A Semantic Approach to Describe Geospatial Resources

Sidney Roberto de Sousa

Institute of Computing, University of Campinas, 13083-970, Campinas, SP, Brazil
sidney@lis.ic.unicamp.br

Abstract. Geographic information systems (GIS) are increasingly using geospatial data from the Web to produce geographic information. One big challenge is to find the relevant data, which often is based on keywords or even file names. However, these approaches lack semantics. Thus, it is necessary to provide mechanisms to prepare data to help retrieval of semantically relevant data. This paper proposes an approach to attack this problem. This approach is based on semantic annotations that use geographic metadata and ontologies to describe heterogeneous geospatial data. Semantic annotations are RDF/XML files that rely on a FGDC metadata schema, filled with appropriate ontology terms, and stored in a XML database. The proposal is illustrated by a case study of semantic annotations of agricultural resources, using domain ontologies.

1 Introduction

The Web became an immense repository of geospatial data in different geographic formats like remote sensing images, maps, sensor data temporal series, textual data files, among others [1, 2]. The retrieval of these data requires special attention due the geographic distribution of the sources and the heterogeneity of the data. Geographic metadata standards and geospatial information portals were created as an initiative to attack this problem. In these portals, users can create their own queries using keywords and metadata fields from some metadata schema such as ISO 19115 and FGDC Metadata. These metadata fields are often filled with natural language text, which can cause ambiguities, while keywords can restrict the result of the queries if different terminology is used or if terms are homonymous [3].

One solution to overcome these problems is the use of domain ontologies - as can be seen in [4] - to identify and associate common concepts. Ontologies are frequently used to explain knowledge about some domain of interest. In the geographic domain, an ontology must have terms and concepts about useful issues to describe geospatial resources, for instance, spatial references, time periods, geographic formats details, and other kinds of meta-information that may improve the retrieval of geospatial information.

C.A. Heuser and G. Pernul (Eds.): ER 2009 Workshops, LNCS 5833, pp. 327–336, 2009.

The World Wide Web Consortium (W3C) proposed the Resource Description Framework (RDF) to describe resources available in the Web as an initiative for providing semantic interoperability. RDF identifies resources using their URIs and describes them using statements. A statement is a triple <*subject, predicate, object*>. From the geospatial point of view, a subject is a geospatial resource, a predicate is a metadata field of this resource, and an object is the value filling the metadata field. Applying this model in a way so ontologies could be included, the object can be an ontology term that semantically associates the metadata field content to some appropriate concept.

Based on this approach, this paper discusses the use of semantic annotations to describe geospatial data, extending the work of [2] to cover implementation aspects. This work defines a semantic annotation as a set of RDF triples, where each triple is basically composed of a FGDC metadata schema, where each metadata field is filled with appropriate terms from domain ontologies. The annotations are stored in an XML database, where they can be retrieved using XQuery and XPath statements.

The rest of this paper is organized as follows. Section 2 describes the approach for semantic annotation of geospatial resources presented in this paper. Section 3 explains how annotations are stored in an XML database. Section 4 shows how the presented approach is applied in a case study of semantic annotations of agricultural resources. Section 5 describes related work. Finally, section 6 contains conclusions and ongoing work.

2 An Approach for Representing Semantic Annotations

A semantic annotation of a geospatial resource must provide semantic descriptions about geographic characteristics of this resource. Such characteristics are structurally organized using geographic metadata standards. The role of ontologies in this scenario is to enhance the annotations, providing appropriate terminology. This section describes the representation of the semantic annotations in RDF/XML format, detailing geographic metadata schema, and ontology concepts.

2.1 Geographic Metadata Schema

Metadata can be considered as data about other data. Their principal role is to add important information to a resource so that ambiguities can be avoided and the retrieval of the resource can be done in an easier way. Absence of metadata may lead to unreliability and re-work when it comes to interoperability among distinct systems, hampering data exchange and integration [5]. Geographic metadata describe geospatial resources, enhancing them with useful information such as reference system used, producer identification, and location information.

