A framework for semantic annotation of geospatial data for agriculture

C.G.N. Macário*

Embrapa Agriculture Informatics,
Brazilian Agriculture Research Corporation,
P.O. Box 6041, Embrapa 13083-886, Campinas, SP, Brazil
E-mail: carlamac@ic.unicamp.br
*Corresponding author

C.B. Medeiros

Institute of Computing,
University of Campinas – UNICAMP,
P.O. Box 6176, 13083-970, Campinas, SP, Brazil
E-mail: cmbm@ic.unicamp.br

Abstract: The Web is a huge repository of geospatial information. Efficient retrieval of this information is a key factor in planning and decision-making in many domains, including agriculture. However, standards for data annotation and exchange enable only syntactic interoperability, while semantic heterogeneity presents challenges. This work describes a framework that tackles interoperability problems via semantic annotations, which are based on multiple ontologies. The framework is being developed within a project to support agricultural planning in Brazil. The paper discusses design and implementation issues using a real case study, provides an overview of annotation mechanisms and identifies requirements for annotating agricultural data.

Keywords: semantic annotation; geospatial data; interoperability; agricultural planning.

Reference to this paper should be made as follows: Macário, C.G.N. and Medeiros, C.B. (2009) 'A framework for semantic annotation of geospatial data for agriculture', *Int. J. Metadata, Semantics and Ontologies*

Biographical notes: Carla G.N. Macário is an IT Reseacher at Embrapa (the Brazilian Agriculture Research Agency), since 1989, working in Research Projects applied to Agriculture to improve the generation of technologies to increase the offer of food while conserving natural resources. She is presently working towards her PhD, which was started in 2006 at the Institute of Computing, UNICAMP, Brazil.

Claudia Bauzer Medeiros is a Full Professor of Computer Science at the Institute of Computing, University of Campinas (UNICAMP), Brazil. She is the Head of the Laboratory of Information Systems (LIS) and coordinates projects focused on Scientific Databases, with Applications in Agriculture and Biodiversity.

1 Introduction

Agriculture is an important activity all over the world. According to the Brazilian Geographic Institute – (IBGE, 2008), in 2007 approximately 25% of Brazilian GNP of U$ 1477 billion corresponded to agricultural activities. This could even increase, if geospatial data became more reliable, thus supporting enhanced 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 questions such as "*When will be the best time to start planting coffee in this area?*" or "*What is the expected sugar cane yield in a region?*". These questions are important for production planning and definition of public policies concerning agricultural practices, furthermore allowing the environmental control of protected areas. Spatio-temporal factors vary widely and are crucial in decision making.

The Web plays an important role in this scenario, having become a huge repository of geospatial information distributed all over the world, collected and stored by different organisations. Such distributed data may be retrieved and combined in an *ad hoc* way, from any source available, extrapolating their local context. Usually, the search for these data and methods is done by their syntactic content, focusing primarily in keyword matching. Semantic interoperability is a key issue needed in this context.

There is a large amount of research on the management of geospatial data, including proposals of models, data structures, exchange standards and querying mechanisms. However, relatively few computer scientists are concerned with the specific requirements of applications in agriculture – e.g., the dependence on spatio-temporal correlations as well as social and cultural constraints.

The notion of semantics is often associated with ontologies, which help the so-called *semantic search* – see, for instance (Mangold, 2007). Our solution is based on exploring the use of *semantic annotations*. In our work, a semantic annotation is a set of one or more metadata fields, where each field describes a given digital content using ontology terms. An ontology formally describes the elements of a domain and the relationships among them, providing a common understanding of the domain (Gruber, 1993).

Semantic annotations are subject of extensive research, in distinct contexts. Their use has many goals, such as data discovery, integration and adding meaning to data. As will be seen, most research focuses on annotation of textual content, without considering spatial issues. When other kinds of content are treated, they are manually annotated by the user. Even when spatial ontologies are used, the spatial description is inserted manually. Finally, most approaches do not direct their research towards a specific domain. We on the other hand, focus our work on many kinds of content, with emphasis on geospatial information, for the agricultural domain. This leads us to annotations that can be useful for activities like crop management and monitoring. Furthermore, by providing semi-automatic annotation process, we liberate users from tedious manual tasks.

Our research is centred on a framework to support

- creation, validation and management of semantic annotations of geospatial data on the Web, for agricultural planning; and consequently
- discovery and search for data in agricultural contexts.

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

The rest of this paper is organised as follows. Section 2 introduces concepts used. Section 3 presents our semantic annotation framework, and its role within WebMAPS. Section 4 contrasts our proposal with related work. Section 5 describes conclusions and ongoing work.

2 Related concepts

2.1 Geospatial Semantic Web

The Semantic Web was initially proposed by Berners-Lee et al. (2001) as a way to bring structure to the meaningful content of Web pages, creating an environment where users can obtain information based on semantics and not only in syntax. Computers would have to access structured collections of information available on pages, and sets of inference rules that they would use to conduct automated reasoning. To make this a reality, some basic issues were posed:

- to adopt standardised metadata to describe and exchange the data
- to describe information in terms that allow common understanding (e.g., ontologies)
- to expose data so that they can be found and retrieved
- to design efficient retrieval mechanisms.

The Semantic Web for geographic information, called Geospatial Semantic Web by Egenhofer (2002), is a way to process requests involving different kinds of GI. This process requires multiple spatial and domain ontologies, to be used in semantic query processing. This leads to the search for a GI retrieval framework that relies on ontologies.

In spite of extensive research, the Semantic Web is far from becoming a reality (Shadbolt et al., 2006). Although several standards have been developed and adopted, there are too many views, interests and needs of people that publish and share content in the Web. Consensual vocabularies and ontologies are hard to establish and maintain. So far, most retrieval engines are restricted to text, and other kinds of media pose countless challenges to the effective implantation of the Semantic Web.

2.2 Semantic annotations

Metadata – often called data about data – can describe an information resource, a part or a collection thereof. It can be embedded in digital content as a header or as part of a HTML or XML file. This allows updating both at the same time. However, to store metadata separately from data can facilitate its management. Hence, metadata and data itself are usually stored in different repositories, with the metadata referring to the described data.

In computing, an *annotation* is used to describe a resource (usually textual) and what it does, by means of formal concepts (e.g., using entities in an ontology)

(Ontotext Lab, 2007). An annotation is represented by a set of metadata that provide a reference to each annotated entity by its unique Web identifier, like a URI. In other words, annotations formally identify resources (in the text we use the term 'digital content') through the use of concepts and the relationships among them, and can be processed by a machine. However, names can vary through time, or in their usage, and distinct users may adopt different ontologies. Therefore, the simple adoption of ontologies during the annotation process is not enough.

In geographic applications, annotations should also consider the spatial component, since geographic information associates objects and events to localities. Hence, the geospatial annotation process should be based on geospatial evidences – those that conduct to a geographic locality or phenomenon.

Reeve and Han (2005) point out that there are two primary types of annotation methods: pattern-based and machine learning-based. Pattern-based systems are those that have an initial set of entities defined, manually or not. These entities are taken as patterns to be found on the content. If new entities are discovered, they may become new patterns. This process continues recursively until no more entities are discovered, or the user stops the process. Machine learning systems utilise two methods: probability and induction. The first use statistical models to predict the locations of entities within text – e.g., to identify address components (number, building, county). The induction methods extract rules and patterns from the data sets, reusing them in subsequent annotation processes.

The annotation process should be as automatic as possible, since a manual process can be slow and subject to errors. This remains as a challenge that has been addressed by a number of research projects (Greenberg et al., 2006). However, most of the proposed mechanisms consider annotations only of textual content, not taking into account other kinds of content. In the geospatial domain, there is also non textual content with important information to consider, e.g., satellite images and data from sensors. There is a scarcity of mechanisms to annotate these data, motivating our research.

