the ontology, including *docName* (the official name for the plan) and *docNr* (the number of a resolution). Information about the author and the date of the adoption of a plan can be provided through the Dublin Core vocabulary. The *elaborationAreaGeometry* property represents the area covered by the plan. The example presented in 3.3 includes objects retrieved from the WFS services. Spatial planning objects (class *SpatialPlanningObject* - see Fig.3) in a plan are divided into zones (class *PlanningZone*) and the area covered by the plan (class *ElaborationArea*). Each zone can be annotated with a set of properties. These are for instance a symbol for the zone, which can be a literal, number or both. The domain of the symbol property is *SpatialPlanningObject*. For zones, there is a property – *zoneSymbol* – whose domain is PlanningZone. It has two subproperties: *zoneLiteralSymbol* and *zoneNumber*.

In spatial development plans, a specific land use is designated to each zone. In the ontology this is modelled by the *designationForZone*, *developmentRules* and *landUseDesignationDescr* properties. Designation can be primary (property *baseLandUseDesignation*) or supplementary (property *supplementaryLandUseDesignation*) – see Fig.4. Different land use classes are modelled in the ontology as instances of the *SpatPlanLandUseDesignation* class. Examples of land use classes from the land use classification are single-family housing, agricultural and service development.

[4] http://www.wymeditor.org/

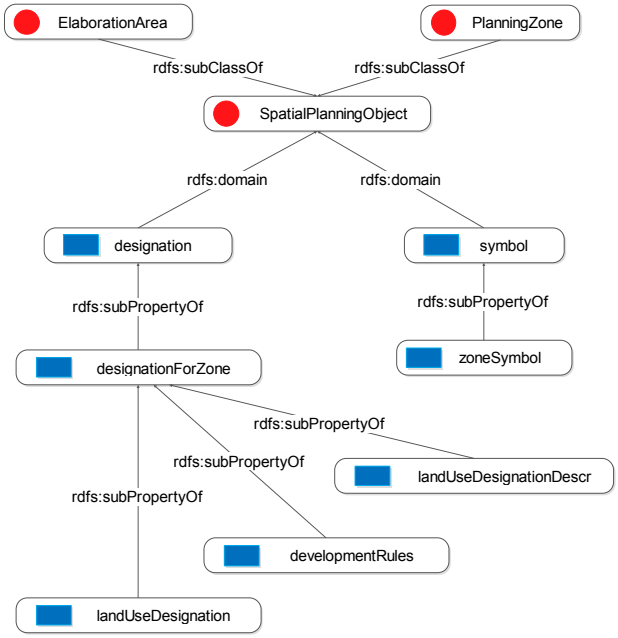

Figure 3. Graphic representation of subclasses and related properties and subproperties of the SpatialPlanningObject class.

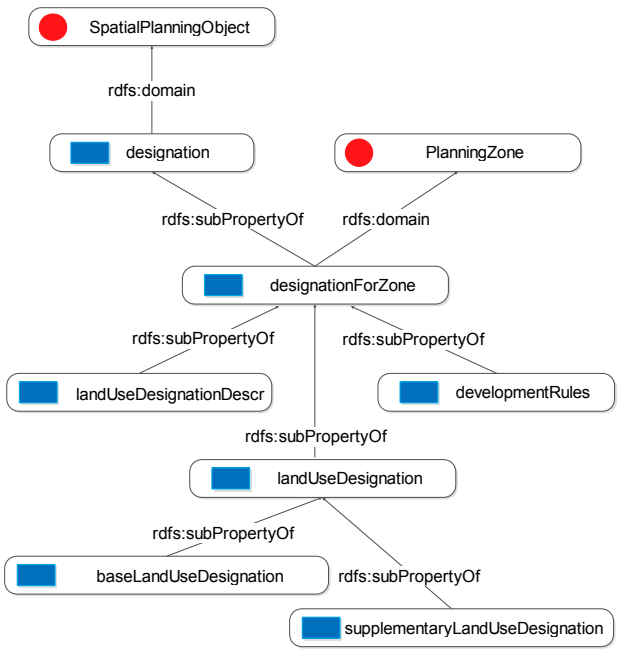

Figure 4. Domain and subproperties of designation

3.3 Registering temporal context

A spatial planning document, which is a legal act, has long-term validity. Changes in land use designations are gathered in subsequent legal acts in the form of resolutions. Approval of a resolution requires voting by a municipal council. The point in time from when a regulation is in force is a crucial aspect of a resolution. This determines the validity period of a given designation. In the simplest situation, when a resolution does not refer to individual regulations, the whole legal act becomes valid at the date of acceptance. Irrespective of the construction of the legal act, each regulation is valid for the period defined by the start and the end times.

