Annotating data to support decision-making: a case study

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

Jefersson A. dos Santos,
Claudia Bauzer Medeiros,
Ricardo da S. Torres
Institute of Computing - P.O.Box 6176
University of Campinas - UNICAMP
13083-970, Campinas, SP, Brazil
{jsantos, rtorres, cmbm}@ic.unicamp.br

ABSTRACT

Georeferenced data are a key factor in many decision-making systems. However, their interpretation is user and context dependent so that, for each situation, data analysts have to interpret them, a time-consuming task. One approach to alleviate this task, is the use of semantic annotations to store the produced information. Annotating data is however hard to perform and prone to errors, especially when executed manually. This difficulty increases with the amount of data to annotate. Moreover, annotation requires multidisciplinary collaboration of researchers, with access to heterogeneous and distributed data sources and scientific computations. This paper illustrates our solution to approach this problem by means of a case study in agriculture. It shows how our implementation of a framework to automate the annotation of geospatial data can be used to process real data from remote sensing images and other official Brazilian data sources.

Categories and Subject Descriptors

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

Keywords

Semantic Annotation, Geospatial data, Remote Sensing Image Classification, Geospatial standards

1. INTRODUCTION

Decision making systems based on georeferenced data have been considered an important tool in a wide range of domains, from studies on global warming to those on urban planning or consumer services. Although this kind of data corresponds to about 80% of available data on the web [25], they usually are not ready for use. In most cases, they have to be analyzed and interpreted according to user context and application. These interpretations produce new information, which is often never recorded. Hence, every time a user wants such information, the data have to be interpreted again. Interpretations not only help understanding data – they can also support a wide range of retrieval possibilities. However, more often than not, retrieval is based on keyword matching and can lead to of irrelevant files.

Consider, for instance, a satellite image which has already been georeferenced. Normally, this image is delivered together with some kind of metadata, which is used to qualify its contents – e.g., date taken, image quality, satellite that produced it, and so on. Distinct users may want to analyze this image according to their decision needs. For instance, environmentalists may be interested in studying deforestation patches, demographers will look for urban patterns, transportation experts may be concerned with viable corridors. Each such need requires specific kinds of processing to detect the objects of interest. Moreover, each user will possibly attach descriptive text to the image, e.g., to help subsequent retrieval, or support work with other colleagues. In many cases, several experts will collaborate in such annotation procedures – e.g., those concerned with environmental issues will contribute with distinct expertises, such as identification of type of vegetation, knowledge of the region, or recognition of additional impacting factors. This further complicates the annotation process, since such collaborations may happen across continents, and people may employ distinct vocabularies or have specific annotation usages. Once the image is thus annotated, retrieval is based on either keyword matching or text-based techniques. However, such annotations are based on natural language, which complicates finding the adequate information.

Semantic annotations, which are a combination of metadata labels and ontology items, have been adopted to attack such problems, enhancing information sharing. They are useful to store the data interpretations, providing a description free of ambiguity. Moreover, they support search based on semantic concepts. However, data annotation is a hard task and prone to errors. Annotation of geospatial data, in particular, requires collaboration of multiple experts, and is time consuming. Most related research focus on annotation of textual resources. When other resources are treated, like images, their are manually annotated by the user.

In this work we present our approach to automate the semantic annotation of different kinds of georeferenced data, using a case study in agriculture. In particular, our work is geared towards any kind of georeferenced data file – e.g., spatio-temporal series, satellite or radar images, maps, networks. Our case study shows how the framework we developed for semantic annotation of geospatial data [19, 20] can

GIR'10 18-19th Feb. 2010, Zurich, Switzerland

be used to accelerate the annotation of satellite images, a prime data source, for decision support in agriculture. It relies on scientific workflows to drive the annotation process, in a context- and domain-dependent manner. As will be seen, our contributions therefore support decision-making in speeding up the annotation process. Moreover, since our annotations are based on ontologies, they can serve as the basis for semantic retrieval.

2. THE SEMANTIC ANNOTATION PROCESS

2.1 Semantic Annotations

Semantic annotations combine concepts of metadata and ontologies: metadata fields are filled with ontology terms, which are used to describe these fields. We define semantic annotations as follows [20]:

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

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

Annotation Schema and Content. An annotation/semantic annotation has a *schema* and a *content*. The schema is its structure, specified through its metadata fields; the content corresponds to the values of these fields.

