Annotating Geospatial Data based on its Semantics

Carla Geovana N. Macário
Institute of Computing - P.O.Box 6176
University of Campinas - UNICAMP
13083-970, Campinas, SP, Brazil
Embrapa Agriculture Informatics - P.O.Box 6041
Embrapa, Brazil
carlamac@ic.unicamp.br

Sidney Roberto de Sousa,
Claudia Bauzer Medeiros
Institute of Computing - P.O.Box 6176
University of Campinas - UNICAMP
13083-970, Campinas, SP, Brazil
sidney@lis.ic.unicamp.br,
cmbm@ic.unicamp.br

ABSTRACT

Geospatial information (GI) constitutes a significant portion of available data and are a key factor in planning and decision-making in a variety of domains,such as emergency management and agriculture. However, to be used, these data have to be interpreted, sometimes producing new data and information. This new information is generally embedded on additional files, or remains on experts' brains. Hence, every time a user wants to use its knowledge, data have to be interpreted again. This paper presents a framework for alleviating this problem based on semi-automatic annotation of geospatial data. This framework is described in detail, as well the choices made in its design and implementation. At the end, we present a case study in agriculture, used to validate our proposal.

Categories and Subject Descriptors

H.2 [**Database Management**]: Database Applications—*Spatial databases and GIS*

Keywords

Semantic Annotation, Geospatial data, Semantic Interoperability, Geospatial standards

1. INTRODUCTION

The term *geospatial data* refers to all kinds of data on objects and phenomena in the world that are associated with spatial characteristics and that reference some location on the Earth's surface. Examples include information on climate, roads, or soil, but also maps or telecommunication networks. According to [31], this kind of data corresponds to about 80% of the available data. Therefore, geospatial data contribute significantly to human knowledge. They constitute a basis for decision making in a wide range of domains, from studies on global warming to those on urban planning or consumer services.

ACM GIS '09 November 4–6, 2009. Seattle, WA, USA

However, to be used, these data have to be analyzed and interpreted. These interpretations are context and domain dependent and performed several times. Interpretations produce new information, which is stored in technical files and often never recorded. Hence, every time a user wants to use such information, the data have to be interpreted again. The absence of solutions to efficiently store these interpretations leads to problems such as rework and difficulties in information sharing.

One approach to alleviate these problems is the use of *annotations*. An annotation, in this paper, is defined as data that describe other data and, in this sense, can be used to store interpretations of geospatial data. However, the simple adoption of annotations is not enough, as each expert or researcher, company or country has its own language and description methods, which can create barriers for understanding the meaning of the description. Hence, semantics are needed. This gave origin to the notion of *semantic annotations*, in which ontologies are used to eliminate ambiguities and promote a common understanding of concepts. This moreover, promotes semantic interoperability among data producers and consumers.

There are several initiatives based on this approach. However, they focus on offering a methodology for manual annotation of data. This is a hard task, especially considering the volume of data to be processed. It is also prone to errors, when it is manually done. Our work goes a step further, presenting a computational framework for semantically annotating geospatial data. Our approach takes advantage of specific kinds of information embedded in geospatial data. This information is stored within semantic annotations, thereby enhancing information sharing and reducing the rework of data interpretation. This framework has been partially implemented and is being tested for distinct kinds of data, for agricultural planning.

The main contributions of our work are therefore: (1) the proposal of a semantic annotation mechanism for different kinds of geospatial data; (2) the definition of processes to produce annotations in a semi-automatic way; (3) the annotation framework, which supports creation, validation and management of semantic annotations of geospatial data. Our proposal follows Semantic Web standards, thereby fostering the sharing of annotated geospatial data.

The rest of this paper is organized as follows. Section 2 presents our semantic annotation framework, giving details of its architecture. Section 3 discusses implementation aspects. Section 4 presents a case study in agriculture. Section 5 contrasts our proposal with related work. Section 6

describes conclusions and ongoing work.

2. THE ANNOTATION FRAMEWORK

2.1 Semantic Annotations

This work combines characteristics of metadata and annotations into semantic annotations: metadata fields are filled with ontology terms, which are used to describe these fields. Based on this, and following [28], we define semantic annotations as follows.

Annotation Units. An *annotation unit* a is a triple <s,m,v>, where s is the subject being described, m is the label of a metadata field and v is its value or description.

Annotation. An *annotation* A is a set of one or more annotation units.

Semantic Annotation Units. A *semantic annotation unit* sa is a triple <s,m,o>, where s is the subject being described, m is the label of a metadata field and o is a term from a domain ontology.

Semantic Annotation. A *semantic annotation* SA is a set of one or more semantic annotation units.

Annotation Schema and Content. An annotation (or semantic annotation) has a schema and a content, or instances. The schema is it structure, given by its metadata fields; the content corresponds to the values of these fields.

In fact, annotation units describe data using natural language; semantic annotations use ontology classes and can be processed by a machine. Natural language content of annotations is also part of an ontology: we use instances (individuals) of the ontology classes.

2.2 Framework Overview

The basic premise of our work is that geospatial information can be used to speed up the annotation process, alleviating the task of expert analysis. Another basic premise is that, for very many kinds of geospatial data, there are core annotation procedures that can be specified by experts. Such procedures can be subsequently tailored to meet context – specific annotation demands.

Given these premises, our annotation scenario is the following. First, experts need to predefine core annotation procedures for each kind of geospatial data source (e.g., thematic maps, satellite images, sensor time series). Each such procedure is specified and stored as a workflow. Then, every time a given data source needs to be annotated, the corresponding workflow is executed, generating a basic annotation, which may be subsequently validated by experts. Moreover, such workflows can be specialized for special needs (e.g., considering a given crop in agriculture).