Semantic annotations can be represented as an RDF graph made up of triples. Relationships between regulations and spatial objects are statements expressed as triples in such a graph. The temporal context can be referred to by each statement. For expressing such context, a named graph can be considered. Named graphs with default graphs are components of RDF datasets. This form of organizing data as named graphs forms the basis for building RDF repositories, called quad stores. Named graphs defined by the context for each triple allow data partitioning. In the case of a spatial planning document, each named graph represents those triples that are valid at some point in time or during some time period. The name of the named graph is specified as an Internationalized Resource Identifier (IRI). Such a name could be part of other triples, providing us with meta-information about conditions of validity of triples gathered in the named graph.

Named graphs could contain triples describing the state of land use regulations at a given point in time. Such a named graph could be considered as a snapshot of land use designations for some time slice. Alternatively, collecting regulations in the form of a named graph of triples identifies the subset of regulations valid during some period of time. The description of the state of legal regulations is contained in a distinct named graph. It allows comparison between different states stored in subsequent named graphs. The result of such comparisons could be recorded as a graph of inconsistencies, performing a similar role to the "diff" utility for listing differences between text documents. Expressions of these dissimilarities can be realized using Delta ontology (Berners-Lee and Connolly, 2004), suitable for such purposes. The differences between graphs could be the subject of analysis about dynamics and the pace of land use changes. The method of organizing data in RDF datasets with temporal context and the methods of discovering changes in such resources was presented by Łukowicz and Iwaniak (2015).

3.4 Example of a spatial development plan with embedded semantic metadata

A spatial development plan consists of a resolution header, a preamble, a resolution title and the content. The latter is divided, depending on the needs, into chapters, paragraphs, sections, letters and/or indents. The individual chapters contain separate provisions of general arrangements or of certain parts of the study area. Individual paragraphs or passages may include provisions for separate planning zones or groups. A sample paragraph (before annotation), describing provisions for a planning zone, is provided in Fig. 5. Fig. 6 shows the (presentation-only) HTML format for the corresponding paragraph.

§ 22. For zone 2-MN shall be determined:

1. Symbol of zone: 2-MN;
2. Surface area of zone: 0.98 ha;
3. Land use designation for zone: single-family residential housing.

Figure 5. A sample paragraph from a spatial development plan

```
<div class="paragraf" title="22">For zone 2-MN shall
be determined:</div>
<div class="punktowanie">
<span class="punkt" title="1">Symbol of zone: 2-
MN;</span>
<span class="punkt" title="2">Surface area of zone:
0.98 ha;</span>
<span class="punkt" title="3">Land use designation
for zone: single-family residential housing.</span>
</div>
```

Figure 6. HTML snippet for the sample paragraph in Fig.5

Semantic annotation experiments were conducted with a specially prepared web-based editor, WYMeditor. The technology used in the editor is based on HTML and Javasript and therefore extendible with new features. The flexible architecture of WYMeditor made it possible to extend it with the required features and HTML elements, allowing the creation of XHTML as recommended by the W3C for XHTML and RDFa documents.

The modified WYMeditor allows one to manually prepare a document in XHTML with 'div' and 'span' tags to specify objects and their corresponding properties. RDFa annotations can be added to existing HTML elements: a piece of text in natural language is enclosed by tags to identify a planning object, descriptional object or referenced entity. That is, such a tag indicates that this part of the text, written in natural language, refers to the specified object or entity and corresponds to an individual from an ontology.

An object is defined by the tag attribute @about, whose value is set to a URI. This creates a unique and identifiable RDF resource of the object. Similarly, the type of the object is defined by the @typeof attribute. The name of a class from the vocabulary is assigned to this attribute. This is an assertion of the individual to the class. Within the tag defining the individual, another tag describing its property may be inserted. For example, 'div' could contain 'span' tags with attributes indicating property types, such as *DataProperty* and *ObjectProperty*. The @property, @rel, @rev attributes define property names defined in the vocabulary, @content defines the literal value of a property and @resource (as well as @src, @href) define the URI of the resource (indicating the relationship). An example of a more comprehensive source HTML document is available at http://wogis2.igig.up.wroc.pl:8023/semeditor/res/mpzp-krokowa-srchtml.html and the corresponding annotated document is available at http://wogis2.igig.up.wroc.pl:8023/semeditor/res/mpzp-krokowa-rdfa.html.