2.3 Overview of the WebMAPS project

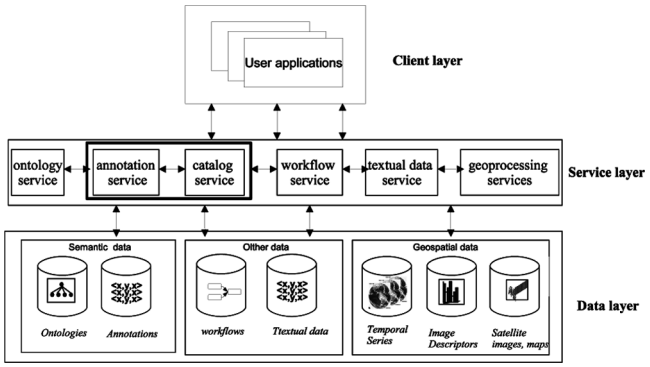
WebMAPS (Macário et al., 2007) is a project that aims to provide a platform based on Web services to formulate, perform and evaluate policies and activities in agro-environmental planning. It involves state-of-the-art research in specification and implementation of software that relies on heterogeneous, scientific and distributed information, such as satellite images, data from sensors and geographic data. This project differs from similar initiatives in the following:

- the emphasis in multidisciplinary research in Computer Science applied to Agricultural Science (whereas in most other initiatives there is almost no computer science research involved)
- the suitability to the Brazilian geographical context
- the real time exploration of image content
- the use of Human Computer Interaction aspects during all project phases.

The project caters to two kinds of users – farmers, and domain experts, such as agronomers or earth scientists. Farmers can enter data on their properties (e.g., production, parcels, crops). As a consequence, they are able to correlate data on these properties to geospatial content available on WebMAPS's repositories – e.g., satellite image series or regional boundaries. Experts may want to investigate distinct kinds of data correlation and propose models to explain, monitor, or forecast crop behaviour – see some of these tools at http://www.lis.ic.unicamp.br/projects/webmaps.

Figure 1 gives an overview of WebMAPS' 3-layer architecture, part of which already implemented. The Client Layer is responsible for processing a user request, forwarding it to be processed by the Service Layer and presenting the returned result. It uses the services provided by the Service Layer, such as: textual and geospatial data management and ontology management. The bottom Data Layer contains digital content provided by WebMAPS, including primary raw data (e.g., county boundaries from Brazilian official sources) and derived data (e.g., NDVI images or time series). Geospatial data include satellite images, region boundaries, crop information. Ontologies provide semantics. Data is stored in the PostGreSQL/PostGIS database management system.

Figure 1 WebMAPS 3-layer architecture



At present, most of the services are being implemented as software modules, to be tested by end-users. The goal is to encapsulate these modules into Web services, to enhance interoperability and support platform flexibility.

The workflow service (Medeiros et al., 2005; Kondo et al., 2007) provides means to edit, execute and manage workflows, including supply chains. It is available as a separate system, which will be incorporated into WebMAPS. The textual data service is responsible for all operations involving textual data, like input and query processing.

The geospatial data service supports functions on geospatial data, such as computation of topologic

predicates or creation of NDVI time series, visualised as graphs.

Ontology management is performed by Aondê – (Daltio and Medeiros, 2008) – a Web service responsible for handling ontologies. It provides a wide range of operations to store, manage, search, rank, analyse and integrate ontologies. If an application is a client of this service, it can enrich its semantics and interoperability by integrating and adopting concepts of ontologies published on the Web and/or available in WebMAPS.

The services surrounded by a box are those that directly concern our work. The catalog service structure was implemented to process biodiversity Web queries (Daltio et al., 2008). Its entries contain ontology terms and URIs of associated resources. It will be extended to publish the semantic annotations provided by WebMAPS' annotation service, enabling discovery and retrieval of annotations and of annotated content. Taking into account the benefits of using standard catalogs – (Nogueras-Iso et al., 2005), this service is based on standards and techniques like the ones proposed by the OpenGIS Consortium (OGC). The annotation service, discussed next, is the core of the paper.

3 The annotation service

3.1 Overview

The goal of the annotation service is to semantically annotate different kinds of geospatial data, such as satellite images and maps. According to Agosti and Ferro (2007), an annotation model should be as uniform as possible, considering all kinds of content, but also flexible, making it possible to exploit the semantics of each content.

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 productivity trends.

Our semantic annotations are composed of:

- an *annotation schema* of metadata labels
- *annotation content* – ontology terms from official Brazilian sources.

The backbone for the annotation schema uses (FGDC, 1998) geospatial metadata standards. Since this is a general purpose standard, we are extending it to support the complex requirements of agricultural applications.

We are dealing with different kinds of digital content, each with distinct geospatial features. The service considers these differences, defining a specific annotation process for each kind of content. Although expert systems are frequently used in annotation systems (Klien and Lutz, 2005; Reeve and Han, 2005), not all of our 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 (Tsalgatidou et al., 2006; Fileto et al., 2003). Each workflow contains information on the annotation schema that will be used during the process, the ontologies that describe these data, operations to perform and how to store the generated annotations.

First, the *annotation schema* is defined (i.e., the metadata fields that will be used to annotate a particular kind of content) and next the schema is filled with ontology terms. In addition, some annotations are defined manually. For instance, if the content is the graph of section Figure 4, 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 – see Figure 3), and geospatial evidence (extracted from content, metadata, provenance and process).

An important issue while constructing the annotation workflow is the nature of the content to annotate. In the example, the graph is what the user sees, but it can be stored in many ways. It can, for instance, be an image file – and thus the file is annotated. Alternatively, as in WebMAPS, it is computed dynamically and stored as a time series when so requested.

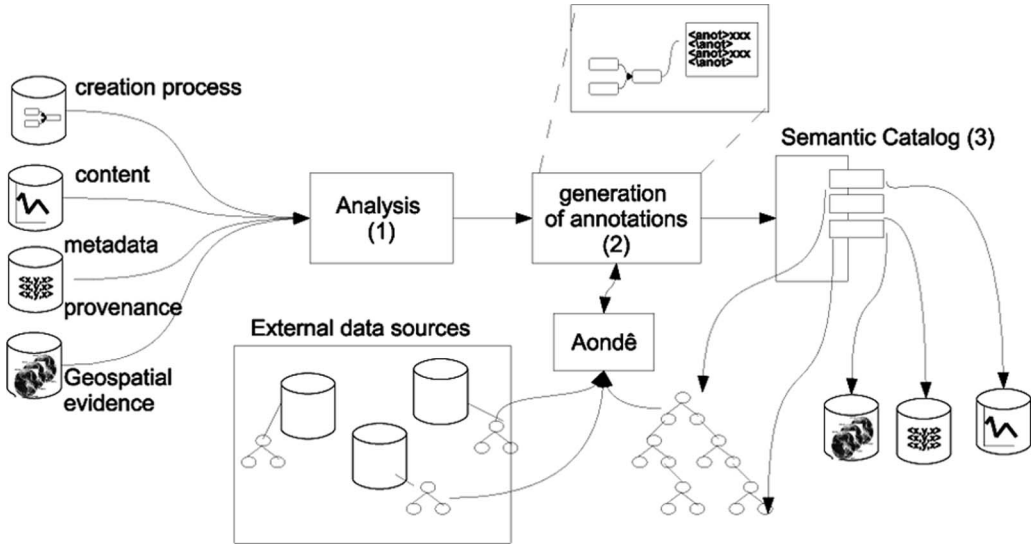
Figure 2 gives an overview of the annotation service, which comprises three basic steps. Step 1 selects the annotation workflow to be performed, based on the content to be annotated. Step 2 comprises the execution of the selected workflow. Once the annotations are generated, in Step 3 the framework publishes them in a semantic catalog, enabling content discovery. Steps 1 and 2 have been implemented and are presented in Section 3.2. Step 3 enables discovery, and requires extending the catalog service (see Section 2.3).

Annotation generation requires accessing several data sources, including external data. The latter will be discovered through metadata catalogs, using WebMAPS catalog service. We consider only those catalogs that use domain ontologies to semantically describe data they represent.

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. For instance, an external data provider may use its own ontology to classify soil units, whereas we use the ontology provided by Embrapa (the Brazilian Agricultural Research Coorporation). In order to annotate the data, both ontologies have to be compared and aligned, generating a new, extended, ontology. Alignment involves identifying term and structure similarities between ontologies, and in our case is ensured by Aondê.

