Specification of a Framework for Semantic Annotation of Geospatial Data on the Web

Carla G. N. Macário[1,2], **Claudia B. Medeiros** [1]

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

[2]Embrapa Agriculture Informatics – Embrapa
PO Box 6041 – 13083-886 – Campinas/SP – Brazil

{carlamac,cmbm}@ic.unicamp.br

Graduate Program: Doctorate in Computing Science
Advisor: Claudia B. Medeiros
Admission: march 2006
Expected Conclusion: march 2010
Concluded Stages: qualification
Keywords: information integration and interoperability; modeling and data semantics; data services for Web; geographic information systems.

Abstract. *Efficient retrieval of geospatial information (GI) available on the Web is a key factor in planning and decision-making in a variety of domains. Specifications such as those provided by the Open Geospatial Consortium (OGC) are concerned with proposing data annotation and exchange standards to enable syntactic interoperability. However, a number of problems caused by semantic heterogeneity still present challenges. One possible approach to tackle these problems is to elicit knowledge by means of semantic annotations, based on multiple ontologies. This work proposes a framework to support management of semantic annotations for digital content on the Web – e.g. satellite images, sensor data, maps and graphs – for agricultural planning and monitoring. This will help end-users (agronomers, farmers, Earth scientists) to work cooperatively in developing integrated practices for land management.*

1. Introduction

Agriculture is an important activity in Brazil. According to [IBGE 2008], in 2007 approximately 25% of Brazil's GNP of U$ 1,477 billion corresponded to agricultural activities. This could even be larger, if experts could enhance their use of geospatial data, thus supporting more accurate prediction and planning methods.

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, soil and temperature, but also maps or satellite images. Such data are a basis for decision making in a wide range of domains, in particular agriculture. Their combined use is useful to answer strategic questions such as '*When will be the best time to start harvesting coffee in this area?*' or '*Given a crop productivity pattern, which regions show the same pattern?*'. These questions are important for production planning and definition of public policies concerning agricultural practices, also allowing the environmental control of protected areas.

There is a large amount of research on the management of geospatial data available on the Web, including proposals of models, data structures, exchange standards and querying mechanisms. However, they mainly consider textual resources and relatively few computer scientists are concerned with the specific requirements of applications in agriculture. Novel solutions must be found to support adequate management and retrieval of geospatial data on the Web – satellite images, maps, graphs –, taking all these factors into consideration, with agriculture in mind.

This thesis proposes a solution based on exploring the use of *semantic annotations* as a key for the semantic interoperability issues in discovery, access and effective search for data. In our work, a semantic annotation is a set of one or more metadata fields, where each field describes a given digital content by ontology terms.

The expected contributions of our work are: (1) the proposal of an annotation mechanism directed to the agricultural context; (2) specification of scientific processes describing the generation of semantic annotations; and (3) a framework, which should consider the following requirements: annotation of any kind of agricultural geospatial data; discovery of geospatial data in the Web, their fusion and/or integration; end user validation of the annotations generated; automate the annotation process as much as possible.

This research is being conducted within the WebMAPS multidisciplinary project under development at UNICAMP, whose goal is to develop a platform based on Web Services for agro-environmental planning [Macário et al. 2007].

The remainder of the text is organized as follows. Section 2 describes related work. Section 3 outlines the proposed annotation service and Section 4 presents an illustrating example. Sections 5 and 6 contain conclusions and the present stage of the work.

2. Related Work and Concepts

2.1. Annotation

"To annotate" means to add notes, to comment. In computing, an *annotation* is used to describe a resource (usually a textual resource) and what it does, by means of formal concepts (e.g., entities in an ontology) [Ontotext Lab 2007]. An annotation is represented by a set of metadata – data about data – that reference each annotated entity by its unique Web identifier, like a URI. Some benefits from the adoption of annotations include the increase in quality of the retrieved information and in interoperability. However, names can vary through time, or in their usage. Therefore, the simple adoption of ontologies during the annotation process is not enough. An ontology is useful to distinguish, for instance, orange fruit from orange color, but it is not enough to describe if a document is about the fruit itself or concerns orange culture management.

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.