While annotation units describe data using natural language, semantic annotations units use ontology terms and can be processed by a machine. We point out that annotation units are specified as tuples, similar to an RDF structure. This helps their subsequent storage and reuse. Users, however, manipulate them in more friendly formats.

2.2 The Annotation Process

The data annotation process is a hard task and prone to errors. To automate this process is not easy. However, geospatial data have some important features that can make this automation a reality. First, availability of coordinates can be used to speed up the annotation process. Second, for very many kinds of geospatial data, there are repeatable core procedures that can be specified by experts to produce annotations. Such procedures can be subsequently tailored to meet context–specific annotation demands.

We took advantage of these features to define our annotation scenario. First, domain experts need to predefine core annotation procedures for each kind of georeferenced data source (e.g., telecommunication networks, satellite images, sensor time series). Focusing on repeatability, sharing, reuse and adaptation to new contexts, we chose to store these procedures as scientific workflows. 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. Each workflow contains information on the annotation schema and ontologies to be used, the operations to perform and how to store the generated annotations. The entire process is supported by our framework, whose main components have been implemented.

Figure 1, adapted from [20], gives an overview of the annotation process supported by our framework, which has three main steps: selection of an 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 [9]. 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. Each workflow activity performs one annotation task, executed by invocation of Web services, through a workflow engine. During this process, the annotation units are presented for user validation, usually a domain expert. 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 an XML database, where they can be used for information retrieval, e.g. using XQuery statements.

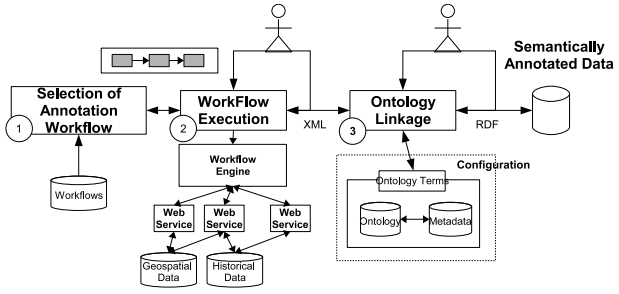

Figure 1: The Semantic Annotation Process

Configuration of the Framework.

The framework has been designed to be generic for different domains. Hence, it is necessary to perform a set of activities to customize the annotation process, such as specification of the annotation schema to be adopted, design of annotation workflows and selection of ontologies, and their terms, to be used for content description. Once the workflows are specified, it is also necessary to implement the workflow activities to produce the desired annotation units. Configuration must be jointly performed by computer scientists and domain experts. Since this is also a hard and time consuming task, it should only be undertaken if experts expect that a given kind of geospatial data source will be frequently annotated for decision support. Our case study concerns on such example, of annotating satellite images for agriculture.

2.3 Implementation Aspects

The framework is being implemented in JAVA, since this language provides several APIs that can facilitate our work. It also is centered on XML files, which facilitates data exchanging. Annotation workflows are specified using WOODSS, a workflow tool [21]. Since WOODSS does not have a native execution engine, we adopted YAWL for this task [30].

Each activity in the workflow is linked to a specific web service and executed on a Tomcat server. Our data repositories have been implemented using real data, stored in the PostGreSQL and PostGIS database system. We have implemented a set of basic services using JAVA and the Axis2 framework. These services encapsulate annotation steps. Our case study details of the implemented services.

3. A CASE STUDY - AGRICULTURAL PLANNING IN BRAZIL

This section presents a case study that concerns the use of our annotation framework to handle remote sensing images, for agricultural planning. Several factors influence crop yield and estimates - e.g., soil, relief, climate, and crop managements practices. Such factors are also used for *agriculture zoning*. This term refers to the partitioning of a given region in subregions *zones* to indicate which crop should be planted where and when. Prediction estimates and zoning are the basis for Brazilian government policies to finance agricultural activities.

Remote sensing images are intensively used for agricultural planning and crop monitoring, providing a basis for decision making. They are used, for instance, to identify the extent of a given plantation, or to detect signs of deterioration in plant health (e.g. due to pests or excess of water). Agricultural experts have to interpret these data to obtain the desired information.

We now show the process of semi-automatic generation of annotations for a remote sensing image of Monte Santo de Minas county, located in one of the Brazilian regions with the highest coffee productivity index. Figure 3 (left side) shows a SPOT satellite image of this area, taken on August 2005.