Given the country's context, our primary ontological sources come from the Brazilian Agriculture Ministry, as defined and maintained by Embrapa – e.g., on soil, live animals, vegetation, agro-ecological relief and other agriculture-related issues. Information on other geographic features, including an ontology with over 16,000 terms concerning Brazil's spatial unit names and relationships, was taken from IBGE (www.ibge.gov.br). Part of this initial set of ontologies

Figure 2 WebMAPS' annotation service



is already being used by WebMAPS (e.g., on produce and on regional and ecological characterisations in Brazil). We are extending them with terms from FAO (Food and Agriculture Organisation of the United Nations) – including FAOSTAT metadata (http://faostat.fao.org) and AGROVOC thesaurus (http://www.fao.org/aims/cs_annotation.htm). Other sources, such as those provided by the SEEK project (http://seek.ecoinformatics.org/) may also be used.

At present, WebMAPS satellite image repository has images of the SPOT sensor for South America, from 1998 to 2006. These images include information on NDVI, humidity, rain, temperature, among others.

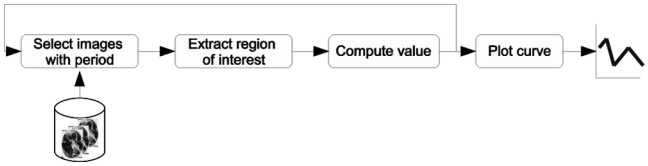
3.2 An illustrating example

This section presents an example to illustrate the requirements and some challenges of WebMAPS' annotation service: annotating an NDVI graph.

Remote sensing has become one of most important research areas in agriculture (Lunetta et al., 2003), taking advantage of satellite imagery. These images require distinct kinds of preprocessing. 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 normalised by the sum of both. NDVI represents the biomass conditions of a plant 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 analysed for yield forecast or to identify regions with problems. 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 and, sometimes, the crop it represents.

Figure 3 presents a high level view of the process used to generate a set of NDVI graphs, for a given period and region, iterating through all images for the period. The process that created the graphs is depicted as a workflow. This follows WebMAPS' design, which uses scientific workflows to specify models in agriculture e.g., to analyse 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 time series). These workflows are stored in a database to be subsequently queried and reused (Medeiros et al., 2005). The annotation service takes advantage of this workflow base.

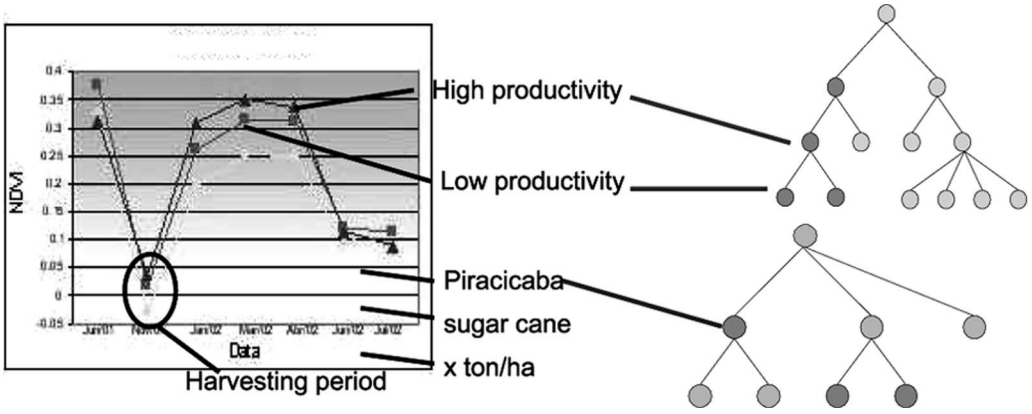
Figure 3 Scientific workflow used to generate a set of NDVI graphs



While WebMAPS uses workflows to specify models, we use workflows to guide the semi-automatic annotation process. Our annotation workflows depend not only on the nature of the content to be annotated but also on its intended used and the availability of process and provenance information. Process information, in WebMAPS, is provided via workflows.

Figure 4 illustrates a set of NDVI graphs, together with a few possible semantic annotations that can be generated for it. These semantic annotations are based on Embrapa's agricultural product ontology, on Brazil's territorial organisation ontology (Fileto et al., 2003) and on production statistics provided by the Brazilian Agriculture Ministry (www.ibge.org.br/concla).

Figure 4 NDVI graph with possible semantic annotations



The figure 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 compare NDVI information on crops for a given region. It is also possible to get the name of the region, through the coordinates provided. Here, the graph was annotated with county name 'Piracicaba'. Finally, annotations can identify production phases, like sowing and harvesting, or yield for that period. Each of these annotations is linked to ontology terms and can be used to answer some of the queries mentioned in Section 1.

We point out that the example shows at least two kinds of annotations – those that apply to the entire series (e.g., yield, region, or crop) and those that concern just part of a curve (e.g., harvesting). The first kind of annotation can be stored using, for instance, a mechanism similar to CREAM's (see Section 4), where an XML file is attached to the file containing the series – with terms such as <region> Piracicaba <\region> and <crop> Sugar cane <\crop>, for metadata fields *region* and *crop*. This kind of annotation storage mechanism is relatively straightforward, the challenge being which annotations to generate and how. The second kind of annotation, however, must be linked to the appropriate regions in the graph. This presents another level of research challenges – not only are annotations linked to parts of a graph, but these parts correspond to computed (derived) information obtained from computing average pixel values in images. We still do not know how to attack this problem in a general way; it appears frequently in agricultural applications, which are highly dependent on dynamically derived content. So far, for geospatial time series (such as those underlying our NDVI graphs), we annotate associated points.

3.3 Implementation aspects

Consider that a user wants to produce an answer to the question "*What is the expected yield of my sugar cane farm?*", Then, the user has to:

- enter the information on the farm in the WebMAPS database, including its geometry (see screen copy of data entry on Figure 5)
- generate the NDVI series for the region of the farm – see Figure 6, showing the NDVI graph dynamically generated by WebMAPS for that farm, for a given period
- use tools that mine time series to retrieve other NDVI series with similar behaviour – see Figure 7, a screen copy of our series mining tool
- analyse the annotations for these series, looking for information on the *yield* ontology term – Figure 8.

Figure 9 shows the workflow we implemented with help of expert users, to generate semantic annotations for a NDVI graph. At the moment, these workflows are being designed using the YAWL Workflow management system (Van der Aalst and Ter Hofstede, 2005), an environment that allows us to specify, simulate, validate and execute scientific workflows. During the design task, agricultural experts have suggested and revised the workflows, having agricultural issues in mind. First, the *annotation schema* is created. Next, provenance information is obtained, like coordinates of the region and sensor name (task *Get Provenance Data*). This information will serve as input for other tasks. Coordinates are used as input to task *Obtain County Name*. This task, implemented as a simple Web service, accesses a WebMAPS repository that contains data from IBGE and determines the county name. *Get Similar Curves* uses our tool for time series mining (Mariotte et al., 2007). Subsequent tasks get annotations on the associated data. Each of these tasks produces part of the annotation, which will be ready for validation at *Validate Annotation* task, performed by expert users.