2.2. Existing Annotation Tools

Annotation of digital content, due to the volume of available information, is not an easy task, always subject to errors. This leaded 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 [Reeve and Han 2005] and [Uren et al. 2006]. Embrapa Information Agency [Souza et al. 2006], Amaya [W3C and IRIA 2007], KIM [Ontotext Lab 2007] 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 [Chakravarthy et al. 2006] and CREAM [Handschuh and Staab 2002] present methods for semantic annotation of visual resources. E-Culture [Hollink et al. 2003], OnLocus [Borges 2006], SPIRIT [Jones et al. 2004] and Semantic Annotation of Geodata [Klien 2007] 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, [Souza et al. 2006],[Handschuh and Staab 2002] and [Klien 2007] 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 the test we also observed that when the data to be annotated is 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 in the content 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. We also analyzed the tools considering the storage feature, since the efficiency of the annotation process is measure by the results of a content search. Annotations stored in an annotation server, like a catalog – as in [Ontotext Lab 2007] and [Handschuh and Staab 2002] – facilitate content discovery, different from those stored in local files ([Chakravarthy et al. 2006]). On the other hand, annotations stored in a relational database, as in [Souza et al. 2006], will not enable content discovery, unless they are also published in another media, like web pages.

A detailed description and comparison of these tools are presented on a paper submitted to the Int. J. Metadata, Semantics and Ontology - Special Issue on Agricultural Metadata and Semantics.

3. Proposed Annotation Service

3.1. The WebMAPS Project

WebMAPS [Macário et al. 2007], in which this research project is been conducted, is a multidisciplinary project that aims to provide a platform based on Web Services to formulate, perform and evaluate policies and activities in agro-environmental planning. It involves state-of-the-art research in specification and implementation of software that

relies on heterogeneous, scientific and distributed information, such as satellite images, data from sensors and geographic data.

Although supporting a wide range of queries, WebMAPS is still limited in terms of semantic support. The goal of this research is to provide such a support via semantic annotations. The annotation service will be responsible for getting information from other data sources and combine them to produce a more meaningful result. A catalog service will be responsible for the management and publication of the produced annotations.

3.2. The Annotation Service

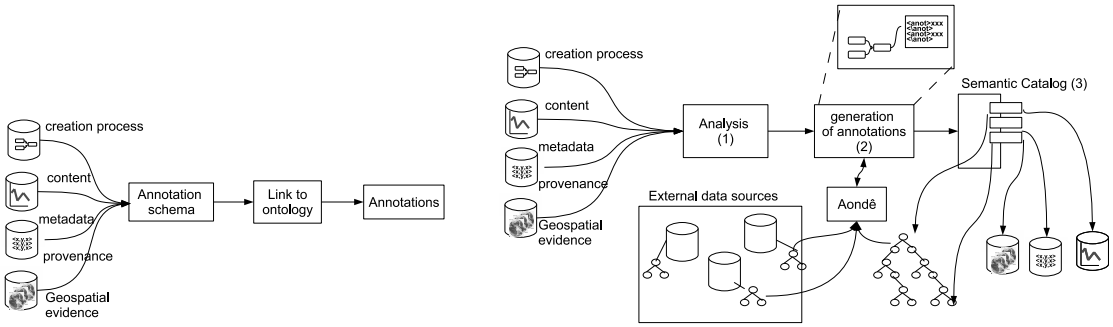
The goal of the annotation service is to semantically annotate different kinds of geospatial data, such as satellite images, maps and graphs. Agosti and Ferro [Agosti and Ferro 2007] propose a formal model for annotation of different kinds of digital content, such as textual documents, images, and multimedia documents in general. According to them, an annotation model should be as uniform as possible, considering all kinds of content, but also flexible, making it possible to exploit the semantics each content has, providing an effective collaboration tool for users.

Taking this into account, our annotation service should not only be based on explicit geospatial features, like geographic coordinates, but also on features that can be derived from the content, like climate and temperature or productivity trends. We are dealing with different kinds of digital content, each one with distinct geospatial features. The service should consider these differences, defining a specific annotation process for each kind of content. Although expert systems are frequently used in annotation systems [Klien 2007, Reeve and Han 2005], not all of our processes can be described by decision systems. Hence, we have decided to use scientific workflows to describe each annotation process [Tsalgatidou et al. 2006, Fileto et al. 2003]. Each workflow contains information on the data annotation schema that will be used during the process, the ontologies that describe these data, which operations must be performed and how to store the generated annotations.

Figure 1(a) presents a high level view of the workflow that annotates content. For instance, if the content is an image mosaic, it uses information from the graph's metadata (e.g., it is a JPG file), its provenance (e.g., the satellite images used to create it), its creation process (recorded as a scientific workflow), and geospatial evidence (extracted from content, metadata, provenance and process). First, the *annotation schema* to be used is selected (i.e., the metadata fields that will be used in an annotation) and next the schema is filled with ontology terms. Additional annotations can be defined manually.