The annotated document in Fig.7 contains built-in triples of concrete objects of specified types, with specific properties indicating their values. These are assertions of objects to classes and properties of objects which are derived from the vocabulary. The graphs extracted from multiple annotated spatial development plans can be stored in a triple-store and accessed via SPARQL queries. There are various ways in which the annotation process can be streamlined. For example, the semantic annotation process can be simplified by preparing XHTML templates containing the skeleton of the entire spatial planning document or repetitive parts of the document. The template may contain text for standard clauses of the Act in natural language, to which concrete designations may be added. It could also define the editorial structure of the text. It may also provide tags with attributes defining prototype objects with type assertions and with properties relating to the mandatory designations taken from the spatial planning ontology. The customized editor can be used to prepare such templates.

```
<div class="typeof mpzp_PlanningZone"
typeof="mpzp:PlanningZone"
about="http://example/gm_krokowa/rdf/mpzp_xv-160-
2011#Zone_2-MN">
<div class="paragraf" title="22">
<a class="property mpzp_zoneGeometry"
rel="mpzp:zoneGeometry"
href="http://example/geoserver/krokowa_sdi_2180/ows?
service=WFS&amp;amp;version=1.0.0&amp;amp;request=GetFeature
&amp;amp;typeName=krokowa_sdi_2180:tereny_jadm_razem_opi
s_sdi&amp;amp;FILTER=&amp;lt;Filter&amp;gt;&amp;lt;PropertyIsEqualTo
&amp;gt;&amp;lt;PropertyName&amp;gt;idpk&amp;lt;/PropertyName&amp;gt;&amp;lt
;Literal&amp;gt;274&amp;lt;/Literal&amp;gt;&amp;lt;/PropertyIsEqualT
o&amp;gt;&amp;lt;/Filter&amp;gt;">For zone </a>
<span class="property mpzp_zoneNumber"
property="mpzp:zoneNumber" content="2">2</span>-
<span class="property mpzp_zoneLiteralSymbol"
property="mpzp:zoneLiteralSymbol"
content="MN">MN</span> shall be determined:
</div>
<div class="punktowanie">
<span class="punkt" title="1">Symbol of zone:
<span class="property mpzp_zoneSymbol"
property="mpzp:zoneSymbol">2-MN</span>;
</span>
<span class="punkt" title="2">Surface area of zone:
<span class="property mpzp_spatObjProp"
property="mpzp:spatObjAreaHa"
content="0.98">0.98</span> ha;
</span>
<span class="punkt" title="3">Land use designation
for zone:
<span class="property mpzp_landUseDesignation"
rel="mpzp:landUseDesignation"
resource="http://example.pl/
tbox/vocabs/mpzp_voc#LandUseDesignation_MN"></span>
<span class="property mpzp_landUseDesignationDescr"
property="mpzp:landUseDesignationDescr"
title="single-family residential housing"
content="single-family residential housing">
single-family residential housing </span>.
</span>
</div>
</div>
```

Figure 7. Example of the spatial development plan with embedded semantic metadata

One way of preparing a template is to obtain an ontology describing the editorial structure of a legal act by a simple transformation from XML Schema (defining the structure of the legal Act) to OWL or RDFS. Such an ontology could be imported into the editor and used as a template. Eventually, the annotated document is exported to a format compatible with the XML schema of the legal act. To start with, an existing spatial planning document, such as an HTML page, can be imported into the editor, or the content of a document in MS Word or OpenDocument format can be pasted into the editor. The document can be annotated by any person independently from the authors. The customized editor in its current stage of development is available at http://wogis2.igig.up.wroc.pl:8023/semeditor/ latest/spatplan-editor.html. Currently, the editor allows for simple, hand-made annotations with a statically attached spatial planning ontology and selected entities from standard ontologies, such as Dublin Core and FOAF. For the future we plan efficiency improvements, such as importing any ontology; importing a list of zones, which are the subject of an annotation retrieved from WFS services; importing spatial document templates; and importing an existing document from the legal register. Other improvement plans include visualization of the structure of a source document while annotating; a validation of nesting elements; and logical schema validation.