Figure 10 shows part of an annotation produced for an NDVI graph, using metadata schema from the FGDC standard. It shows values assigned to the standard's *Locality information* field: *Place_Keyword*, *Spatial_Reference_Information* (latitude and longitude). Field *Spatial_Data_Organisation_Information*, uses IBGE ontology terms. We extended the FGDC standard to include other annotation fields, such as *productivity*, *crop*

Figure 5 Insertion of a farm in WebMAPS

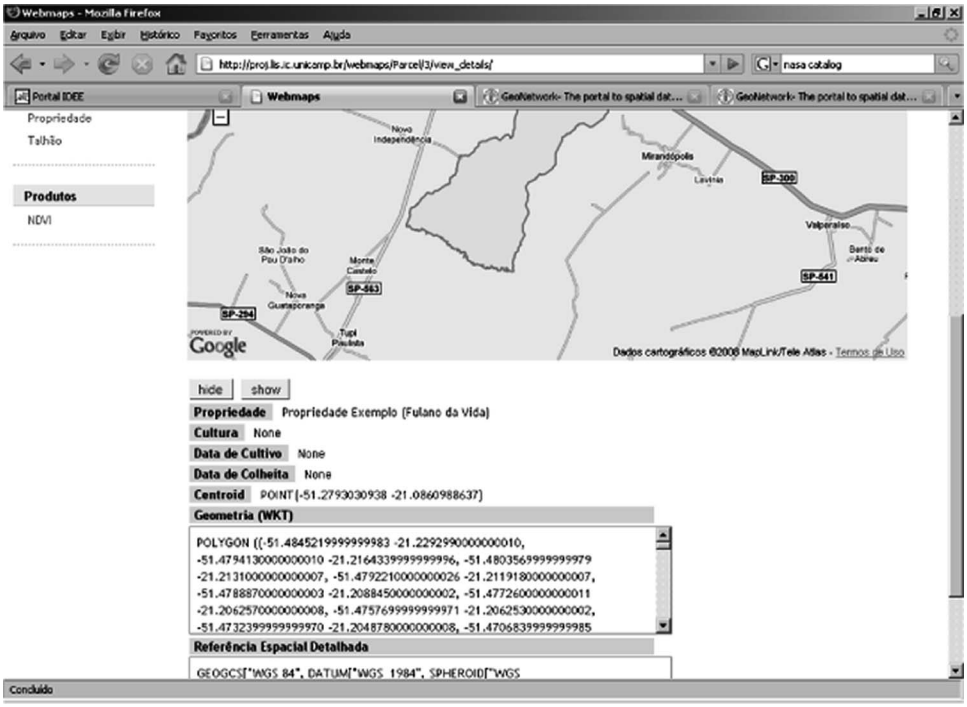
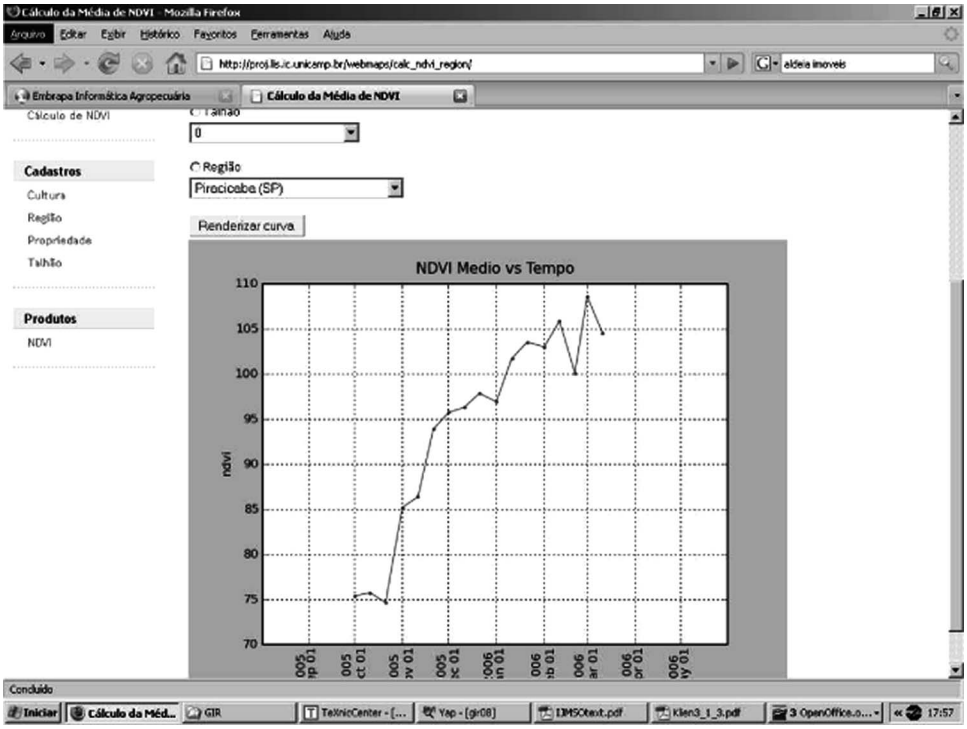


Figure 6 A NDVI graph dynamically generated by WebMAPS for the farm of Figure 5

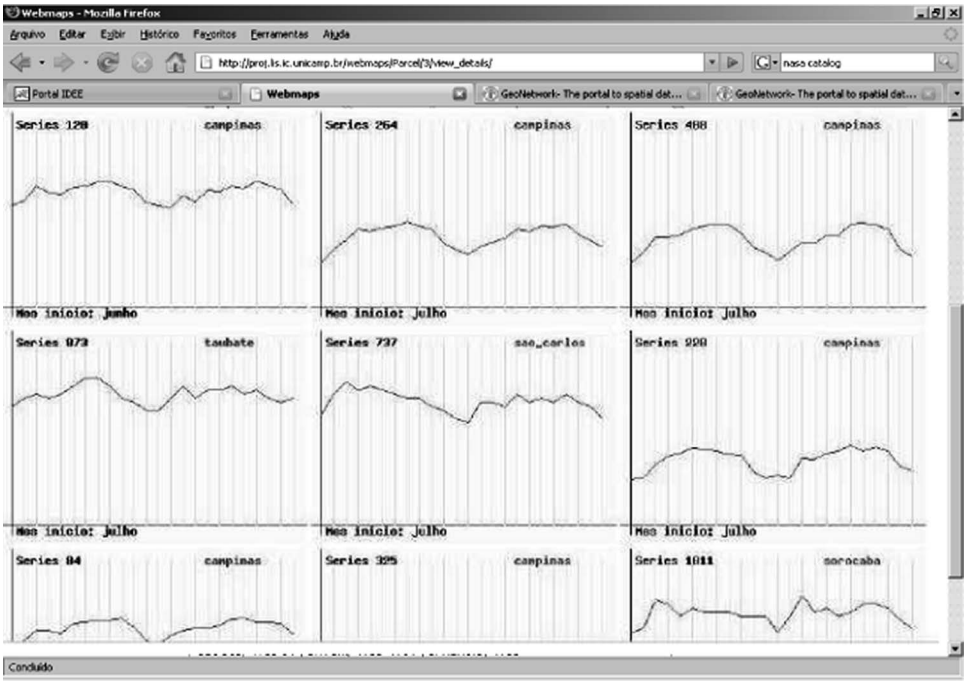


identification and *harvesting_period*. The annotation schema depends on the kind of the content being annotated. In the definition of these elements, we also considered the FAOSTAT/AGROVOC metadata.

Since our first goal is to validate the annotation process, annotation workflows are not yet executed automatically, though each step is automated. Rather, each task is invoked manually. In the example of Figure 9, the branch (*Get Provenance – Obtain County Name*) is automated, resulting in data shown in Figure 10. Similarity of curves is obtained by manually invoking our similarity tool (Mariotte et al., 2007). Extraction of annotations is obtained by SQL queries on annotations associated to the files.

Let us comment on some design and implementation challenges. First, *yield* annotates the series – but it depends on the crop and region characteristics (in particular, soil and climate variables). Thus, it is not enough to find

Figure 7 Retrieval of similar NDVI series



similar series to forecast a crop's yield: they must all refer to the same kind of soil and climate constraints. Hence, before mining for similar series, the series database has to be restricted to series for the same kind of crop, and compatible soil and geographic characteristics (activity *Retrieve Series by Soil/Crop*). Crop and soil are kinds of annotation attached to a series, so all series that have the same annotation are selected.