Although expert systems are frequently used in annotation systems [21, 30], not all of our annotation processes can be described by decision systems. Moreover, we are dealing with geographic phenomena. Hence, we have decided to use scientific workflows to describe each annotation process [33, 12]. Each workflow contains information on the annotation schema that will be used during the process, the ontologies to describe these data, the operations to perform and how to store the generated annotations.

Our steps of semi-automatic annotation follow procedures of manual annotation available in Geographic Portals, such as FAO[1] and GOS[2]. First, an annotation schema is chosen; next, it is filled with information. The resulting annotation is presented to domain experts for validation.

Figure 1 gives an overview of the annotation process supported by our framework, which has three main steps: selection of annotation workflow, workflow execution and ontology linkage. The workflow orchestrates the generation of annotation units. In the last step (linkage) each annotation unit is transformed into a semantic unit, replacing the natural language content by a reference to the associated ontology term. Users may intervene to validate the annotations being generated.

In more detail, the framework receives as input a geospatial data file to be annotated and also some provenance data. The type of data is identified and a specific workflow is selected to be executed. This workflow indicates the annotation schema, and the operations to be performed to produce annotation content. During this process, the annotation units are presented for user validation, usually a domain expert, who may choose another workflow or define a new one. In the third step, appropriate ontology terms are chosen to assemble the semantic annotations (linking annotation units to ontology terms). The semantic annotations are stored as RDF triples in a XML database, where they can be used for information retrieval, e.g. using XQuery statements.

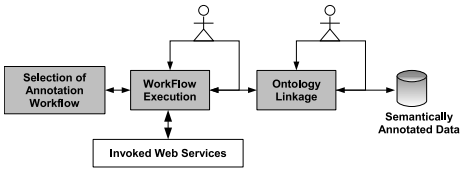

Figure 1: The GeoSpatial Data Annotation - Main steps

Configuration.

Configuration consists in a set of activities that have to be performed by domain experts to customize the annotation framework. One of the challenges we face is the specification of annotation workflows, whose purpose is to identify features to be considered for each kind of geospatial data. This is a very difficult task, and depends on experts knowledge. Hence, to produce context-dependent annotation workflows, we have to interview these experts, identifying the different information sources to be used and actions to be performed. Once the workflows are specified, it is necessary to implement the workflow modules to produce the desired annotation units.

Configuration also involves selection of ontologies, and their terms, to be used for content description. They have to be well-known, consensual, ontologies and adherent to the domain. Good examples are POESIA [12] (for agricultural zoning) and SWEET [25] (for various domains such as geography, physics, and chemistry).

2.3 Architecture of the Framework

The architecture of our framework is divided in two parts:(1) the annotation manager, annotation services and the on-

[1]www.fao.org/geonetwork/srv/en/main.home

[2]gos2.geodata.gov

tology linker, and (2) persistence layer, which includes the database manager. Figure 2 presents this basic architecture, which was designed taking into account interoperability issues. White boxes correspond to external modules invoked by the framework.

The *Annotation Manager* is responsible for managing the execution of the steps presented on figure 1, working as an event controller. It receives a request for data annotation, identifies the type of the data and makes a request for the retrieval of the corresponding workflow. This workflow will be executed by a Workflow Management System (WfMS) and once the annotation is ready and validated, it is forwarded to the Ontology Linker, for association with ontology terms. *Annotation Services* are responsible for implementing the services that are invoked by an annotation workflow to generate the desired content. The *Database Manager* works as a mediator, providing interoperability for the underlying databases. These databases contain annotation workflows, ontologies, annotated geospatial data and additional spatial data that is used by the services (e.g., historical information on crop productivity or time series for given region and phenomenon such as rainfall or temperature).

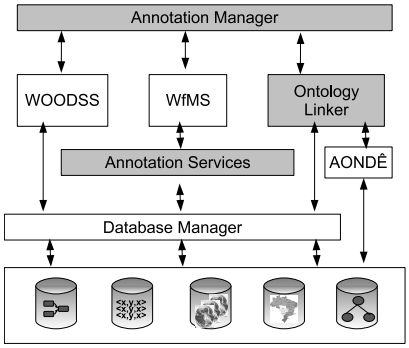

Figure 2: The Architecture of the Framework

2.3.1 Workflow Selection - WOODSS

An annotation workflow specifies the process of producting annotations tailored to each kind of geospatial content, for a given use context. These workflows are specified using WOODSS, a workflow tool [24] that provides means to edit and manage scientific workflows. All workflows are stored in a specific repository. Figure 3 illustrates a workflow specified using WOODSS, which is used for annotating NDVI time series with county, crop, production, etc.

One can see, for instance, that the generation of annotations begins by retrieving the schema for the particular data source. Once the county name is obtained (e.g., from coordinates) the next step retrieves a set of NDVI series from the same region, which are already annotated and similar to the input series. Each retrieved series is associated with a given crop. Crop names are presented to the user, as annotation suggestions. If there is more than one crop name, the user can choose the most appropriate one. Productivity is next estimated from the similar series.

2.3.2 Workflow Execution – Annotation Units

The WfMS is responsible for executing the selected workflow, through the use of a WfMS, such as the YAWL environment [35].