4. SPATIAL PLAN PROCESSING BY AUTOMATED LEARNING APPROACH

By and large the spatial plans currently in actual use in Poland are textual documents. They have been, and continue to be, written by many different authors, using vastly different language styles, vocabulary, and are often later edited by other authors, or committees. Many of these documents do not even adhere to the guidelines set forth by legislation. A good question is: can these documents be processed automatically? As a minimum, this processing would add semantic markups, or annotations. Additionally, it could be part of a querying system, which would search the specific spatial plan, or a group of relevant plans, to answer questions posed by the user. The ultimate goal would be to create a complete structural representation of a planning document, expressing all its relevant and meaningful information in the form of a semantic network.

Such project has been undertaken and some preliminary results are described here. The required technologies come from the field of Natural Language Processing (NLP). Specifically, the process of interest is referred to as Information Extraction. This is the task of identifying predefined types of information in unrestricted texts (Piskorski et al., 2013). Its main objective is finding specific information in large bodies of textual data, while ignoring irrelevant information. If the text does not contain any 'interesting' information, nothing is to be extracted. Another part of Information Extraction is to provide a uniform structure for the extracted data, which allows querying it and comparing information from different documents.

Most technologies related to Natural Language Processing have been developed originally for the English language and the tools available for other languages are few. This is especially true for the Slavic languages, which are characterized by rich inflection and relatively free word order, making them harder to process.

Information Extraction is related to Information Retrieval. The task of Information Retrieval is to select a subset of documents relevant to a particular query from a collection of documents. It returns selected documents in ranked order, where the rank of a document corresponds to its relevance to the given query. Information Retrieval is universal, since it does not require specifying the structure for any domain, but the returned documents must be subsequently processed by other systems, often by humans. Such mechanism is used in search engines.

4.1 A prototype system for automatic processing of spatial plans

In this approach, the Information Extraction process only examines the textual part of the local land use plans, ignoring the information contained in the map part. The structure of a plan's textual part is typical for legal documents – it is divided into sections, paragraphs, and bullet items. Each plan covers some specific land area, often divided into smaller areas for different purposes. The area's purpose is one of the most important parts of the information specified by the plan, because location of a plot in an area determines what the owner will be able to build on it. A plan's text refers to an area using its symbol.

Based on the analysis of many different Polish local land use plans, and the general Information Extraction system architecture, a prototype system has been built for automatic processing of land use plans to convert them into a semantic network-type structural representation in the form of an RDF graph (Zacharski, 2014). The system uses classification, with the classification being automatically generated in a machine learning procedure. A number of plans are first manually annotated to indicate the key parts of their structure. The original and annotated plans then undergo a machine learning process. Supervised classifier training is used as machine learning model. The features used in the classification are functional expressions evaluated on morphologically annotated text. Once the system has been trained, the processing of the actual plans proceeds automatically. The annotated plans are then automatically converted into the semantic RDF graphs.

The goal for the prototype system, built as a proof of concept, was to extract the purpose of each area designated in the plan. The system was developed in C#, using additional libraries: the PDFBox.NET library to extract the original text from the PDF documents; the TaKiPi parser (Piasecki, 2007) to tokenize and tag text with lexical information; and dotNetRDF to generate the RDF graph image. The overall architecture of the system is presented in Fig. 8.

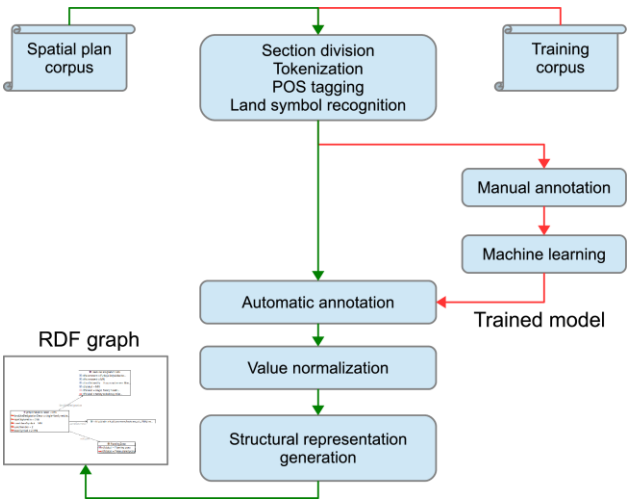

Figure 8. General architecture of the system extracting information from the local land use plans.