Figure 8 The desired answer

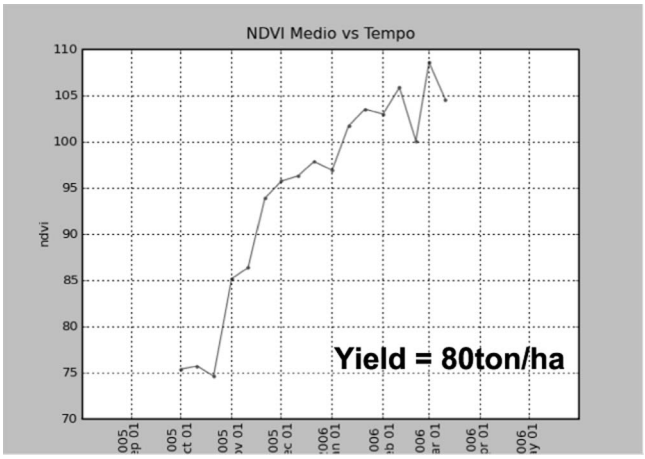


Figure 9 The workflow to annotate an NDVI graph

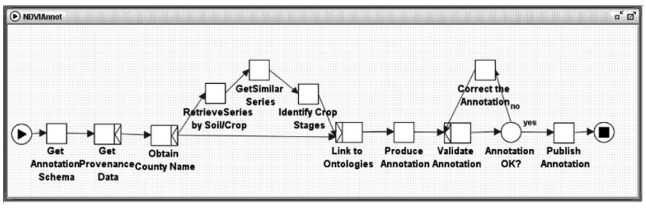
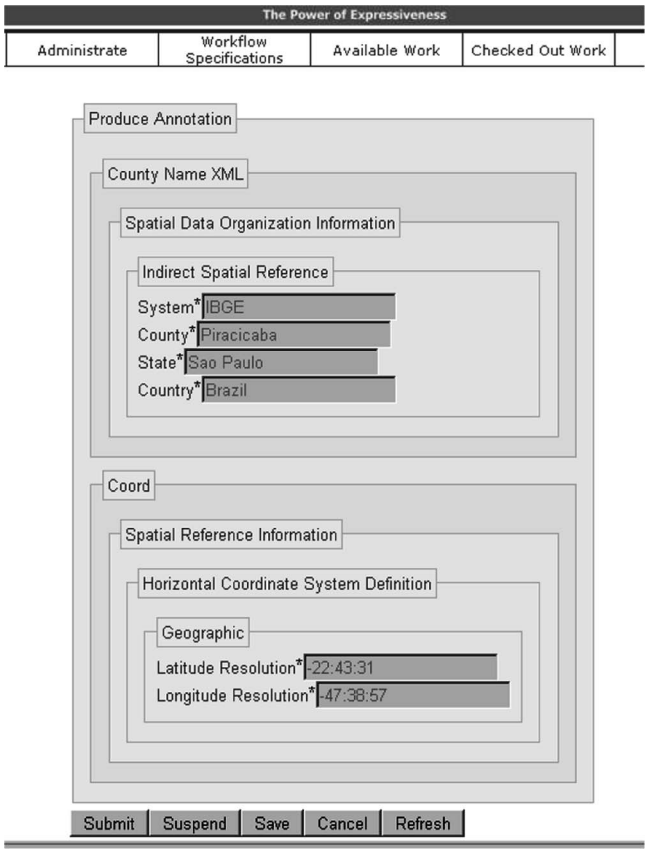


Figure 10 Part of an annotation produced for a geospatial time series



Region compatibility is much more complex. Our experts have defined which counties in Brazil have similar climate behaviour, and our county ontology has been enhanced to include links between regions with such a relationship. Hence, before executing time series mining, only a

subset S of the stored series are retrieved: those whose annotations have the same crop and soil fields (using SQL on annotations) and, for these, the ones that refer to 'compatible' regions. Compatibility search is performed by Aondê: it retrieves the names of all counties that satisfy this relationship, and these names are compared with those that annotate the files in S, to restrict S even further. The final set is used as the basis for similarity matching.

4 Related work

Though there are many annotation mechanisms on the Web, there is little or no comparison among them. This section compares some of these mechanisms.

4.1 Non spatial annotation mechanisms

Embrapa Information Agency (Souza et al., 2006), Amaya (W3C and IRIA, 2007), Knowledge and Information Management (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.

Embrapa information agency (Souza et al., 2006) is a Web system to organise, deal with, store, publish and access the technological information generated by Embrapa and other agricultural research institutes. Information is are organised through a tree branched structure named *knowledge tree*, in which knowledge is organised hierarchically. Each information node can be complemented by information resources (papers, books, image and sound files, etc.) The system uses Dublin Core metadata (Weibel et al., 1995) and allows date retrieval by different user profiles. The annotation process is fully manual and the descriptions are made in natural language, without validation. Hence, only a syntactic search for discovery of the stored resources is available. The annotations are stored in an Oracle database and the annotation process is done by librarians.

Amaya (W3C and IRIA, 2007) is a Web editor that aims to integrate as many W3C technologies as possible. It is a client of Annotea, a W3C project for advanced development in semantics. For Amaya, an annotation is a comment, note, explanation or any other kind of external markup that can be attached to a Web document. It uses an annotation schema based on RDF to describe information through metadata. The metadata currently produced consists of the author's name, title of the annotated document, annotation type, creation date, and last modification date. Annotations can be stored locally or in an annotation server. When a document is browsed, Amaya queries each of these servers, requesting the annotations related to that document.

The WebMAPS main page was annotated using Amaya. The described metadata were automatically created and the page's author could write a text to complement them.

KIM (*Knowledge and Information Management*) (Ontotext Lab, 2007) is a platform for semantic annotation of non structured or semi-structured texts on the Web. It provides an infrastructure and services for semantic annotation, ontology population, indexing and content retrieval. The basic approach is to analyse texts, in a manual or automatic way, to recognise entity references, matching them with those that are already known and have an URI and a description. For those matching references, a document reference is created, annotating the entity URI. Each annotated entity can be explored for its properties and attributes. Figure 11 shows the Kim Annotation Plug-in. In this example, the WebMAPS home page was analysed using the KIM ontology (on the left side). Five entities of class *GeneralTerm* were automatically recognised: *analysis*, *data* (datum), *factors*, *region* and *project*. The plug-in highlighted the annotated entities with the same colour of the related ontology term.

AKTive Media (Chakravarthy et al., 2006) is a system for annotation of images and text. It is based on string similarity, mining information from websites, integrating the obtained information. Initially the user manually annotates text(s) or image, based on a given ontology. The produced annotations are saved as part of a corpus to be used as basis for future annotations, enabling a semi-automatic annotation process. The system stores the collected information in an RDF base, which can be indexed for data retrieval. Figure 12 illustrates the annotation of the WebMAPS page using this framework. In this example, the annotation process was based on an ontology provided by another tool, since AKTive Media did not have one available. The instances *Laboratory of Information Systems* and *CEPAGRI* were annotated as *NonProfitOrganisation*; *Institute of Computing*, *University of Campinas* and *FEAGRI* as *EducationalOrganisation*, and *agro-environmental planning* as *Work*. During the annotation process, the system presents a ruler (upper left of the figure), where the user can inform the accuracy level of the annotation.

CREAM – CREAting Metadata for the Semantic Web (Handschuh and Staab, 2002) – is a framework that allows the creation of metadata that instantiate interrelated definitions of classes in a domain ontology. It provides facilities for page annotation, indicating parts of a text that correspond to parts of its annotation schema. The annotation can be performed manually or automatically, using, for example, geographic dictionaries or the language resources used (in XML format). The annotation schema provides a default schema, with a basic set of metadata such as person, organisation, location. This schema can be modified to cover the desired annotations. Automatic annotations are created using the processing