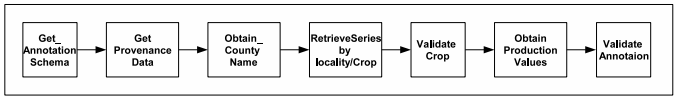

Figure 3: A workflow in WOODSS for semantic annotation of a NDVI time series

During this execution, the annotation schema to be filled is retrieved. The schema indicates which metadata elements should be used for each kind of geospatial data file. Workflow execution will produce information to fill each one of these fields. This schema is based on FGDC's [11] geospatial metadata standard, a general purpose and open standard. However, a full description using all fields from this standard may be too long. Hence, for a core geospatial annotation, we identified the most relevant parts of the schema, taking into account the metadata usually provided by some well known Geographic Portals, such as INSPIRE[3], IDEE[4], FAO[5] and GOS[6]. We also realized that the FGDC standard needs to be extended for some special domains, like agriculture. Thus, for the kinds of data we are working with, in our testbed, we have provided additional schema fields, to account for domain requirements.

Our annotation schema is divided into two parts: Identification and Extended Information. Figure 4 illustrates this schema. Section idinfo corresponds to Identification information from the FGDC standard, including citation (*citation*), description (*descript*), period that the data comprehends(*timePerd*), status of data (*status*), information of locality (*SpDom*) and keywords (*keywords*). The second part (*extendinfo*) is used to describe the information resulting from data interpretation and can vary according to the kind of data being annotated, domain being considered or usage context. In the example, for agricultural issues, it includes information on location (*location*) and on crop production (*product*).

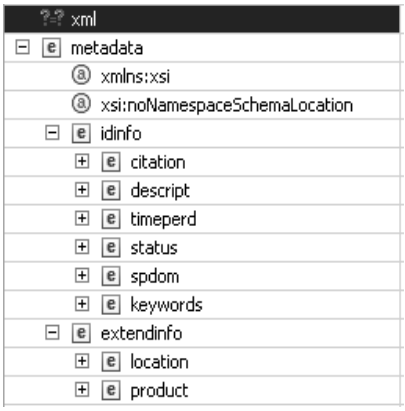

Figure 4: The adopted Annotation Schema

During workflow execution, each annotation unit is produced as a triple <*resource identification*> <*metadata schema label*> <*content*>, using natural language to describe the

[3]www.inspire-geoportal.eu
[4]www.idee.es
[5]www.fao.org/geonetwork/srv/en/main.home
[6]gos2.geodata.gov

content. A group of services of the *Annotation Services* are executed to produce the content to fill the fields. These services have to access the persistence layer to obtain information for annotation content. Part of this information comes from provenance data, e.g. the creation process of a file; part comes from the geospatial data file, like coordinates; and part are produced by the interpretation of the data, like a name of a place or the productivity of a crop. The produced annotation units are presented to the user (domain expert) for validation, and that is the reason for natural language usage. The user may change the content, or request the execution of another annotation workflow. The user may also add new annotation units.

At the end of this step, the resulting annotation is ready to be linked to ontology terms, i.e., to be transformed into a *semantic annotation*.

2.3.3 Ontology Linker

This module is responsible for linking each annotation unit to a term in an ontology. In other words, an annotation unit <*resource identification*> <*metadata schema label*> <*natural language content*> will be transformed into a semantic annotation unit by linking the content to an ontology term. The module thus deals with our second challenge: automatic identification of the ontology terms to be used. Existing tools for semantic annotation, such as [27], [5] and [15], yield this responsibility to the user performing the annotation task.

Before linkage, our annotation units contains terms in natural language. Although convenient, this approach can lead to ambiguities: users can fill the fields as they like, producing annotations that may not be machine or software understandable.

For example, consider that we have a remote sensing image containing a crop region. Also consider our FGDC-based annotation schema to describe this image, where the *origin* field describes the name of the organization/individual that created the file. Now, consider that the annotation workflow fills the *origin* field with the text "UNICAMP", based on the coordinates associated with the input file. If the annotation unit is intended to be used just for (human) users to browse, and moreover within a specific work environment, this may be satisfactory. However, if it is intended to be reused by software or outside users, or integrate this data set with others, such software will have to somehow interpret the content of the *origin* field to infer that it means a university.

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 [21]. The use of ontology terms guarantees unique meaning, associating annotation units to concepts that semantically represent their content. Ontologies also provide a hierarchical structure that helps to understand their concepts. Figure 5 shows the solution for this example, using terms of POESIA Agricultural Zoning ontology [12]. It indicates that *University of Campinas* is a public university and furthermore it is an organization categorized as a public institution. Here, an annotation unit might be <*resource_id*><*origin*><*UNICAMP*> while its semantic interpretation is <*resource_id*><*fgdc:origin class= "http:// www.lis.ic.unicamp.br/poesia#PublicUniversity"*> <*'University of Campinas'*>.

The Aondê ontology Web service [9] plays an important role in the linkage process, looking for and querying appropriate ontology terms, or aligning ontologies available within the framework to those used by external sources. For instance, suppose the annotation field *origin* is filled with "State University of Campinas". However, this is not a term on the used ontologies. Hence, using AONDÊ alignment services, it is possible to look for synonyms or the correct term – in fact just *UNICAMP*. Alignment involves identifying term and structure similarities between ontologies, and in our case is ensured by Aondê.

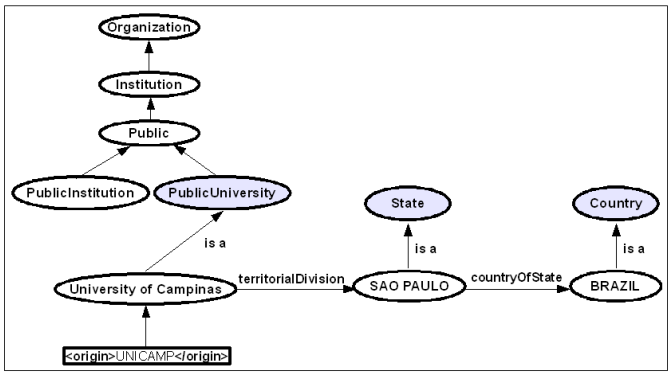

Figure 5: Associating an ontology term to an annotation field

Given the country's context and our domain context, our primary ontological sources come from the Brazilian Agriculture Ministry – e.g., on soil, live animals, vegetation, agro-ecological relief and other agriculture-related issues. Information on other geospatial features, including an ontology with over 16,000 terms concerning Brazil's spatial unit names and relationships, was taken from IBGE[7] – Brazil's National Geographic Institute.