Use of geographic metadata is strongly disseminated by geographic catalogs, such as GeoNetwork[1], which use geographic metadata standards. ISO 19115 is a

[1] http://sourceforge.net/projects/geonetwork Accessed in March 30th, 2009.

proprietary standard of geographic metadata, developed by the ISO Committee. It has a UML based structure, where each metadata element is defined in context of a class and is characterized by a *name*, *definition*, *obligation*, *multiplicity*, *data type*, and a *domain*. This standard has a minimal set of elements which is defined for the most important information needed to describe some resource, called *core data*. It is possible to extend this set of elements to serve special needs [6].

The Federal Geographic Data Committee Metadata (FGDC Metadata) is an open standard which defines some particularities needed to catalog and publish geographic meta-information. It provides knowledge about the kind of the resource, indicating whether it meets the users expectation, and where/how to find it. Use of a specific section or element is either mandatory or optional [7].

2.2 Using Ontology Terms

In geographic catalogs, metadata fields are filled with natural language text, which most times can lead to ambiguities or bad understanding. Despite the structure and semantics that metadata can provide, the content of the fields may not be able to avoid this and other kinds of problems [3]. The use of ontology terms guarantees unique meaning, associating metadata fields to concepts that semantically represent their content. Ontologies also provide a hierarchical structure that helps to understand their concepts. Figure 1 shows the solution for the example seen, using terms of NASA SWEET Numerics ontology[2]. It indicates that the *Graph* term is a 2D distribution.

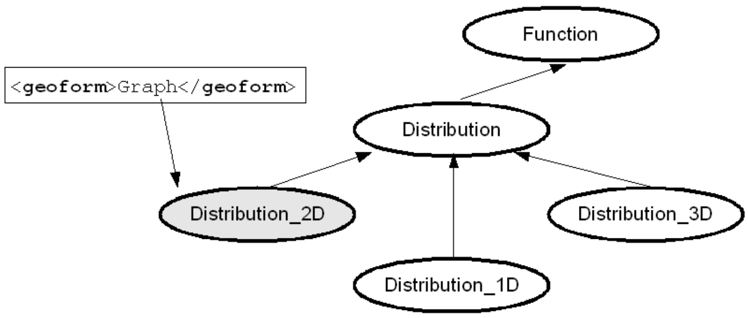

Fig. 1. Use of an ontology term to represent a metadata field

2.3 Representation in RDF

Once a metadata schema is chosen, it is possible to use RDF to semantically describe a resource. Figure 2 illustrates a possible representation in RDF/XML (without the use of ontology terms) of a graph that shows the evolution of some phenomenon with time, as measured per seasons. It uses metadata fields

[2] http://sweet.jpl.nasa.gov/ontology/ Accessed in March 31st, 2009.

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:fgdc="http://www.fgdc.gov/metadata/fgdc-std-001-1998.xsd#">

   <rdf:Description
    rdf:about="http://www.lis.ic.unicamp.br/efarms/NDVI_graphs/graph01.jpg">
      <fgdc:citeinfo rdf:parseType="Resource">
         <fgdc:origin>eFarms</fgdc:origin>
         <fgdc:pubdate>20080526</fgdc:pubdate>
         <fgdc:title>NDVI Graph</fgdc:title>
         <fgdc:edition>Digital image version</fgdc:edition>
         <fgdc:geoform>Graph</fgdc:geoform>
         <fgdc:serinfo rdf:parseType="Resource">
            <fgdc:sername>NDVI graphs set</fgdc:sername>
            <fgdc:issue>NDVI calculus of rural areas</fgdc:issue>
         </fgdc:serinfo>
         <fgdc:pubinfo rdf:parseType="Resource">
            <fgdc:pubplace>Campinas - SP</fgdc:pubplace>
            <fgdc:publish>LIS, IC-UNICAMP</fgdc:publish>
         </fgdc:pubinfo>
      </fgdc:citeinfo>
   </rdf:Description>
</rdf:RDF>
```

Fig. 2. Representation in RDF of metadata for a graph, using fields from the FGDC metadata standard