Figure 11 Annotation the WebMAPS main page using the Kim Annotation Plug-in

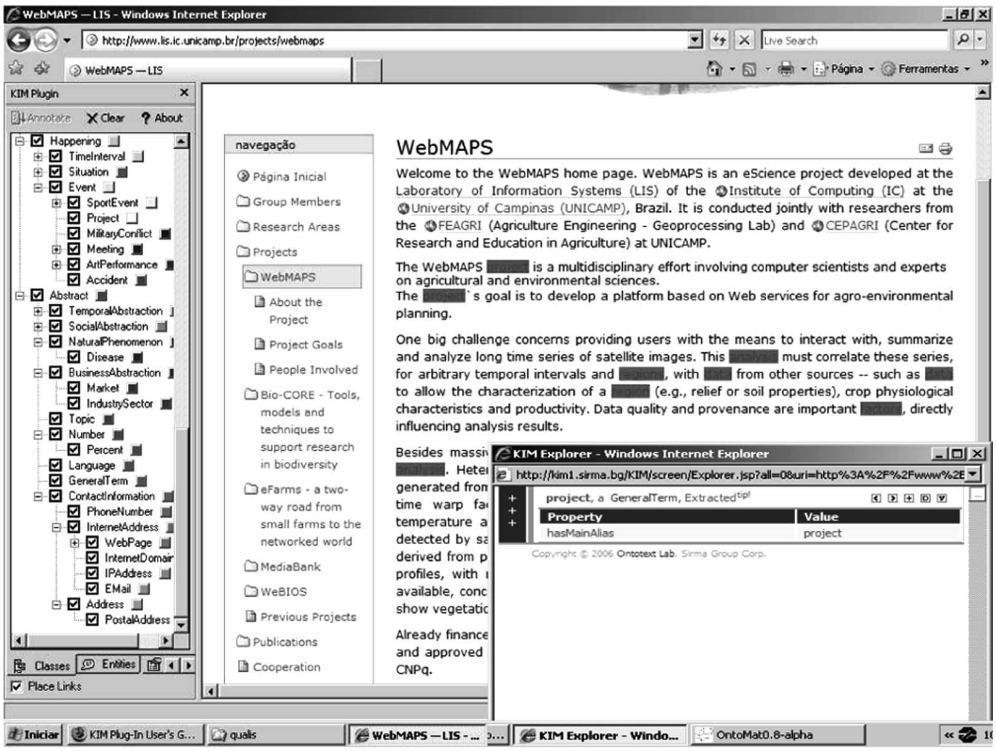
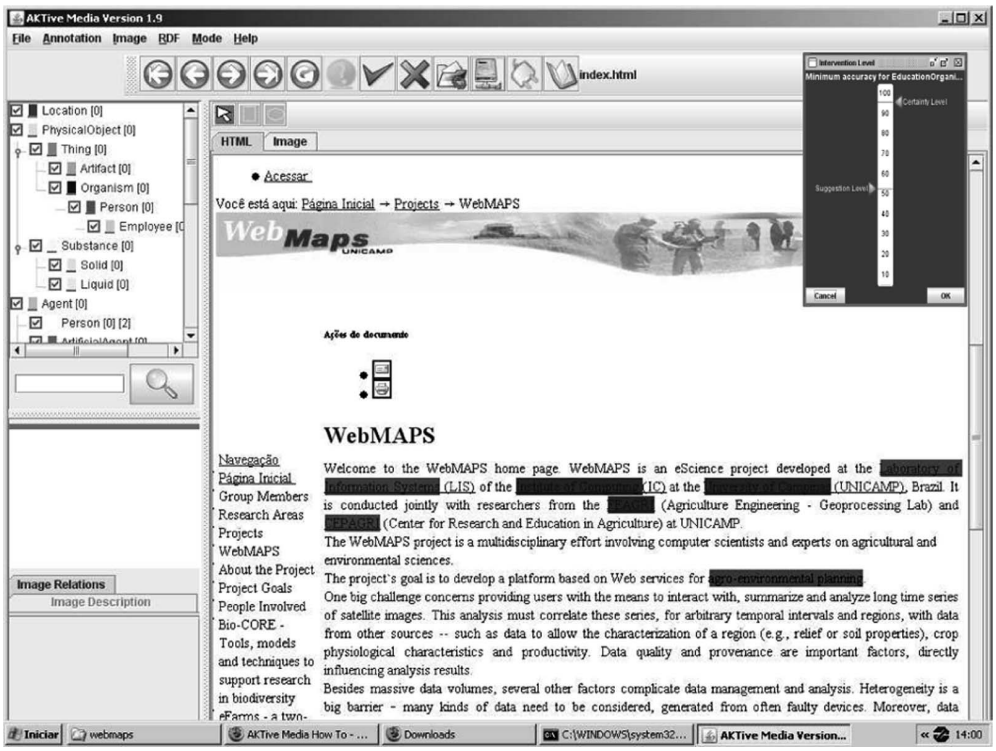


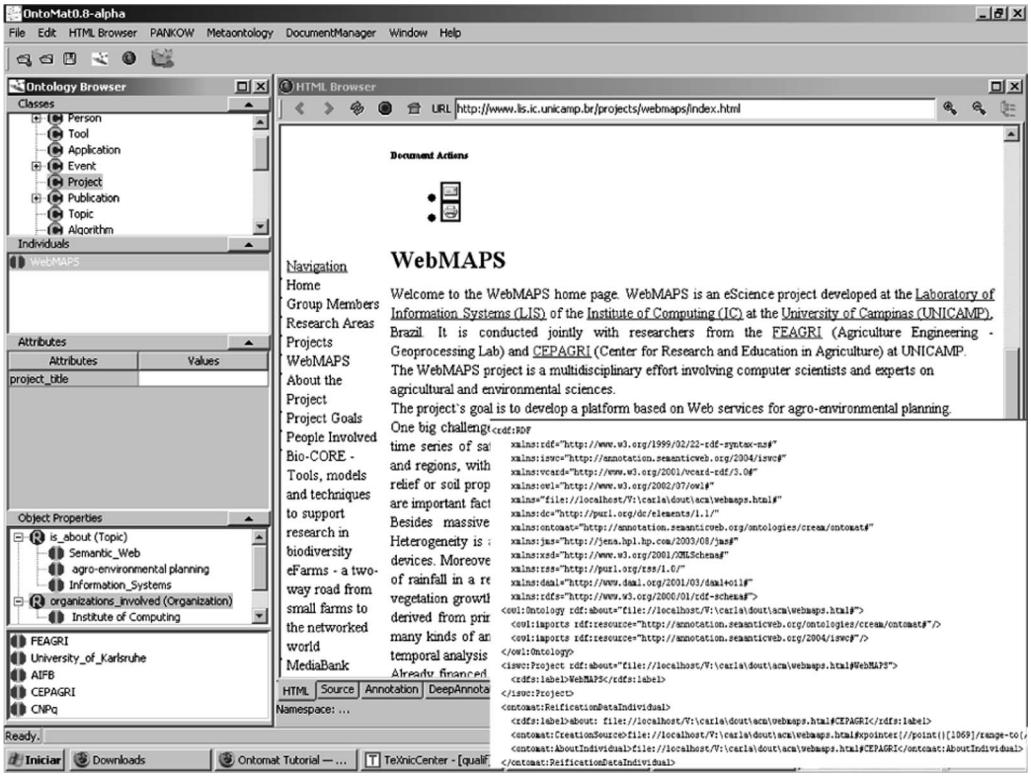
Figure 12 Annotation the WebMAPS main page using the AKTive Media



resources available. The manual annotation associates each term to a class in a given ontology. Doing this, individuals are created for the classes and the user is requested to give values to the existing attributes. This is repeated until the user is satisfied. The annotations are saved in OWL or RDF, as part of the annotated page.

Figure 13 illustrates OntoMat – CREAM's annotation tool – annotating the WebMAPS web page and part of the annotation file generated. In this example, *agro-environmental planning* was annotated as an instance of entity *Topic* and Institute of Computing, CEPAGRI and CNPq were annotated as instances of

Figure 13 Annotation the WebMAPS main page using the OntoMat tool



class *Organisation*, the last one as a *research-funding organisation*. Next, WebMAPS was annotated as an instance of entity *Project* and the previous annotations appear as available options for the instance properties, creating a relation among them.

4.2 Spatial annotation mechanisms

The traditional systems described in Section 4.1 are not able to mine for information based on spatial components, mainly because their search mechanisms do not have features to deal with spatial relationships. We now present some approaches that consider the spatial component.

E-culture (Hollink et al., 2003; Hollink, 2006) is a project that proposes an approach for semantic annotation and searching of images of paintings, sometimes considering spatial properties within an image. There are two types of spatial concepts that are considered: absolute positions (north, south, east, west, ...) – represented by WorldNet ontology – and spatial relations (right, left, above, near) – represented by terms of the SUMO ontology.

