Management of Semantic Annotations of Data on Web for Agricultural Applications

Sidney R. de Sousa[1], **Claudia B. Medeiros**[1]

[1]Institute of Computing – University of Campinas (Unicamp)
PO Box 6176 – 13084-971 – Campinas/SP – Brazil

sidney@lis.ic.unicamp.br, cmbm@ic.unicamp.br

Graduate Program: Master degree in Computer Sciences
Advisor: Claudia B. Medeiros
Admission: february 2008
Expected Conclusion: march 2010
Concluded Stages: qualification
Keywords: Semantic annotations, geospatial data, ontologies, RDF, XML databases.

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 work proposes an approach to attack this problem. This approach is based on semantic annotations that use geographic metadata and ontologies for describing heterogeneous geospatial data. Semantic annotations are RDF/XML files that rely on a FGDC metadata schema, filled with appropriate ontologies terms, and stored in a XML database. The proposal takes as case study semantic annotations of agricultural resources, using domain ontologies.*

1. Introduction and Motivation

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 [Fonseca and Rodriguez 2007, Macário and Medeiros 2008]. 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 for attacking 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 [Klien and Lutz 2005].

One solution to overcome these problems is the use of domain ontologies - as can be seen in [Klien et al. 2004] - 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.

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 URI's 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 work discusses the use of semantic annotations to describe geospatial data, extending the work of [Macário and Medeiros 2009] 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, and 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.

2. Related 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 in a way so ambiguities can be avoided and the retrieval of this 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 [Nogueras-Iso et al. 2003].

Geographic metadata describe geospatial resources, enhancing them with useful information for the geographic domain, such as reference system used, producer identification, and location information. A useful geographic metadata schema must have a readable structure and a grammar that indicates the semantics and the structure of the elements. For interoperability reasons, its format should be compatible to XML.

Use of geographic metadata is strongly disseminated by geographic catalogs, such as

GeoNetwork[1], which makes use of geographic metadata standards. One of such standards is the ISO 19115, which is a proprietary standard of geographic metadata, developed by the ISO Committee. This standard has a minimal set of elements which is defined for the most important information needed to describe some resource, called *core data*; however, it is possible to extend this data model to serve special needs [Karschnick et al. 2003]. Another very important standard is the Federal Geographic Data Committee Metadata (FGDC Metadata), 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 user's expectation, and where/how to find it.

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 [Klien and Lutz 2005]. Use of ontology terms guarantee unique meaning, associating metadata fields to concepts that semantically represent their content. Ontologies also provide a hierarchical structure that helps to understand their concepts.

3. Characterization of the Contribution

This work intends to provide semantic interoperability among geospatial resources describing them with semantic annotations, where each semantic annotation is a set of RDF statements, and a RDF statement is a triple <*subject, predicate, object*>. In this work, differently of the majority of works that intend to provide description of geospatial data, *object* is an ontology term that makes reference to some issue in the geographic domain. To validate our approach, we consider as case study resources that are useful for agricultural applications.

3.1. Representing Semantic Annotations in RDF

Figure 1 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 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[2]. 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 2 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".

[1]http://sourceforge.net/projects/geonetwork. Accessed in May 24th, 2009.
[2]http://www.w3.org/TR/rdf-schema/. Accessed in June 23rd, 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>
```

Figure 1. Semantic annotation in RDF, without the use of ontology terms.

```
<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>
```

Figure 2. Adding an ontology term to *fgdc:origin* element.

4. Storing RDF annotations

RDF can be represented by more human-readable languages like Notation3[3] (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 [Weibel et al. 1998] and e-GMS [Alasem 2009]. 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 [XML:DB Initiative]. 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[4]. 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 graph of the previous example, he could retrieve this information using an XPath statement (*/rdf:RDF/rdf:Description/fgdc:citeinfo/fgdc:origin*).

5. Related Work

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

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

[4]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.

and retrieval. There are some efforts to provide interoperability among metadata standards, as can be seen in [Nogueras-Iso et al. 2003, Chandler and Foley 2000]. Use of ontologies to deal with interoperability problems in the geospatial domain is discussed in [Visser et al. 2002, Fonseca and Rodriguez 2007, Klien et al. 2004].

Another area is representation of information. RDF is being widely used for representing geographic meta-information. In [Córcoles et al. 2003], RDF is used to define a catalog of geographic resources from various Web sites. Córcoles and Gonzáles [Córcoles and González 2004] 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 [Baru et al. 1999], 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[5] format is discussed in [Westermann and Klas 2003], 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 [Jones et al. 2001].

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 work proposes 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 from the geographical and agricultural domains, 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 - like Sesame [Broekstra et al. 2002] and Jena [Wilkinson et al. 2003] - 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.

References

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

Baru, C., Chu, V., Gupta, A., Ludäscher, B., Marciano, R., Papakonstantinou, Y., and Velikhov, P. (1999). XML-based information mediation for digital libraries. In *DL '99: Proceedings of the fourth ACM conference on Digital libraries*, pages 214–215, New York, NY, USA. ACM.

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

Broekstra, J., Kampman, A., and van Harmelen, F. (2002). Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. pages 54–68. Springer Berlin / Heidelberg.

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

Córcoles, J. E. and González, P. (2004). Using RDF to Query Spatial XML. In *Web Engineering*, pages 316–329. Springer Berlin / Heidelberg.

Córcoles, J. E., González, P., and López-Jaquero, V. (2003). Integration of Spatial XML Documents with RDF. In *ICWE*, pages 407–410.

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

Jones, M. B., Berkley, C., Bojilova, J., and Schildhauer, M. (2001). Managing Scientific Metadata. *IEEE Internet Computing*, 5(5):59–68.

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

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

Klien, E. and Lutz, M. (2005). The Role of Spatial Relations in Automating the Semantic Annotation of Geodata. In Cohn, A. and Mark, D., editors, *Spatial Information Theory: International Conference, COSIT 2005,Ellicottville, NY, USA, September 14-18, 2005*, Lecture Notes in Computer Science (LNCS), pages 133–148. Springer.

Macário, C. G. and Medeiros, C. B. (2009). Specification of a framework for semantic annotation of geospatial data on the web. *SIGSPATIAL Special*, 1(1):27–32.

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

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

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

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

Westermann, U. and Klas, W. (2003). An analysis of XML database solutions for the management of MPEG-7 media descriptions. *ACM Comput. Surv.*, 35(4):331–373.

Wilkinson, K., Sayers, C., Kuno, H., and Reynolds, D. (2003). Efficient RDF Storage and Retrieval in Jena2. In *EXPLOITING HYPERLINKS 349*, pages 35–43.

XML:DB Initiative. Frequently Asked Questions About XML:DB. Available in http://xmldb-org.sourceforge.net/faqs.html. Accessed in April 4th, 2009.