3. IMPLEMENTATION ASPECTS

We are implementing a framework that supports the whole annotation process to validate our proposal. Its design and construction followed the main principles of adoption of standards and ontologies to provide interoperability. The framework is being implemented in JAVA, since it provides several APIs that can facilitate our work. It also is centered on XML files, which facilitates data exchanging. Since WOODSS does not have a native execution engine, we adopted YAWL for this task [35]. Each activity in the workflow is linked to a Java annotation service.

3.1 Configuring the Framework

3.1.1 Editing Workflows

We use WOODSS [24] to edit the workflows, since this is an environment easy to use and it supports annotations of workflows and their storage in a database. In WOODSS, workflows (which are themselves annotated to allow their reuse) are stored in the PostGreSQL DBMS. This allows the automatic selection of the appropriate workflow to execute, which can be retrieved according to the annotations attached to it (e.g., indicating that it is a workflow that orchestrates the annotation of a satellite image, for crop identification in agriculture). WOODSS does not have a native

[7]www.ibge.gov.br

execution engine, and its workflows have to be exported for execution.

3.1.2 Choosing Ontology Classes

Recall that the configuration process involves the specification of annotation workflows, but also of the ontologies and ontology terms to be used when semantically annotating a specific geospatial dataset, for a given usage context.

Our semantic annotations use ontology terms – classes and their instances. For example, *Brazil* is an instance of the class *Country* and is used to identify a Country, in natural language. The semantic description is given by the class' URI. Hence, during production of annotation units production, these ontology terms should be available for use. This part of the callibration process is responsible for this.

Ontology selection is performed by an expert, using a Web interface. Figure 6 illustrates this process, which has three main steps: selection of ontologies, selection of ontology terms and their association to annotation fields and storage of this information. In the first step, the user types the URL of some ontology of interest to be used for the annotations. The module loads this ontology and extracts all the URI's of the ontology terms, using the Jena Ontology API[8]. Having all these URI's, the user is asked to indicate which term can be used to fill each annotation field. Note that one term may be associated to one or more annotation fields. At the end, the module stores the URI of the chosen terms, and the label of associated annotation fields in a database.

At this part of the framework, the expert has to indicate the ontology classes to be used in each annotation field, for a semantic description. As most of these classes have instances associated, the name of these instances will work as a controlled vocabulary of natural language terms to be used during the generation of the annotation units. However, in case of absence of appropriate instances, classes can be used to characterize the content. Another option is the usage of AONDÊ, for ontology alignment. Considering the example of figure 5, "*University of Campinas*" is a natural language description for *origin*, whose semantic description is *"http://www.lis.ic.unicamp.br/poesia#PublicUniversity"*.

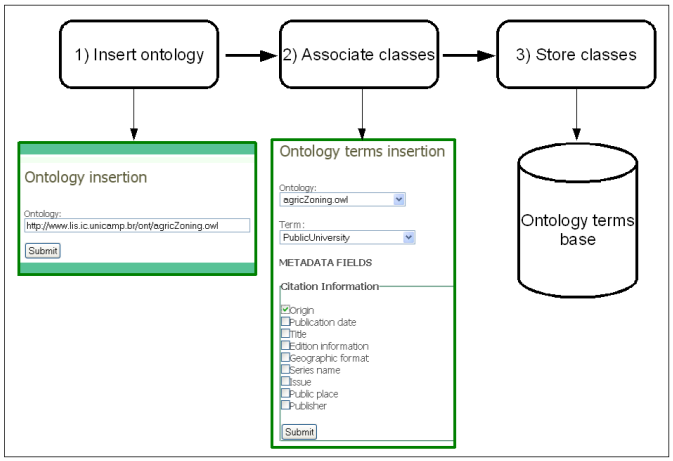

Figure 6: Process of association of ontology terms to annotation fields

[8]http://jena.sourceforge.net/ontology/index.html. Accessed in June 15th, 2009.

This implementation option enables us to easily change the used ontology whenever needed, without damage to previously annotated data. It also makes this feature generic for any domain being considered.

3.2 Creating Annotation Units

During the annotation process, the annotations units are stored in XML files. We used the Java Architecture for XML Binding (JAXB), a java API that easily maps Java classes to XML representations. Through JAXB, we just had to define a XML schema (XSD file) for the adopted annotation schema and the API generates java classes to read and write an XML file in accordance with the given XSD file. Since FGDC provides the corresponding XSD files for their geospatial metadata standard, we just had to adapt theses files to our needs.

Figure 7 presents part of the XML Schema for our annotation schema presented in section 2.3.2. For example, the annotation schema in XML to be generated is composed of a field *metadata*, which has two kinds of metadata: *idinfo* and *extedinfo*. Field *idinfo* is of *idinfoType*, which indicates that it composed by other six metadata fields: *citation, descript, timeperd, status, spdom* and *keywords*.

```
<xsd:element name="metadata" type="metadataType"/>

<xsd:complexType name="metadataType">
<xsd:sequence>
<xsd:element name="idinfo" type="idinfoType" />
<xsd:element name="extendinfo" type="extendinfoType"/>
</xsd:sequence>
</xsd:complexType>

<xsd:complexType name="idinfoType">
      <xsd:sequence>
<xsd:element name="citation" type="citeinfoType"/>
<xsd:element name="descript" type="descriptType"/>
<xsd:element name="timeperd" type="timeperdType"/>
<xsd:element name="status" type="statusType"/>
<xsd:element name="spdom" type="spdomType"/>
<xsd:element name="keywords" type="keywordsType"/>
   </xsd:sequence>
</xsd:complexType>
```

Figure 7: Partial XML Schema – FGDC

The processing of this specific XML schema by the JAXB API produced 43 Java classes. These classes are responsible for the creation and reading of XML files containing our FGDC metadata schema.