In WebMAPS, scientific workflows are used to specify models in agriculture (e.g., to analyze erosion trends, or to define areas suitable for a given crop [Fileto et al. 2003]). Workflows may also be used to specify how to create some kinds of content within WebMAPS (e.g., erosion maps or NDVI graphs). These workflows are stored in a database to be subsequently queried and reused [Medeiros et al. 2005]. Hence, the annotation service can take advantage of this workflow base to determine information on content.

Figure 1(b) gives an overview of the annotation service, comprising 3 basic steps. Step 1 selects the annotation workflow to be performed, based on the nature of the content to be annotated. Step 2 comprises the execution of the selected workflow. Finally, once the



(a) A workflow for content annotation (b) The proposed annotation process

Figure 1. NDVI graph with possible semantic annotations

annotations are generated, in step 3 the framework publishes them in a catalog, enabling the data discovery and analysis provided by WebMAPS.

Annotation generation will require accessing several data sources, including external data. The desired data will be discovered through metadata catalogs, using WebMAPS catalog service. We will only consider those catalogs that use domain ontologies to semantically describe data they represent. After the new metadata are generated, the framework has to relate them to one or more ontologies, giving them a semantic meaning, thus creating the annotations. The Aondê Web Service [Daltio and Medeiros 2008] plays an important role in the annotation process, looking for and querying appropriate ontologies, or aligning those available within WebMAPS to those used by external sources.

Since we are focusing on interoperability, our framework will take advantage of the standards provided by the Open Geospatial Consortium, like the Geographic Markup Language [OGC 2007]. The backbone for the annotation schema will probably use FGDC's [FGDC 1998] or ISO 19115 [ISO 2008] geospatial metadata standards. However, we expect that it will be necessary to extend it to support the complex requirements of agricultural applications.

The research project combines work on ontologies, annotation mechanisms and scientific workflows. Hence, the main challenges comprise: how to deal with heterogeneity questions? how to yield the desired results, using distinct filtering and aggregation criteria? which annotation schema is best? how to combine the available data? how to design the workflows? how to store, to manage and to maintain annotations? what annotations each kind of data should have? which are the most important for the agriculture domain?

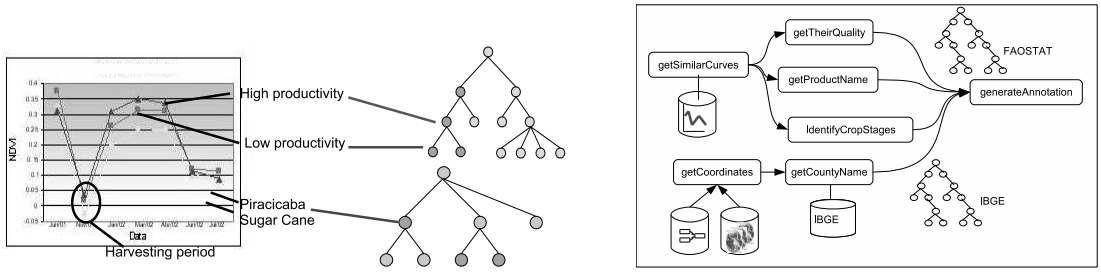
4. An Illustrating Example

Traditional approaches for annotation of geospatial data focus on textual content, where the spatial component is explicit. In agricultural domain, a lot of strategic data like satellite images, maps and graphs cannot be annotated using them. This section presents an illustrating example of our approach, showing how the produced annotations can be useful for answering strategic question.

Remote sensing has become one of most important research areas in agriculture, taking advantage of satellite imagery. These images require distinct kinds of preprocess-

ing. An example are the so-called NDVI images, whose pixels contain NDVI values, calculated by the difference of the spectral reflectance of red and near-infrared regions and normalized by the sum of both. NDVI represents the biomass conditions of a vegetation area and is widely used in distinct kinds of analysis – e.g. agriculture, biodiversity. An NDVI graph plots the average NDVI pixel value in a region though a temporal series of images. This can be used for crop monitoring and prediction. For example, in the sugar cane culture, a curve with higher values may indicate a product with better quality. Curves can be compared and analyzed for yield forecast, for instance, quality and productivity, or to identify regions with problems. Considering this, it is possible to select curves indicating similar productivities to identify which is the best period for harvesting the crop. This usually involves series of steps, but using the proposed annotations it becomes an easier operation.