```
<fgdc:origin rdf:parseType="Resource">
      <rdfs:comment>eFarms</rdfs:comment>
      <rdf:type rdf:resource="http://sweet.jpl.nasa.gov/1.1/data.owl#Project"/>
</fgdc:origin>
```

Fig. 3. Adding an ontology term to *fgdc:origin* element

from FGDC. The *rdf:Description* element indicates a description of some Web resource. The *rdf:about* attribute identifies the resource using its URI. After this, come the metadata fields, using the following rule: if an element is composed of one or more elements, it must have a *rdf:parseType="Resource"* attribute indicating that it contains other elements.

Now, imagine that we want to add ontology terms to the metadata fields, but we want to preserve the natural language content for future use in a publication interface: how to do this, using RDF? One way to solve this problem is to keep the natural language text as a human readable description of the metadata field's content, using the property *rdfs:comment* from RDF Schema (RDFS), an extension to RDF for defining application-specific classes and properties[3]. In addition, we can specify that the content of the metadata field is an instance of an ontology class (the ontology term), using the property *rdf:type*. Figure 3 shows this solution. In this example, the field *origin* contains a human readable description that says that the resource was originated by "eFarms" and a reference to the class *Project* that specifies that the originator of the resource is an instance of this class. Thus, we want to say that "the resource was originated by a project called eFarms".

[3] http://www.w3.org/TR/rdf-schema/ Accessed in June 23rd, 2009.

3 Storing RDF Annotations

RDF can be represented by more human-readable languages like Notation3[4] (N3) or by more structured languages like RDF/XML, which is the most used one. An essential characteristic of a good quality geographic metadata standard is that it should be XML compatible. Both FGDC Metadata and ISO 19115 have this feature, as well as metadata standards from other domains such as Dublin Core [8] and e-GMS [9]. These facts lead towards the use of XML databases to store RDF/XML.

An XML database is a data persistence software that allows storage of data in XML format, generally mapping these data from XML to some storage format, which can be a relational database or even other XML documents [10]. Queries over a XML database are generally executed using XPath or XQuery statements. It is possible to retrieve RDF/XML data using XQuery, once this language was designed to query XML data not just from XML files, but anything that is structured in XML.

Both XPath and XQuery allow retrieval of full XML-based documents or subtrees of these, using their DOM trees[5]. If we know the schema of an annotation of interest, we can retrieve the full annotation or parts of these. For instance, if someone wanted to know who originated the NDVI graph of the previous example, he could retrieve this information using an XPath statement (*/rdf:RDF/rdf:Description/fgdc:citeinfo/fgdc:origin*).

Another solution for storing and querying RDF is to use some framework for these purposes, like Sesame [11] and Jena [12]. These frameworks play the role of a layer that manage persistent storage of RDF in files or relational databases and provide queries over RDF in SPARQL or in other specific languages. Moreover, such frameworks provide reading and writing of RDF in different notation languages.

4 A Case Study: Semantic Annotation of Agricultural Resources

We propose an architecture for semantic annotation of agricultural geospatial data to illustrate the approach of this paper, taking as example a NDVI graph. Normalized Difference Vegetation Index (NDVI) is a numerical indicator used to analyze whether some region of interest has live green vegetation or not. Using this index, it is possible to verify some aspects like density of vegetation or crops in some area of interest. A NDVI graph is a 2D distribution containing return values of the NDVI function in a certain time period, where the y axis is the NDVI index and the x axis is a date, thus characterizing a time series. This kind of annotation can be useful for activities like crop management and monitoring [2].