Annotation services fill the schema fields. Implemented as Java classes, they are grouped by their functionality. For example, there are services related to region naming issues, such as to obtain the name of a county for a given location or to provide names for macro or micro region or state. Hence, these services are part of *Locality* java class. Other services are related to crops, such as, given a temporal series, to identify the crop it refers to, or to obtain productivity values for a given crop, in a specific place and year. These are specified in the *Crop* class.

When one of these services is executed, it produces some kind of description in natural language. Such descriptions are instances of ontology classes, which were selected on the configuration phase. The identification of the candidate term can be done based on different issues: by the geospatial component – e.g., for a county name; by previously annotated data – e.g., when comparing historical series; by the use of some predefined patterns – e.g., for some descriptions fields.

These services have to access different kinds of data during

their execution, such as spatial information, historical data and temporal series. This could be a problem, as the service has to know how this data is stored and in which database. To facilitate this task, the framework provides the *Database Manager* layer, which works as a mediator, being responsible for accessing all the used DBMS, such as PostGreSQL for relational data and workflows, PostGIS for spatial data and XML databases. Hence, through the methods provided by this layer, the access to the data is performed in a transparent way, regardless on how the data is stored.

3.3 Creating Semantic Annotation Units

Our semantic annotations are represented using the *Resource Description Framework*[9] (RDF). RDF/XML is a language for RDF, structured in XML. RDF identifies resources using their URI's and describes them using statements. A statement is composed of a subject, a predicate, and an object. From the geospatial point of view, a subject is a geospatial resource (e.g. 'Image 1'), a predicate is an annotation unit field of this resource (e.g., 'origin'), and an object is the value filling this field – e.g. *'University of Campinas'*.

Figure 8 illustrates an annotation unit of a remote sensing image, considering the schema presented on figure 7. The *rdf:Description* element indicates a description of some resource. The *rdf:about* attribute identifies the resource by its URI. Next, come the annotations units fields, using the following rule: if an element is composed of one or more elements, it must have a *rdf:parse Type="Resource"* attribute indicating that it contains other elements.

```
<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#"
         xmlns="http://www.lis.ic.unicamp.br">

    <rdf:Description
rdf:about="http://caramuru.lis.ic.unicamp.br/efarms/remotesensing/images
/image01.tiff">
        <fgdc:citeinfo rdf:parseType="Resource">
            <fgdc:origin>University of Campinas</fgdc:origin>
            <fgdc:pubdate>20070114</fgdc:pubdate>
            <fgdc:title>Coffee crop region</fgdc:title>
            <fgdc:edition>1.0</fgdc:edition>
            <fgdc:geoform>Remote sensing image</fgdc:geoform>
            <fgdc:serinfo rdf:parseType="Resource">
                <fgdc:sername>Remote sensing images from Sao
Paulo state</fgdc:sername>
                <fgdc:issue>Crops monitoring</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 8: RDF annotation of a remote sensing image

In order to link annotation content to ontologies, we use the ontology instances of the annotation units to identify the ontology terms that will be used on the mapping to the semantic annotation units. As these instances are related to ontology classes, it is quite simple to provide the semantic description for the annotation units. As we want to maintain the "natural language" description of the annotation units, we use the predicate *rdfs:comment* from RDF Schema[10] (RDFS), which represents a human-readable description . Hence, a semantic annotation unit is a triple, using the property *rdf:type* to specify that the content of the semantic annotation unit is an individual of an ontology class. In the example of figure 5, the field *origin* contains a human readable description (content of *rdfs:comment*), which says that the resource was originated by "University of Campinas", and a reference to the class *PublicUniversity* (*rdf:resource=http://www.lis.ic.unicamp.br/poesia#PublicUniversity"*), specifying that the originator of the resource is an instance of this class (via *rdf:type*). Thus, we want to say that "the resource was originated by UNICAMP, which is a public university".