In this project, each image is annotated by VRA Metadata (VRADSC, 2007), an extension of Dublin Core (Weibel et al., 1995) for images. This schema has at last four terms – agent, action, object and recipient – where each object is associated to terms of WordNet, AAT, ULAN and Iconclass ontologies, providing semantics to the content. Each image can be described by more than one sentence. A query is processed using ontology elements. In special, during the search process, concepts like class equivalence and ontology alignment are considered, to increase the searching coverage. Although the annotation process is manual, some issues are considered to improve it, like suggesting terms. Like this proposal, we intend to take advantage of operations on ontologies to augment annotation capabilities. Unlike them, we will also use other operations on ontologies.

OnLocus (Borges, 2006) consists of a GI retrieval approach supported by the OnLocus ontology for recognising, extracting and geotagging of geospatial evidences of local features such as address, postal codes and phone numbers available on the Web. These evidences represent implicit locations, which are capable to correlate the content of a Web page, or part of it, to an urban geographic location. Search machines may use this information to retrieve pages of urban services and activities in a specific locality or near it. The OnLocus ontology consists of a set of concepts (place, territorial division, reference point), a set of spatial and traditional relationships (topological ones, all-part, location) and a set of axioms to conceptualise the domain of interest D. This domain defines urban and intra-urban places associated to the Web pages. The system was validated by experiments, using real data corresponding to a set of four million Web pages. Like our proposal, it is based on ontological spatial knowledge. Unlike ours, it is centred on annotating Web pages and is applied to urban applications.

SPIRIT – Spatially-Aware Information Retrieval on the internet (Jones et al., 2004) – is an european project whose goal is to design and implement a mechanism to help search on the Web for documents and data

sets related to places and regions. Software tools and techniques were developed to produce search agents able to recognise geographic terms that are present in Web pages and retrieve them. A prototype to validate the search mechanism was developed, working as a platform to test and evaluate new geographic information retrieval techniques.

Some challenges of this project are name disambiguation, treatment of imprecise terms and spatial query interpretation, considering ranking problems based on the relevance of the result. During the process of adding geographical identification metadata to pages being analysed (geotagging process), metadata can be associated with Web sites or images, and also with geographic information, like addresses. These metadata are usually latitude and longitude coordinates, but can also include altitude and place names. Similar to our proposal, geospatial and domain ontologies are used to eliminate name ambiguity, expand queries, rank results and extract metadata from textual sources. We extend this to other kinds of media.

Semantic annotation of geodata (Klien, 2007; Klien and Lutz, 2005) propose an approach to automatically extract semantic knowledge from geographic data, to semantically annotate them. This is part of the SWING Project, which aims at the development of Semantic Web service technology in the geospatial domain (http://www.swing-project.org/). The key to this approach is the use of multiple ontologies defined by homogeneous themes (like hydrology, geology, ecology, transportation planning) (Lutz et al., 2008). Each ontology is complemented by a set of rules that directs the information extraction process. The information sources are spatial information objects, like maps that are stored in a database. They can have spatial analysis methods associated, which are used on the extraction process.

The authors exemplify their approach with a study of floodplain areas, which can be analysed according to different aspects, such as topography, hydrology and geology. Figure 14 illustrates the procedure for annotation of existing floodplains in a map considering the geomorphology domain. The left part of the figure shows a reference dataset that already has an annotation of a river. As a floodplain, in geomorphology domain, is adjacent to a river, the system uses GIS spatial operations to identify if the dataset to be annotated has a river. Hence, if it has, the adjacent areas are considered as floodplain. An ontological description is automatically created and stored as an annotation.

Like this work, we use geographic ontologies, and also some spatial relations, during our annotation process. However, we not base the whole annotation process on them. Moreover, we will also tailor annotations to the kind of content.

4.3 Analysis of the presented tools

Table 1 shows a comparative analysis of the presented tools, taking into account the requirements pointed by Reeve and Han (2005) and Uren et al. (2006) for semantic annotation tools, to which we added criteria on the spatial component. Blank slots in the table represent information not provided.

The first column informs the format in which annotations are saved. It is an important feature, as standards increase interoperability. Column *ontology* indicates if the tool uses some ontology during the annotation process. As we have already seen, this can eliminate ambiguity of meaning. Column *Storage* informs how the annotations are stored: using a local file, a relational database or an annotation server. The next two columns are related: the first one indicates if the annotation process is automated and the next one, for which automated annotation technique (ML stands for machine learning). The *Annotated data* column describes the kind of data that can be annotated and the last one indicates if it considers some kind of spatial information. Most of the tools analysed focus on annotation of textual

Figure 14 Procedure for (semi-)automated annotation of geodata from Klien and Lutz (2005)

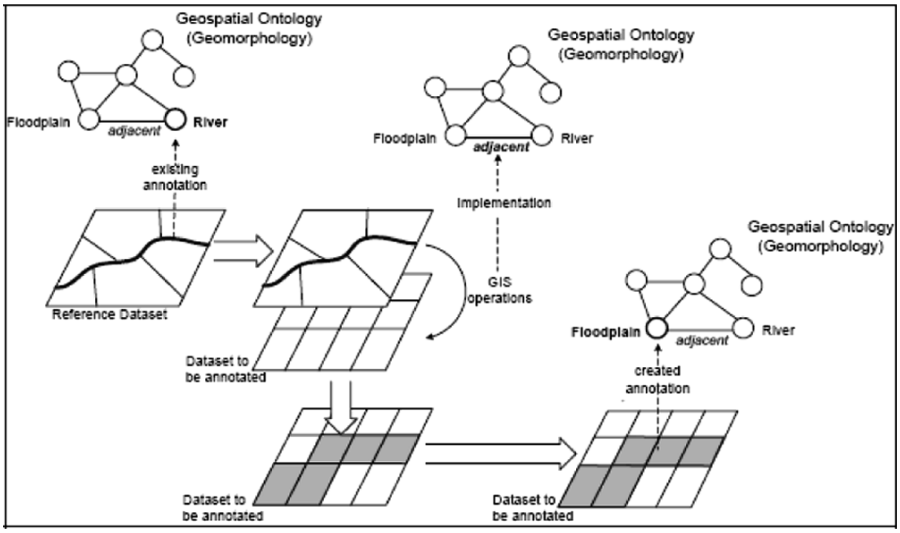


Table 1 Summarisation of the analysed annotation tools

Tool	*Format*	*Ontology*	*Storage*	*Automated*	*Annotation method*	*Annotated data*	*Spatial component*
Embrapa information agency	XML, using Dublin core metadata	No	Relational data base	No	Manual, using natural language	Textual Web pages, videos, images and documents	No
Amaya	XML, RDF	No	Local files	Yes, but very limited	Based on given parameters	Textual Web pages	No
Kim	RDF, OWL	Yes	Local files or in an annotation server	Yes	String matching and ML	Textual Web pages	No
AKTive media	RDF	Yes	Local files	Yes	ML (induction), with continuous manual training	Textual Web pages and images	No
CREAM	RDF, OWL	Yes	Local files or in an annotation server	Yes, with supervised learned	ML (induction) manual training	Textual Web pages, videos and images	No
E-Culture	RDF, OWL, using VRA metadata	Yes		No	Manual, using a structured schema	Images of painting	Yes
OnLocus	XML	Yes		Yes	Geospatial evidences (addresses)	Textual Web pages	Yes
SPIRIT		Yes		Yes	Geospatial evidences	Textual Web pages	Yes
Geodata annotation	XML, using ISO 19115 metadata	Yes		Yes	Spatial methods, string matching	Geographic data	Yes

resources, even the ones that consider the geospatial component. When a visual resource is considered, like a map or a painting, it is necessary to explore its content manually or through the use of specific operations.

5 Conclusions and ongoing work

Geospatial data available on the Web are very useful to answer important questions for production planning and definition of public policies concerning agricultural practices. However, the retrieval of this kind of data is not a trivial task. One solution pointed out in the literature is to associate enhanced annotations to such data, often taking advantage of ontological knowledge. Then, distinct kinds of retrieval solutions may be used to access relevant data. Nevertheless, as shown in Section 4, present annotation mechanisms are centred on text, and content semantics are often lost. Moreover, annotations are usually performed manually for more complex kinds of digital content, such as those used for decision processes in agriculture.