The results obtained in the experiments conducted with the system are as follows. An example of a local land use plan processed with the system (not part of the training batch) is presented in Fig. 9. (MPZP for Kowalewo from May 15th, 2014). The resulting RDF graph obtained for this plan is presented in Fig. 10. It should be noted that the main purpose of the specific areas are left as textual descriptions. This is due to the limited scope of the system and the ontology used in it. The parameters associated with the extracted purpose of the plan, such as the maximum percentage of built-up parts of an area, the minimum percentage of the biologically active part of the area, minimum and maximum heights of the buildings, etc., are extracted correctly from the plan document and represented in the resulting RDF graph. In this particular case, the minimum and maximum building heights have not been extracted correctly, since the specific wording was not recognized.

§11.Tereny oznaczone symbolami 8MN, 9MN, 10MN, 11MN, 12MN, 13MN, 14MN i 17MN przeznacza się na cel zabudowy mieszkaniowej jednorodzinnej; obowiązują następujące ustalenia:

1) wysokość zabudowy mieszkaniowej do dwóch kondygnacji nadziemnych oraz maksymalnie 9,0 m;

2) dopuszcza się podpiwniczenie budynków;
3) dachy budynków mieszkalnych o nachyleniu od 30° do 50°;
4) dopuszcza się budowę wolnostojących budynków garażowo-gospodarczych o architekturze nawiązującej do budynku mieszkalnego, wysokość budynków maksymalnie 5 m;
5) kąt nachylenia dachów budynków garażowo-gospodarczych od 15° do 45°;
6) dopuszcza się realizację infrastruktury technicznej związanej z podstawową funkcją terenu;
7) minimum 60% powierzchni działki budowlanej należy pozostawić w formie biologicznie czynnej (zieleń użytkowa lub ozdobna);
8) powierzchnia zabudowy do 40% powierzchni działki budowlanej;
9) minimalna powierzchnia nowo wydzielanych działek budowlanych 0,1 ha;
10) wskaźnik intensywności zabudowy od 0,1 do 0,8.

Figure 9. A fragment of the local land use plan for Kowalewo describing the purpose of some areas.

```
<rdf:RDF  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
          xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
          xmlns:mpzp="http://mpzp.pl/"
          xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-
ns#"
<mpzp:Plan rdf:about="&mpzp;KowalewoClear.spdf">
    <mpzp:describesLand
rdf:resource="http://mpzp.pl/KowalewoClear.spdf/8MN"/>
    <mpzp:describesLand
rdf:resource="http://mpzp.pl/KowalewoClear.spdf/9MN"/>
  </mpzp:Plan>
<mpzp:Land rdf:about="http://mpzp.pl/KowalewoClear.spdf/8MN">
    <mpzp:hasPurpose>
      zabudowy mieszkaniowej jednorodzinnej ;obowiązują
      następujące ustalenia :
    </mpzp:hasPurpose>
<mpzp:maximalBuildingArea>40%</mpzp:maximalBuildingArea>
<mpzp:maximalBiologicalActiveArea>60%</mpzp:maximalBiologicalAct
iveArea>
  </mpzp:Land>
  <mpzp:Land rdf:about="http://mpzp.pl/KowalewoClear.spdf/9MN">
    <mpzp:hasPurpose>
      zabudowy mieszkaniowej jednorodzinnej ; obowiązują
      następujące ustalenia :
    </mpzp:hasPurpose>
<mpzp:maximalBuildingArea>40%</mpzp:maximalBuildingArea>
<mpzp:maximalBiologicalActiveArea>60%</mpzp:maximalBiologicalAct
iveArea>
  </mpzp:Land>
</rdf:RDF>
```

Figure 10. The RDF graph (abbreviated) obtained from processing the local land use plan for Kowalewo.

Running the system on a larger number of local land use plans brought limited success. For example, in a batch of 45 documents, of which 5 were manually annotated and used for training, among the remaining 40 test documents only 5 were converted correctly (but the information extracted has not been 100%, as illustrated in the above example), while the other 35 documents did not produce any output at all. Increasing the training set to 10 documents, allowed 17 testing documents to be processed, however, almost half of them had partially incorrect results. This is due to the flexibility of the language and different wording used in the documents.

Further extending the training corpus gave increasingly better results on the testing set. The general conclusion, however, is that the outlook for this approach is limited. With the increase of the training set, the body of automatic processing rules grows, but there are significant interactions between these rules, and the number of incorrect results starts to grow as well. Another limiting factor lies in the natural-language documents themselves. For example, some plans turned out to have a description in the unexpected tabular format, while others had area symbols breaking the rules set out in the spatial planning legislation. Finally, there are ambiguous descriptions, for which different interpretations are possible.