[4] http://www.w3.org/DesignIssues/Notation3.html Accessed in June 23rd, 2009.

[5] The XML DOM (Document Object Model) defines a standard way for accessing and manipulating documents compatible to XML, presenting them as a tree structure where elements, attributes, and text are nodes.

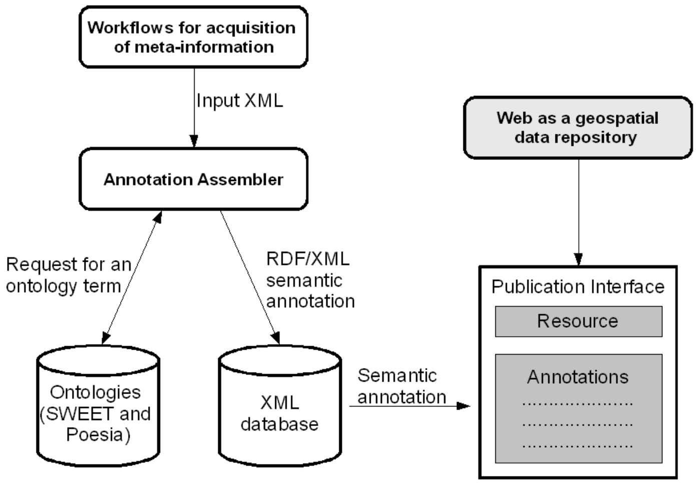

Fig. 4. Proposed architecture

Figure 4 illustrates the proposed architecture. It is composed by the following modules:

– ***Workflows for acquisition of meta-information:*** Acquisition of meta-information depends on the kind of resource. Here, the acquisition of such information is performed by specific workflows described in [2], where a specific workflow is activated for each kind of geospatial data. For a NDVI graph, crop identification is done comparing the curve of the graph to other existing curves where crops were already identified. A more detailed explanation about this work is given in section 5. The meta-information is organized in a set of FGDC metadata, encapsulated in a simple XML file, and submitted to the annotation assembler;
– ***Annotation assembler:*** This module receives the meta-information submitted by a specific workflow, which contains a specific set of FGDC metadata fields filled with natural language text. An ontology term is associated to each metadata field using the approach presented in section 2.3. The choice of the ontology term is done by a mechanism that queries the ontology base for URI of terms;
– ***Ontology base:*** The ontology base is composed by geospatial and agricultural ontologies. NASA SWEET ontologies provide terms about issues in various domains like geography, physics, chemistry, among others. Poesia Agricultural Zoning ontology [13] provides terms about crops and Brazilian locations. Some of these ontologies were extended to attend specific needs like terms about crop production;

Metadata element	Description	Short name	Metadata Schema	Obligation	Ontology used
Citation Information	Reference to be used for the data set	citeinfo	FGDC	Yes	Nasa SWEET
Indirect Spatial Reference	Means which locations are referenced	indspref	FGDC	Yes	POESIA Agricultural Zoning
Horizontal Coordinate System Definition	System which linear/angular quantities are measured and assigned to the position that a point occupies	horizsys	FGDC	Yes	Nasa SWEET
Time Period Information	Time period for which the data set corresponds	timeperd	FGDC	Yes	Nasa SWEET
Digital Transfer Information	Description of the form of the data to be distributed	digtinfo	FGDC	Yes	Nasa SWEET
Crop Identification	Information about Identification of crops	cropid	Agricultural Extension	No	POESIA Agricultural Zoning
Soil Identification	Information about Identification of soils	soilid	Agricultural Extension	No	POESIA Agricultural Zoning
Productivity Identification	Information about productivity issues	productivity	Agricultural Extension	No	POESIA Agricultural Zoning

Fig. 5. Composition of a semantic annotation of a NDVI graph

– ***XML database:*** After an annotation in RDF/XML is created, it is stored in a XML database from where it can be retrieved using XQuery statements;
– ***Publication interface:*** A Web interface where agricultural researchers can see the Web resource and its semantic annotation.