We propose an annotation framework to attack these problems, which supports semi-automatic *Semantic annotations* of various kinds of digital content, directed towards the agriculture context. This framework, under implementation, is part of the WebMAPS project. It relies on four major concepts: the use of authoritative domain ontologies to provide a consensual annotation vocabulary; the adoption of scientific workflows, designed by domain experts, to guide a semi-automatic annotation process; the exploration of spatial information derivable from a given content to help narrow down annotation alternatives; and the availability of catalogs that publish data and annotations, thus helping external users to perform semantic search for content.

As shown in the paper, we have already implemented part of the framework, which is being validated by real case studies and expert users. Our implementation takes advantage of tools available in WebMAPS. Several challenges have still to be considered. First, though we can annotate entire digital objects, and parts of specific kinds of objects (e.g., the time series of our example) we still need to devise workflows that support annotation of parts of objects, especially for multimedia data. For instance, distinct users may select different parts of a satellite image to annotate the phenomena of interest – this raises issues such as annotation storage management, and on associating annotation content to user context. Another issue is the annotation of virtual content – e.g., when users annotate NDVI graphs, it is the underlying series/points that are actually annotated, though users want to annotate the graphs themselves. This is moreover associated with a third challenge: the series are derived from annotated images. Hence, one needs to handle correlations among annotations of primary and derived data. We hope that the use of ontologies will help derive such correlations, by means of inference and ontology manipulation operations,

such as alignment or view generation. We furthermore restrict ourselves to annotations of stored (as opposed to virtual) data, thereby ignoring the second issue for the moment.

Last but not least, ontology management is a topic in itself. Open problems include languages to specify them, mechanisms to manage and generate them, and implementation of efficient operations. Aondê (Daltio and Medeiros, 2008) was developed to meet some of these challenges, but much remains to be done. For more info on open problems, the reader is referred to Euzenat and Shvaiko (2007).

Acknowledgements

The research described in this paper was partially financed by CNPTIA-EMBRAPA, CNPq (WebMAPS project) and the FAPESP-Microsoft Research Virtual Institute (eFarms project).

References

Agosti, M. and Ferro, N. (2007) 'A formal model of annotations of digital content', *ACM Trans. Inf. Syst.*, Vol. 26, No. 1, p.3.

Berners-Lee, T., Hendler, J. and Lassila, O. (2001) 'The semantic web', *Scientific American*, pp.34–43.

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

Chakravarthy, A., Ciravegna, F. and Lanfranchi, V. (2006) 'AKTiveMedia: cross-media document annotation and enrichment', *Fifteenth International Semantic Web Conference (ISWC2006) – Poster*, Georgia, USA.

Daltio, J. and Medeiros, C.B. (2008) 'Aondê: an ontology web service for interoperability across biodiversity applications', *Information Systems*, Vol. 33, Nos. 7–8, pp.724–753.

Daltio, J., Medeiros, C.B., Gomes Jr., L.C. and Lewinsohn, T. (2008) 'A framework to process complex biodiversity queries', *SAC '08: Proceedings of the 2008 ACM Symposium on Applied Computing*, ACM, Fortaleza, Brazil, pp.2293–2297.

Egenhofer, M.J. (2002) 'Toward the semantic geospatial web', *GIS '02: 10th ACM International Symposium on Advances in Geographic Information Systems*, ACM Press, Virginia, USA, pp.1–4.

Euzenat, J. and Shvaiko, P. (2007) *Ontology Matching*, Springer-Verlag, New York, Inc.

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

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*, Vol. 12, No. 4, pp.352–367.

Greenberg, J., Spurgin, K. and Crystal, A. (2006) 'Functionalities for automatic metadata generation applications: a survey of metadata experts' opinions', *Int. J. Metadata, Semantics and Ontologies*, Vol. 1, No. 1, pp.3–20.

Gruber, T.R. (1993) 'A translation approach to portable ontology specifications', *Knowl. Acquis.*, Vol. 5, No. 2, pp.199–220.

Handschuh, S. and Staab, S. (2002) 'Authoring and annotation of web pages in CREAM', *WWW '02: Proceedings of the 11th International Conference on World Wide Web*, ACM Press, Hawaii, USA, pp.462–473.

Hollink, L. (2006) *Semantic Annotation for Retrieval of Visual Resources*, PhD Thesis, Vrije Universiteit Amsterdam.

Hollink, L., Schreiber, G., Wielemaker, J. and Wielinga, B. (2003) 'Semantic annotation of image collections', *Workshop on Knowledge Markup and Semantic Annotation – KCAP'03*, Florida, USA, pp.0–3.

IBGE (2008) *Geographic and Statistical Brazilian Institute (IBGE)*, IBGE/USP. <http://www.ibge.gov.br/english/>. Accessed in: 25 March.

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

Klien, E. (2007) 'A rule-based strategy for the semantic annotation of geodata', *Transactions in GIS*, Vol. 11, No. 3, pp.437–452.

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

Kondo, A.A., Medeiros, C.B., Bacarin, E. and Madeira, E.R.M. (2007) 'Traceability in food for supply chains', *3rd International Conference on Web Information Systems and Technologies (WEBIST)*, INSTICC, Barcelona, Spain, pp.121–127.

Lunetta, R., Johnson, D., Lyon, J. and Crotwell, J. (2003) 'Impacts of imagery temporal frequency on landcover change detection monitoring', *Remote Sensing and Environment*, Vol. 89, No. 4, pp.444–454.

Lutz, M., Spradob, J., Klien, E., Schubertd, C. and Christ, I. (2008) 'Overcoming semantic heterogeneity in spatial data infrastructures', *Computers and Geosciences*, in Press.

Macário, C.G.N., Medeiros, C.B. and Senra, R.D.A. (2007) 'The WebMAPS project: challenges and results', *IX Brazilian Symposium on GeoInformatics – Geoinfo 2007*, Campos do Jordão, Brazil, pp.239–250.

Mangold, C. (2007) 'A survey and classification of semantic search approaches', *Int. J. Metadata, Semantics and Ontology*, Vol. 2, pp.23–34.

Mariotte, L., Medeiros, C.B. and Torres, R. (2007) 'Diagnosing similarity of oscillation trends in time series', *International Workshop on Spatial and Spatio-Temporal Data Mining – SSTDM*, Nebraska, USA, pp.643–648.

Medeiros, C.B., Pérez-Alcazar, J., Digiampietri, L., Pastorello Jr., G.Z., Santanchè, A., Torres, R.S., Madeira, E. and Bacarin, E. (2005) 'WOODSS and the web: annotating and reusing scientific workflows', *SIGMOD Record*, Vol. 34, No. 3, pp.18–23.

Nogueras-Iso, J., Zarazaga-Soria, F., Bjar, R., lvarez, P. and Muro-Med, P. (2005) 'OGC catalog services: a key element for the development of spatial data infrastructure', *Computers and Geosciences*, Vol. 31, pp.199–209.

Ontotext Lab (2007) *The KIM Platform: Knowledge and Information Management*, <http://www.ontotext.com/kim/index.html>

Reeve, L. and Han, H. (2005) 'Survey of semantic annotation platforms', *SAC '05: 2005 ACM Symposium on Applied Computing*, ACM, New Mexico, USA, pp.1634–1638.

Shadbolt, N., Berners-Lee, T. and Hall, W. (2006) 'The semantic web revisited', *IEEE Intelligent Systems*, Vol. 21, No. 3, pp.96–101.

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', *II Digital Libraries Workshop*, Brazil, pp.51–56 (in Portuguese).

Tsalgatidou, A., Athanasopoulos, G., Pantazoglou, M., Pautasso, C., Heinis, T., Gronmo, R., Hoff, H., Berre, A-J., Glittum, M. and Topouzidou, S. (2006) 'Developing scientific workflows from heterogeneous services', *SIGMOD Record*, Vol. 35, No. 2, pp.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*, Vol. 4, No. 1, pp.14–28.

Van der Aalst, W.P. and Ter Hofstede, A. (2005) 'YAWL: yet another workflow language', *Information Systems*, Vol. 30, No. 4, pp.245–275.

VRADSC (2007) *VRA Core 4.0*, <http://www.vraweb.org/index.html>

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

Weibel, S., Godby, J., Miller, E. and Daniel, R. (1995) *OCLC/NCSA Metadata Workshop Report*.