5. EVALUATION OF THE EMBEDDED SEMANTIC METADATA APPROACH THROUGH A USE CASE

5.1 Evaluation of embedded semantic metadata approach for the use case in Poland

The preparation of a spatial development plan is a complex procedure. After adopting the spatial development plan through a resolution by the commune council, the plan is verified to conform to general legal requirements and published in the Official Journal of Voivodeship, which is a legal register. The spatial development plan life cycle is completed with the repeal or amendment of a resolution. The approach of embedded semantic metadata has benefits in various stages of the life cycle of a spatial development plan, from acceding to its preparation and to its amendment or repeal. The following actors have been identified for the evaluation discussion:

1. Authorities authorized to preparing spatial plan
2. Plan designers (planners) or design teams
3. Authorities responsible for nature protection
4. Authorities responsible for heritage and culture values protection
5. Transportation administration and management
6. Infrastructure services and management
7. Authorities responsible for nature protection
8. Authorities responsible for construction supervision or building inspection (issuing building permissions, authorization to for building usage, decisions for change of building usage etc.)
9. Citizens
10. Non-government organizations (NGO)
11. Developers
12. Real estate brokers
13. Lawyers representing various legal entities

Firstly, the RDFa editor could play a useful role in the preparation of the spatial development plan. It can be used to enrich the plain HTML document with a logical structure based on the objects and classes defined in the spatial planning ontology. The objects in this logical structure can be linked to other objects defined in the spatial planning ontology or to other resources, such as cadastral resources provided through a WFS or Linked Open Data resources.

In the first mode of handling spatial development plans, actors 1 and 2 create documents (they are the creators) and share them with other actors (consumers) in different forms and ways. The authority preparing the spatial development plan does not work in isolation, but is forced to cooperate with a design team led by a certified urban planner. The planner applies his/her competence, knowledge and experience to design a legally sound solution in compliance with urban design art and with reference to the rules of spatial order.

Annotated documents could be published and shared many times during the preparation procedure. For proper management of subsequent versions, each stage of the document development should be distinguished by a named graph. Creators and consumers can extract RDF graphs from HTML with embedded RDFa to build a quad/triple store repository, accessible via a SPARQL end-point. Data published in such a way can be retrieved and integrated with third party LOD resources using federated queries.

Consumers could use shared spatial development plans in active or passive mode. Passive mode implies that the document is

statically served as HTML with embedded RDFa resources via a web server. In passive mode, consumers search and browse the HTML spatial development plans like any other Web resources indexed by a web search engine. In addition, search engines with appropriate capabilities can interpret the embedded RDFa semantic metadata to provide more precise search results and to enrich search results with useful descriptions found in the semantic metadata. For example, to compare and analyse the spatial distribution of a specific land use class across more than one jurisdiction.

Other examples of passive use are the extraction of RDF graphs from annotated documents with the purpose of integrating these with other resources, and the exploration and integration of resources provided in the quad/triple store format through a SPARQL end-point, for example, to compare descriptions from an external resource, such as rights, restrictions and responsibilities recorded in the cadastre, to proposed land use classes in a spatial development plan.

Active consumers could enrich the raw or annotated documents with their own annotations, for example by classifying document objects or by linking them to quite different resources than the creators have. These could be shared, in a similar fashion as the creators who shared the documents they annotated. For instance, authorities responsible for environmental protection could annotate raw HTML with embedded RDFa by classifying planning objects according to an environmental ontology or by linking planning objects to spatial objects in an external register accessible via WFS. The relevant authorities can re-share the annotated documents. They can publish their own resources linked to objects in the spatial development plan, in the LOD manner.

6. DISCUSSION

Spatial planning documents are heterogeneous and can be complex. In order to meaningfully interpret multiple spatial planning documents it is essential to accurately identify objects, such as areas, sites or zones and to extract semantic information about these objects directly from the document. Embedding semantic metadata in spatial planning documents makes it possible to discover objects and associated semantics described in the document. It also has the add-on benefit of lowering the risk of metadata being out of synch with the documents they describe.