```
<fgdc:origin rdf:parseType="Resource">
    <rdfs:comment>University of Campinas</rdf:comment>
    <rdf:type rdf:resource="http://www.lis.ic.unicamp.br/poesia#PublicUniversity">
</fgdc:origin>
```

Figure 9: Referencing an ontology term to *fgcd:origin* element.

3.4 Storing Semantic Annotations in RDF

Another issue we faced was to choose how to store annotations. RDF can be represented by various languages, the RDF/XML language is the most common. One of the essential characteristics 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 e-GMS [1]. These facts made us choose a XML database to store RDF/XML semantic annotations.

An XML database is a data persistence software that allows storage of data in XML format, mapping these data from XML to some storage format, which can be a relational database or even other XML documents [41]. Queries over a XML database are generally executed using XPath or XQuery statements. It is possible to retrieve RDF/XML data using XQuery.

XPath and XQuery allow retrieval of full XML-based documents or subtrees thereof, using their DOM trees[11]. If we know the schema of an annotation that we want to retrieve, we can retrieve the full annotation or a part of interest. For example, if someone wanted to know who originated the remote sensing image of the example from figure 8, he could retrieve this information using the XPath statement (*/rdf:RDF/rdf:Description/fgdc:citeinfo/fgdc:origin*).

4. CASE STUDY - AGRICULTURAL PLANNING IN BRAZIL

Brazil is a big country, with a diversity of soil, relief, crops, crop management practices, climate conditions and diseases which can break productivity. These several factors influence crop prediction and estimates. They are also used for zoning issues, indicating which crop should be planted in a locality in the country, given a period of time, which information – prediction, estimates and zoning – are the basis

[9]http://www.w3.org/RDF. Accessed in June 10th, 2009.

[10]An extension to RDF for defining application-specific classes and properties

[11]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.

for Brazilian government polices to finance agricultural activities. Besides this, at reaping time, the follow up of this information ensures the payment of insurance, when needed, and allows new financings.

All of this led to the search for more objective and efficient estimation and prediction methods. Remote sensing images are intensively used for crop monitoring, providing a basis for decision making based on soil occupation changes. Examples of their use are the identification of extension and kind of crop, diseases, or management actions, such as soil treatment.

Agricultural experts have to manually interpret these data to obtain the desired information. We are now using our framework to automate part of this interpretation, taking into account the geospatial component. For example, through the coordinates of an image, and using some historical data, it may be possible to derive not only the region's name, but also the crop and its productivity. Semantic annotations are then used to record these annotations, allowing their reuse by information consumers.

Figure 10 presents a remote sensing image of Monte Alto county, located in one of the Brazilian regions with the highest coffee productivity index. Annotations that are result of the our process are, for instance, the county name, and production and climate factors.

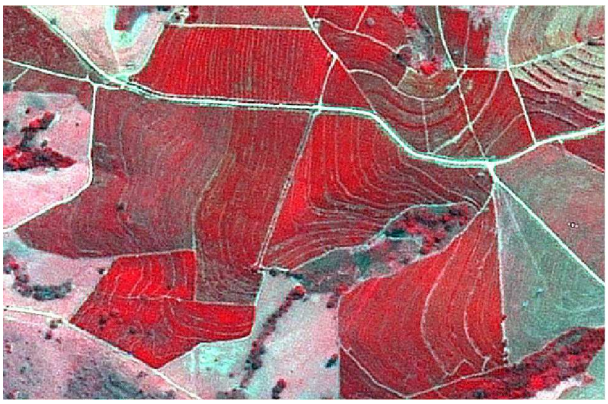

Figure 10: Remote sensing image for arabica coffee in Monte Santo county

Figure 11 presents the workflow for annotation of a remote sensing image. After the selection of the schema, an image classification tool is invoked. This tool [10] uses image processing techniques, and based on spatial and texture information, provides vegetation cover identification (here, crop name). If the user validates the crop, historical productivity values are obtained for this crop in the same region. These values are obtained from IBGE database, which maintains information of productivity for different crops, grouped by geographic region – macro and micro region, state and county – and by year.

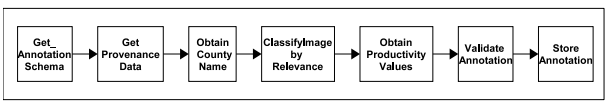

Figure 11: The core workflow for annotation of Remote sensing images

Figure 12 presents part of these annotations. This corresponds to the Extended Information of the schema. For example, the image is related to arabica coffee crop (Crop Identification), the pair *<crop>*, *<rdf:li rdf:resource="http://www.lis.ic.unicamp.br/ont/agricZoning.owl#Arabica"/>*.