Figure 2(a) illustrates a set of NDVI graphs, together with a few possible semantic annotations that can be generated for them, associated with ontologies and figure 2(b) presents the corresponding workflow for generation of these annotations. Figure 2(a) shows two curves, respectively representing graphs for periods with high and low productivity, for the same region and months of a year. Productivity is a kind of semantic annotation that has been added to the curves. One can use tools that mine time series (e.g., see [Mariotte et al. 2007]) to determine information on crops for a given region, based on NDVI value or oscillation behavior; here, this resulted in identifying crop = "sugar cane". Given an NDVI graph, by its period and locality (latitude and longitude), it is also possible to obtain other information such as season, temperature and climate conditions, geographic region. So, through the coordinates provided, the graph was annotated with county name "Piracicaba". Finally, annotations can identify production phases, like planting and harvesting. Each of these annotations are linked to ontology terms, like those provided by FAOSTAT and IBGE.



(a) An annotated NDVI graph (b) The annotation workflow

Figure 2. NDVI graph with possible semantic annotations

5. Concluding Remarks

Semantic annotations are subject to much research, in distinct contexts. Their use has many goals, such as data discovery, integration and adding meaning to data. Most research focuses on annotation of textual resources, without consider the spatial component. When others resources are treated, like images, they are manually annotate by the user. The same occurs in most cases in which the spatial component is taken into account: although spatial ontologies are used, the spatial description is manually done. Finally, most of

approaches are not concerned with a specific domain. The generality of an approach can decrease the possibilities of annotations.

The goal of this research is to provide a mechanism for semantic annotations of geospatial, distributed and heterogeneous data available on the Web, geared towards generation of strategic information for agriculture. It should support all steps of the annotation process, helping experts to document data sources, but also supplying means to query and extract knowledge from the annotations. Content to be annotated in this context includes, among others, satellite images, sensor data temporal series (e.g., from sensors or weather stations), and all kinds of textual data files (e.g., crop productivity reports). This will help actors in the decision-making process (policy-makers, agronomers, farmers, Earth scientists) to work cooperatively in developing integrated practices for land management.

This framework will be implemented as a web service within WebMAPS, considering the automation of the annotation process, the integration of heterogeneous data to generate annotations and the queries that will be posed on the annotated data.

6. Present Stage of Work

We have tested existing annotation and catalog tools, performing a comparative analysis. We are now interviewing agriculture experts for identifying the possible annotations each geospatial content can provide. After that, we will specify the scientific workflows for generation of the annotations and start the implementation phase. The framework validation will be done by these experts, using real data.

Acknowledgments This thesis is partially financed by Embrapa, FAPESP and CNPq.

References

Agosti, M. and Ferro, N. (2007). A formal model of annotations of digital content. *ACM Trans. Inf. Syst.*, 26(1):3.

Borges, K. A. V. (2006). *Using an Urban Place Ontology to Recognize and Extract Geospatial Evidence on the Web (in portuguese).* PhD thesis, UFMG.

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

Daltio, J. and Medeiros, C. B. (2008). Aondê: An ontology web service for interoperability across biodiversity applications. *Information Systems.* Accepted for publication.

FGDC (1998). *FGDC-STD-001-1998. Content Standard for Digital Geospatial Metadata.* Washington, D.C.

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

Handschuh, S. and Staab, S. (2002). Authoring and annotation of web pages in CREAM. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 462–473, New York, NY, USA. ACM Press.

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

IBGE (2008). *Geographic and Statistical Brazilian Institute (IBGE)*. IBGE/USP. <http://www.ibge.gov.br/english/>.

ISO (2008). *ISO 19115:2003 Geographic information – Metadata*. ISO. Available on:<http://www.iso.org/iso/home.htm>.

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

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

Macário, C. G. N., Medeiros, C. B., and Senra, R. D. A. (2007). The webmaps project: challenges and results (in portuguese). In *IX Brazilian Symposium on GeoInformatics - Geoinfo 2007*, pages 239–250.

Mariotte, L., Medeiros, C. B., and Torres, R. (2007). Diagnosing similarity of oscillation trends in time series. In *Int. Workshop on spatial and spatio-temporal data mining - SSTDM*, pages 643–648.

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

OGC (2007). *Geography Markup Language*. The Open Geospatial Consortium. <http://www.opengeospatial.org/standards/gml>.

Ontotext Lab (2007). *The KIM Platform: Semantic Annotation*. Ontotext. <http://www.ontotext.com/kim/semanticannotation.html>.

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

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

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

Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., and Ciravegna, F. (2006). 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.

W3C and IRIA (2007). *Amaya, W3C's Editor/Browser*. W3C. <http://www.w3.org/Amaya/>.