The explicit description of the content of the textual document as a number of objects classified according to a vocabulary and with relations to each other transforms the human readable document into a database. This opens up all the possibilities associated with databases, e.g. query and retrieval of specified information, and data processing and integration with data from other sources. The RDFa annotations make it possible to save relationships between objects in a spatial development plan and external resources. The annotated documents are discoverable by general-purpose crawlers and web search engines. Crawlers can detect relationships between different elements within the spatial planning documents, as well as to external sources. This allows users to analyse spatial development plans beyond the domain of spatial planning.

The approach presented in this article allows seamless integration with linked data resources. Publishing spatial development plans as linked data and integrating them with other resources enables their accessibility through distributed SPARQL queries or simply by browsing linked data resources. The question can be raised whether the same could not be achieved through HTML tags and class attributes? However, using HTML tags and class attributes to identify, describe and link objects has significant semantic limitations and constraints. A standard HTML document is annotated with tags to specify the presentation of the document. HTML tags also allow the creation of links from parts of the text to other documents. However, there is no way to describe the role of the link and therefore no possibility of interpreting the purpose of this linkage. Such semantic information can be added by adding classes to tags, but still, this is only understandable within a single document. Also, no information on how the class should be interpreted is available – i.e. a vocabulary description is not provided. Therefore, the HTML approach fails where a large number of documents are published by a variety and broader group of authors. Class interpretations should be consistent – at least among documents from the same domain, created for a specific purpose.

Semantic annotations created with the use of vocabularies and technologies, such as RDFa or microdata, provide the possibility to specify the roles of the selected parts of a document – transforming the free-text document into a structured document through the use of a triple model. This is essential for the process of web scraping (information retrieval), i.e. the process of extracting the structured data (or transforming parts of text into structured data) from web pages by autonomous web agents (spiders, crawlers). Of course, it would be perfect if all structured information were stored in specialized and standardized services, where data is available through web interfaces, but creating and maintaining such web services would require a substantial amount of work.

The question can also be raised whether it is necessary to create a formal model representation of the free format text document? Could the same not be achieved through full-text search methods available in general-purpose web search engines? RDFa annotated HTML documents can be seen as a bridge between free format text documents and structured information stores. Annotating spatial planning documents does not require specialized technical skills and can therefore be performed by urban planners. Annotations created with the use of known technologies and vocabularies allow more efficient and effective extraction of data (than for free format text data). In addition, the RDF approach makes data available in both human and machine-readable form in a single format, simplifying synchronization between different forms.

An obvious limitation of the approach presented in this article is the considerable amount of manual work required to annotate relevant sections of the spatial development plan. A number of ways to streamline this work have been presented and plans to improve the efficiency of the editor were discussed. While the editor needs to be improved, explicitly identifying objects in a spatial development plan is not necessarily a drawback. Related research suggests that explicit annotation of the structure of the document can improve the quality of the document (Coetzee et al., 2011).

In future, the work in this project will be extended with web crawlers that systematically review and analyze annotated spatial planning documents published on the Web. Data obtained from this process will be stored in an RDF repository. The web crawlers will be based on the LDSpider framework, which allows searching documents annotated with RDFa,

extracting RDF triples from them and storing them in repository. The crawler can extract triples based on selected ontologies - in this case the ontology used for the annotation of the spatial planning documents.

7. CONCLUSIONS

The solution presented in this article is a first step towards the automated interpretation and integration of heterogeneous spatial planning documents and the integration of spatial planning documents with other data sources.

In this article we showed how textual and human readable documents can be enriched with a description of the semantic content of the document. A spatial planning ontology was used to formally describe objects in a spatial development plan. Through the use of RDFa annotations references between objects described in the textual part and spatial objects in the graphical part of a spatial development plan were established. References may also point to external resources in separate registers or to any information published on the Web. These references represent well-defined logical connections according to the fundamental principles of linked data. Thus, the solution presented in this article provides a means to integrate spatial planning documents into the linked data environment, which is gaining significance in the field of geographic information.

The formal description of objects in spatial development plans according to a spatial planning ontology makes it possible to integrate data on a wider scale than the local one. Documents with embedded semantic metadata are discoverable and retrievable by browsers enabled with semantic web technology. At the same time, the description of an annotated document in the form of a graph is flexible – it does not impose a uniform scheme or hierarchy of Web resources on the document.

A limitation of the approach presented in this article is the semi-automated process of annotating a spatial development plan. Results of an experiment following an automated learning approach to process spatial plans show limited success, but a number of ways to streamline this work were proposed and plans to improve the efficiency of the editor were discussed.