```
<fgdc:formcont>
-
<rdf:Bag>
<rdf:li
rdf:resource="http://sweet.jpl.nasa.gov/ontology/biosphere.owl#Crop"/>
<rdf:li>Coffe crop</rdf:li>
</rdf:Bag>
</fgdc:formcont>
</fgdc:digtinfo>
</fgdc:digform>
-
<crop>
-
<rdf:Bag>
<rdf:li
rdf:resource="http://www.lis.ic.unicamp.br/ont/agricZoning.owl#Arabica"/>
<rdf:li>Arabica Coffee</rdf:li>
</rdf:Bag>
</crop>
</rdf:Description>
```

Figure 12: Semantic annotation generated for a remote sensing image

Figure 13 shows a table that explains the terms used in the semantic annotation of this image. The first column shows the annotation fields used. Each field shown in the table is composed by other specific fields, which were abstracted in the table. The second column has a brief description of each element. The third column shows its short name, defined in their respective XML Schema. The fourth column indicates from which metadata standard the field belongs. The fifth column specifies whether the presence of the element is mandatory or not. The last column indicates the ontologies used to describe each annotation field.

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

Figure 13: Composition of a semantic annotation of a Remote Sensing Image

The experts just have to validate the created semantic annotations. Using them, a Brazilian government expert may confirm the extension of a crop, producing correct productivity values. Another important use is the identification of diseases, impacting insurance. As an additional gain, our annotations, because of the semantic descriptions, can enhance the number of relevant documents retrieved in a query operation (the recall factor).

5. RELATED WORK

Our paper concerns semantic annotations of geospatial data, including tools and to generate and manage these annotations. This section presents related work concerning these issues, which comprises semantic annotation tools, the

use of semantic annotations to record interpretations and representation and sharing of meta-information.

5.1 Existing Annotation Tools

Annotation of digital content, due to the volume of available information, is not an easy task, always subject to errors. This led to the development of tools, which aim to facilitate the annotation process. We have tested some of them, taking into account the requirements pointed by [30] and [34]. Embrapa Information Agency [32], Amaya [37], KIM [27] are examples of traditional mechanisms for annotation, where the spatial component is not considered. They are mainly based on pattern identification, such as stored strings, and machine learning. AKTiveMedia [5] and CREAM [15] present methods for semantic annotation of visual resources.

In geographic applications, annotations should also consider the spatial component, since geographic information associates objects and events to localities, through a rich vocabulary of places and geographic object names, spatial relationships and standards. Hence, the geospatial annotation process should be based on geospatial evidences – those that conduct to a geographic locality or phenomenon, e.g. see [3, 18]. E-Culture [16], OnLocus [3], SPIRIT [17] and Semantic Annotation of Geodata [20] are approaches that consider the spatial component for the annotation of digital contents.

Except for the SPIRIT project, all the analyzed tools use a *standard format*, like XML, OWL or RDF to save their annotations. Among them, [32],[15] and [20] also adopt standardized metadata (Dublin Core, VRA and ISO 19115), which increases the probability of the annotated content to be found. On the other hand, annotations which are saved on RDF or OWL enable the annotated content to be found during a semantic search, through the use of ontologies. During this comparative study of annotations tools, reported in [22], we also observed that when the data to be annotated are mainly textual, without taking the spatial component into account, the annotation method is based on machine learning. In this case, since the identification of annotations is based on string matching, the use of an ontology is essential for the disambiguation. The same occurs when the spatial component is taken into account: if the process is automated, the use of ontologies is a key factor for the correct identification of spatial evidences. However, if the content is an image or a video, it has to be manually annotated. The analyzed tools do not consider other kinds of content, like maps and graphs, for annotation.

Tools have also to be compared considering storage features, since the efficiency of the annotation process is measured by the results of a content search. Annotations stored in an annotation server, like a catalog – as in [27] and [15] – facilitate content discovery, different from those stored in local files ([5]). On the other hand, annotations stored in a relational database, as in [32], will not enable content discovery, unless they are also published in another media, like web pages.

Like these tools, we rely on ontologies for annotation. Unlike them, we combine several components in our framework to facilitate the annotation process and to foster reuse of annotations. Moreover, our framework is extensible and general purpose.

5.2 Using Annotations to Record Interpreted Information

There are several initiatives that use annotations to store data interpretation. Wang et al. [38] present a framework to annotate medical images, as a way to promote information sharing, in a collaborative annotation process. The annotations can be textual or multimedia. The former ones are based on a limited group of metadata and are used to describe regions of interest on the image. The latter are used to enrich existing information. Unlike us, they do not consider semantic issues.

Rainaud et al. and Mastella et al. [29, 23] deal with recording of interpretations of geological data for oil companies. The authors point that these interpretations, produced by geoscientists, are very important. They propose a methodology to store the interpretation of raw data using a semantic repository. The interpreted data (research papers, public reports) are stored in a repository. A semantic repository is used to relate the raw data and the interpretation, by the use of terms of ontologies. The creation of these ontologies is part of the methodology, considering reservoir studies. The work also concerns automatic generation of data, but different from our work, they just focus on textual resources.

5.3 Management of Metadata

Use of ontologies to deal with interoperability problems in the geospatial domain is discussed in [36, 13, 14, 20], but not focusing on the use of geographic metadata, while [26, 6]discuss interoperability among geographic metadata standards.

Another trend is the representation of geographic meta-information, in which RDF is being widely used. In [8], RDF is used to define a catalog of geographic resources from various Web sites. Córcoles and Gonzáles [7] propose an approach for providing queries over spatial XML resources with different schemas using a unique interface, where the resources are integrated using RDF. Although these works concern aspects like integration and interoperability, they do not explore the use of ontologies.

Our framework uses XML databases to store metadata in RDF/XML, due to the conventional use of XML to share and store meta-information. There are some works that also use XML databases to store other kinds of metadata. In [2], 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[12] format is discussed in [39], with a survey concerning XML database solutions for this issue. A schema-independent XML database used to store metadata about scientific resources is presented in [19].

Another solution for storing and querying RDF is to use some framework for these purposes, like Sesame [4] and Jena [40]. These frameworks play the role of a layer that manages 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. We intend to use a framework like these in the future and so compare this approach to the storage in XML databases.