Figure 5 shows a table that explains the contents of the semantic annotation of a NDVI graph. In order to cover agricultural needs, a agriculture metadata set was created containing elements about crops, soil, and productivity issues (Agricultural Extension).The first column shows the metadata elements used to describe a NDVI graph. Each FGDC element shown in the table is composed by other specific elements, which were abstracted in the table. For details about the acquisition of agricultural meta-information, see [2]. The second column shows a brief description of each element. The third column shows the short name of each element, defined in their respective XML Schemas. The fourth column shows the metadata schema to which each element belongs. The fifth column specifies whether the presence of the element is mandatory or not. Finally, the last column shows the ontologies used to describe each metadata element.

5 Related Work

One of the aims of this work is to provide implementation support to the work of [2], which proposes a framework for semantic annotation of agricultural resources. In that work, each different geospatial resource has a specific workflow for acquisition of spatial and crops meta-information, linkage to ontologies terms, and production and publication of semantic annotations. Our architecture provides the infrastructure needed to associate semantics to annotations, via the linkage module.

There are several research initiatives related to the work reported in this paper. One such trend concerns semantic interoperability in GIS, dealing with problems in data exchange and retrieval. There are some efforts to provide interoperability among metadata standards, as can be seen in [5, 7]. Use of ontologies to deal with interoperability problems in the geospatial domain is discussed in [1, 4, 14, 15].

Another area is representation of information. RDF is being widely used for representing geographic meta-information. In [16], RDF is used to define a catalog of geographic resources from various Web sites. Córcoles and Gonzáles [17] propose an approach for providing queries over spatial XML resources with different schemas using a unique interface, where the resources are integrated using RDF.

Due to the conventional use of XML to represent meta-information, some works have used XML databases to store metadata. In [18], a XML database is used to store metadata in a prototype of a digital library system, which provides queries over metadata from art pieces. The use of XML databases for the management of metadata in the MPEG-7[6] format is discussed in [19], where a survey concerning XML database solutions for this issue was done. A schema-independent XML database used to store metadata about scientific resources is presented in [20].

6 Conclusions and Ongoing Work

Geographic distribution and heterogeneity are issues that hamper the retrieval of geospatial data. Geographic metadata standards were created to solve these problems, but filling metadata fields with natural language text can cause ambiguities. To attack this problem, this paper discussed an approach based on RDF, geographic metadata and ontologies to describe geospatial resources, bringing together Semantic Web and geographic standards technologies. Moreover, it discussed the storage of semantic annotations in XML databases, considering the RDF/XML notation.

Based on this approach, a mechanism is being implemented that chooses and ranks appropriate ontology terms to the metadata fields. At the moment, the choice of terms is done over specific ontologies (Nasa SWEET and Poesia Agricultural Zoning), but the mechanism is intended to be ontology-independent, so that it can choose appropriate ontologies and hence appropriate terms to fill the fields. Once an annotation in RDF is created, the mechanism stores it in a XML database. However, it is intended to use a RDF framework for storing and querying the semantic annotations and so make a comparison about the two approaches.

Acknowledgments

The authors thank FAPESP-Microsoft Research Virtual Institute (eFarms project), CNPq and CAPES for the financial support for this work.