The World Wide Web has evolved into a global space, not only connected through hypertext documents but also through data. The approach presented in this article follows this evolution and provides new opportunities to enrich the Web of Data with spatial planning information.

As we can see, the needs of multipurpose applications fulfilling needs of different users, from scientific to social domains, demand specific tools. Linked Open Data, as structure for publishing content of spatial planning documents is the best way for providing technologically neutral resources, useful for different purposes, from scientific to social.

ACKNOWLEDGEMENTS

This work was supported by the National Science Centre under Grants number DEC-2011/03/N/HS4/03819, DEC-2012/05/B/H/HS4/04197, DEC-2012/05/N/HS4/00642.

REFERENCES

Berners-Lee, T., Hendler, J., Lassila, O., 2001. The semantic web. *Scientific American,* 29-37.

Berners-Lee, T. and Connolly, D., 2004. Delta: an ontology for the distribution of differences between RDF graphs, http://www.w3.org/DesignIssues/Diff

Coetzee, S., Cox, S. and Herring, J., 2011. Configuration management of a system of interdependent standards. *7th International Conference on Standardization and Information Technology (SIIT)*, Berlin, Germany, 28-30 September 2011, pp 47-58.

Heath, T., Bizer, C., 2011. Linked data: evolving the web into a global data space Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool.

Hjelmager, J., Moellering, H, Cooper, A., Delgado, T., Rajabifard, A., Rapant, P., Danko, D., Huet, M., Lauren, D., Aalders, H., Iwaniak, A., Abad, P., Düren, U. & Martynenko, A., 2008. An initial formal model for spatial data infrastructures, International Journal of Geographical Information Science, 1295–1309.

Iwaniak, A., Kaczmarek, I., Kubik, T., Łukowicz, J., Paluszyński, W., Kourie, D.G., Cooper, A.K., Coetzee, S., 2011. An Intelligent Geoportal for Spatial Planning, *25th International Cartographic Conference*, Paris, pp. CO-029

Janowicz K., and Hitzler P., .2012. The Digital Earth as Knowledge Engine. *Semantic Web Journal*, 1–0 1 IOS Press

Kubik, T., Iwaniak, A., 2010. Building and maintaining metadata repositories with the aid of ontology tools and technologies, *GSDI-12 World Conference*

Lukowicz J., Iwaniak A., (2015) From Snapshots to Processes Description – Spatiotemporal Modelling using RDF Datasets and Formal Ontologies. *18th AGILE International Conference on Geographic Information Science*, 9-12 June 2015 - Lisbon, Portugal, http://www.agile-online.org/ Conference_Paper/cds/ agile_2015/shortpapers/61/61_Paper_in_PDF.pdf

Yu L. and Liu Y., 2015. Using Linked Data in a heterogeneous Sensor Web: challenges, experiments and lessons learned. (2015). *International Journal of Digital Earth*, 2015 Vol. 8, No. 1, 17–37, http://dx.doi.org/10.1080/17538947.2013.839007

Piasecki M., 2007. Polish Tagger TaKIPI: Rule Based Construction and Optimisation. Task Quarterly, pp. 151–167

Piskorski J., Yangarber R., 2013. Multilingual Information Extraction and Summarization, *chapter Information Extraction: Past, Present and Future.* Springer

Plan4all D3.2.2, 2010. Plan4all Metadata Profile - Final version, Retrieved December 01, 2014, from Plan4all: http://portal.plan4all.eu/simplecms/?menuID=29&articleID=50 &action=article&presenter=ArticleDetail

Plan4all D4.2, 2010. Conceptual Data Models for Selected Themes, Retrieved December 01, 2014, from Plan4all: http://portal.plan4all.eu/simplecms/?menuID=29&articleID=61 &action=article&presenter=ArticleDetail

Regulation of 26 August 2003 on the required scope of the project of zoning plan (Journal of Laws No. 164, item 1587)

Spatial Planning and Land Development Act of 27 March 2003 (Journal of Laws No. 80, item 717)

Vilches-Blázquez L.M., Villazón-Terrazas B., Corcho O., Gómez-Pérez A., 2014. Integrating geographical information in the Linked Digital Earth, *International Journal of Digital Earth*, 7:7, 554-575, DOI: 10.1080/17538947.2013.783127

Zacharski D., 2014. Translation of written documents in Polish to a structural representation. M.S. Thesis, Wroclaw University of Technology.