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

6. CONCLUSIONS AND FUTURE WORK

Geospatial data are a basis for decision making systems. However, these data have to be interpreted to be used. Even when recorded, this interpretation is hard to understand; this increases the cost of decisions made on such data. The absence of approaches to efficiently store these interpretations leads to problems such as rework and difficulties in information sharing.

This paper presented and discussed an approach for alleviating this problem based on semi-automatic annotation of geospatial data. This approach was outlined in [22] and this paper discusses architectural ans implementation issues. Our proposal, which is being validated in the domain of agricultural planning and monitoring, presents the following characteristics: it is compliant to Semantic Web standards; the descriptions are free of ambiguities in their understanding; and it promotes interoperability.

A real case study for agriculture was presented, discussing the semantic annotations obtained for a remote sensing image. We have implemented part of the framework, which still lacks an appropriate user interface, to help annotation updates. This is part of our ongoing work. The next steps to be followed are: selection of other kinds of content to be annotated, such as maps for erosion control, implementing the services to produce the desired information; implementing the semantic annotation storage in RDF database, just like OpenRDF [13]. An annotation can be extended to multimedia (e.g. voice annotations). However, this remains an open problem to be attacked in the future.

Acknowledgment

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

7. REFERENCES

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

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

[3] K. A. V. Borges, A. H. F. Laender, C. B. Medeiros, and J. C. A. Davis. Discovering geographic locations in web pages using urban addresses. In *GIR '07: Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 31–36. ACM, 2007.

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

[5] A. Chakravarthy, F. Ciravegna, and V. Lanfranchi. AKTiveMedia: Cross-media document annotation and enrichment. In *Fifteenth International Semantic Web Conference (ISWC2006) - Poster*, 2006.

[6] A. Chandler and D. Foley. 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, 2000.

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

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

[9] J. Daltio and C. B. Medeiros. Aondê: An ontology web service for interoperability across biodiversity applications. *Information Systems*, 33(7-8):724–753, 2008.

[10] J. A. dos Santos, R. A. Lamparelli, and R. da S. Torres. Using relevance feedback for classifying remote sensing images. In *Proceedings of Brazilian Remote Sensing Symposium*, 2009.

[11] FGDC. *FGDC-STD-001-1998. Content Standard for Digital Geospatial Metadata*. Washington, D.C., June 1998.

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

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

[14] F. T. Fonseca and M. J. Egenhofer. Ontology-driven geographic information systems. In *GIS '99: Proceedings of the 7th ACM international symposium on Advances in geographic information systems*, pages 14–19. ACM, 1999.

[15] S. Handschuh and S. Staab. Authoring and annotation of web pages in CREAM. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 462–473. ACM Press, 2002.

[16] L. Hollink, G. Schreiber, J. Wielemaker, and B. Wielinga. Semantic annotation of image collections. In *Workshop on Knowledge Markup and Semantic Annotation - KCAP'03*, 2003.

[17] C. Jones, A. Abdelmoty, D. Finch, G. Fu, and S. Vaid. The SPIRIT spatial search engine: Architecture, ontologies and spatial indexing. In *Geographic Information Science: Third International Conference, Gi Science 2004*, pages 125 – 139, October 2004.

[18] C. B. Jones, A. I. Abdelmoty, and G. Fu. Maintaining ontologies for geographical information retrieval on the web. In *OTM Confederated International Conferences - CoopIS, DOA, and OOBASE*, pages 934–951, 2003.

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

[20] E. Klien. A rule-based strategy for the semantic annotation of geodata. *Transactions in GIS*, 11(3):437–452, 2007.

[21] E. Klien and M. Lutz. The role of spatial relations in automating the semantic annotation of geodata. In *Proceedings of the Conference of Spatial Information Theory (COSIT'05)*, volume 3693, pages 133–148, 2005.

[13] www.openrdf.org. Accessed in June 10th, 2009.

[22] C. G. N. Macário and C. B. Medeiros. A framework for semantic annotation of geospatial data for agriculture. *Int. J. Metadata, Semantics and Ontology - Special Issue on "Agricultural Metadata and Semantics"*, 4(1/2):118–132, 2009.

[23] L. S. Mastella, M. Abel, L. F. De Ro, M. Perrin, and J.-F. Rainaud. Event ordering reasoning ontology applied to petrology and geological modelling. In *IFSA 2007 World Congress on theoretical advances and applications of fuzzy logic and soft computing.*, pages 465–475. Springer-Verlag, 2007.

[24] C. B. Medeiros, J. Pérez-Alcazar, L. Digiampietri, G. Z. P. Jr., A. Santanchè, R. S. Torres, E. Madeira, and E. Bacarin. Woodss and the web: Annotating and reusing scientific workflows. *SIGMOD Record*, 34(3):18–23, 2005.

[25] NASA. Semantic web for earth and environmental terminology (sweet).

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

[27] Ontotext Lab. *The KIM Platform: Semantic Annotation*. Ontotext, 2007.

[28] G. Z. Pastorello Jr, J. Daltio, and C. B. Medeiros. Multimedia Semantic Annotation Propagation. In *Proceedings 1st IEEE Int. Works. on Data Semantics for Multimedia Systems and Applications (DSMSA) – 10th IEEE Int. Symposium on Multimedia (ISM)*, 2008.

[29] J.-F. Rainaud, L. S. Mastella, P. Durville, Y. A. Ameur, M. Perrin, S. Grataloup, and O. Morel. Two use cases involving semantic web earth science ontologies for reservoir modeling and characterization. In *W3C Workshop on Semantic Web in Oil & Gas Industry*, 2008.

[30] L. Reeve and H. Han. Survey of semantic annotation platforms. In *SAC '05: Proc.of the 2005 ACM symposium on Applied computing*, pages 1634–1638, 2005.

[31] A. Sonal and A. Sharma. Semantics for decision making. *The Global Geospatial Magazine*, 13(4):42–44, 2009.

[32] M. I. F. Souza, A. D. Santos, M. F. Moura, and M. D. R. Alves. Embrapa information agency: an application for information organizing and knowledge management. In *II Digital Libraries Workshop*, pages 51–56, 2006. (in portuguese).

[33] A. Tsalgatidou, G. Athanasopoulos, M. Pantazoglou, C. Pautasso, T. Heinis, R. Gronmo, H. Hoff, A. Berre, M. Glittum, and S. Topouzidou. Developing scientific workflows from heterogeneous services. *SIGMOD Record*, 35(2):22–28, 2006.

[34] V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, and F. Ciravegna. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(1):14–28, january 2006.

[35] W. P. van der Aalst and A. ter Hofstede. Yawl: yet another workflow language. *Information Systems*, 30(4):245–275, 2005.

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

[37] W3C and IRIA. *Amaya, W3C's Editor/Browser*. W3C, 2007.

[38] F. Wang, C. Rabsch, and P. Liu. Native web browser enabled svg-based collaborative multimedia annotation for medical images. In *Proceedingds of 24th International Conference on Data Engineering - ICDE*, 2008.

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

[40] K. Wilkinson, C. Sayers, H. Kuno, and D. Reynolds. Efficient RDF Storage and Retrieval in Jena2. In *Exploiting Hyperlinks 349*, pages 35–43, 2003.

[41] XML:DB Initiative. Frequently Asked Questions About XML:DB. http://xmldb-org.sourcefourge.net/faqs.html.