[6] A standard for the description of multimedia content.

References

[1] Fonseca, F., Rodriguez, A.: From Geo-Pragmatics to Derivation Ontologies: new Directions for the GeoSpatial Semantic Web. Transactions in GIS 11(3), 313–316 (2007)

[2] Macário, C.G.N., Medeiros, C.B.: A Framework for Semantic Annotation of Geospatial Data for Agriculture. Int. J. Metadata, Semantics and Ontology - Special Issue on "Agricultural Metadata and Semantics" (2008) (accepted for publication)

[3] Klien, E., Lutz, M.: The Role of Spatial Relations in Automating the Semantic Annotation of Geodata. In: Cohn, A.G., Mark, D.M. (eds.) COSIT 2005. LNCS, vol. 3693, pp. 133–148. Springer, Heidelberg (2005)

[4] Klien, E., Einspanier, U., Lutz, M., Hübner, S.: An Architecture for Ontology-Based Discovery and Retrieval of Geographic Information. In: Proceedings of the 7th Conference on Geographic Information Science (AGILE 2004), Heraklion, Greece, pp. 179–188 (2004)

[5] Nogueras-Iso, J., Zarazaga-Soria, F.J., Lacasta, J., Bejar, R., Muro-Medrano, P.R.: Metadata Standard Interoperability: Application in the Geographic Information Domain. Computers, environment and urban systems 28(6), 611–634 (2003)

[6] Karschnick, O., Kruse, F., Topker, S., Riegel, T., Eichler, M., Behrens, S.: The UDK and ISO 19115 Standard. In: Proceedings of the 17th International Conference Informatics for Environmental Protection EnviroInfo. (2003)

[7] Chandler, A., Foley, D.: Mapping and Converting Essential Federal Geographic Data Committee (FGDC) Metadata into MARC21 and Dublin Core: Towards an Alternative to the FGDC Clearinghouse. D-Lib Magazine 6 (2000)

[8] Weibel, S., Kunze, J., Lagose, C., Wolf, M.: RFC2413: Dublin Core Metadata for Resource Discovery (1998)

[9] Alasem, A.: An Overview of e-Government Metadata Standards and Initiatives based on Dublin Core. Electronic Journal of e-Government 7, 1–10 (2009)

[10] XML:DB Initiative: Frequently Asked Questions About XML:DB, `http://xmldb-org.sourceforge.net/faqs.html` (accessed in April 4th, 2009)

[11] Broekstra, J., Kampman, A., van Harmelen, F.: Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema, pp. 54–68. Springer, Heidelberg (2002)

[12] Wilkinson, K., Sayers, C., Kuno, H., Reynolds, D.: Efficient RDF Storage and Retrieval in Jena2. Exploiting Hyperlinks 349, 35–43 (2003)

[13] Fileto, R., Liu, L., Pu, C., Assad, E.D., Medeiros, C.B.: POESIA: An ontological workflow approach for composing Web services in agriculture. The VLDB Journal 12(4), 352–367 (2003)

[14] Visser, U., Stuckenschmidt, H., Schuster, G., Vögele, T.: Ontologies for geographic information processing. Comput. Geosci. 28(1), 103–117 (2002)

[15] Fonseca, F.T., Egenhofer, M.J.: Ontology-driven geographic information systems. In: GIS 1999: Proceedings of the 7th ACM international symposium on Advances in geographic information systems, pp. 14–19. ACM, New York (1999)

[16] Córcoles, J.E., González, P., López-Jaquero, V.: Integration of Spatial XML Documents with RDF. In: Cueva Lovelle, J.M., Rodríguez, B.M.G., Gayo, J.E.L., del Ruiz, M.P.P., Aguilar, L.J. (eds.) ICWE 2003. LNCS, vol. 2722, pp. 407–410. Springer, Heidelberg (2003)

[17] Córcoles, J.E., González, P.: Using RDF to Query Spatial XML. In: Koch, N., Fraternali, P., Wirsing, M. (eds.) ICWE 2004. LNCS, vol. 3140, pp. 316–329. Springer, Heidelberg (2004)

[18] Baru, C., Chu, V., Gupta, A., Ludäscher, B., Marciano, R., Papakonstantinou, Y., Velikhov, P.: XML-based information mediation for digital libraries. In: DL 1999: Proceedings of the fourth ACM conference on Digital libraries, pp. 214–215. ACM, New York (1999)
[19] Westermann, U., Klas, W.: An analysis of XML database solutions for the management of MPEG-7 media descriptions. ACM Comput. Surv. 35(4), 331–373 (2003)
[20] Jones, M.B., Berkley, C., Bojilova, J., Schildhauer, M.: Managing Scientific Metadata. IEEE Internet Computing 5(5), 59–68 (2001)