3.1 Configuring the Framework

Defining an Annotation Schema.

Since we are concerned with geospatial data for decision making in agriculture , the metadata schema chosen is based on FGDC's geospatial metadata standard [1], a general purpose and open standard. We extended this standard to provide additional fields for agricultural issues, such as information on crop production. Workflow execution produces information to fill each one of these fields.

Selecting Ontologies.

Experts were requested to choose the appropriate ontology terms that could be used to produce each semantic annotation unit, i.e., to fill metadata fields. This is performed by the ontology extraction tool. For each ontology chosen, all terms are extracted, tagged indexed in one of the framework's repositories, for future retrieval. As in the linkage step, possible terms are associated with each annotation field. For example, for annotation schema field *crop*, some possible terms are *"www.lis.ic.unicamp.br/ont/agricZoning.owl #coffee"* or *"www.lis.ic.unicamp.br/ont/agricZoning.owl #sugarCane"*, all extracted from POESIA [13], an ontology for agricultural zoning.

Designing and Developing Workflows.

Experts have to define annotation processes, which are tailored to each kind of geospatial data source. For instance, for the same region, there are distinct annotation processes for files containing satellite images, photos, time series (e.g. rainfall) or crop characteristics.

Figure 2 presents the core workflow for annotation of a remote sensing image, describing the main tasks to be performed.

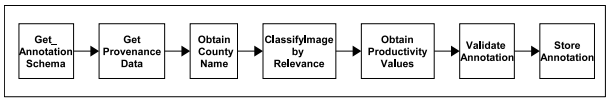

Figure 2: The core workflow for annotation of Remote sensing images

3.2 Producing Annotations - Workflow Execution

We adopted the YAWL workflow engine for workflow execution. The first step of execution, after the retrieval of the annotation schema, is to obtain provenance data (information like satellite name, spatial resolution, acquisition date and information on coordinates). This information is extracted from the image's header. The next activity to be executed is "Obtain County Name" (see figure 2). This is performed by the *CountyNameService*, which is invoked by YAWL. We implemented this Web service to return the name of a county, for a given group of one or more pair of coordinates. This service accesses an IBGE[2] data source, which contains information of all counties, cities and states in Brazil. If the image's coordinates show that it covers more than one county, all their names are returned.

After this, an image classification tool is invoked of service. This tool uses image processing techiniques to identify polygons within the input image that match a given input pattern. Identification and recognition of crop patches in a remote sensing image is a painstaking process, and thus this tool is of great help to users – see [10]. This tool receives as input an annotated image pattern and an image I to be classified. It provides as result a new image I', composed by a set of polygons which were identified as having the pattern. As both images I and I' are georeferenced, it is possible to overlay them to check the result.

In more detail, the input pattern is processed to extract its texture and spectral features, which are encoded into pattern *descriptors*. Next, I is processed to identify similar patterns. In other words, the classification tool uses image descriptors to encode texture and spectral features from I combining them with information from the annotated input pattern. Next, the tool segments I, identifying areas of interest (here, areas being occupied by the same crop). The segmented image, in raster format, is converted into a vector representation to be further processed – image (I').

Figure 3 illustrates this process. Given the input pattern for coffee (small box on the left bottom), the tool divides the image into small parts (considering parameters like usual area size for specific crops) and extracts their image features. These small images are then sorted according to their similarity to the features of the input pattern. The most similar

[1]FGDC-STD-001-1998. Content Standard for Digital Geospatial Metadata/Federal Geographic Data Committee

[2]Geographic and Statistical Brazilian Institute, official organ in Brazil for all aspects concerning territorial and statistical data on the country

parts are then converted into polygons, which are presented as the result to the user for validation (figure at the right).

The tool was able to correctly identify more than 66% percent of the coffee parcels (gray parcels) which had been previously identified by field trips. This is considered an excellent result, mainly when compared to MaxVer, a traditional classification method and the most frequently used one [23].

This identification process can be performed for different inputs for the same image. In this example, for instance, there are also areas in which sugar cane is planted and that can be identified using the same tool with sugar cane patterns for input. We point out that this process is spatially sensitive. A given image pattern may be associated with different kinds of crop, depending on the region being examined, and image resolution.

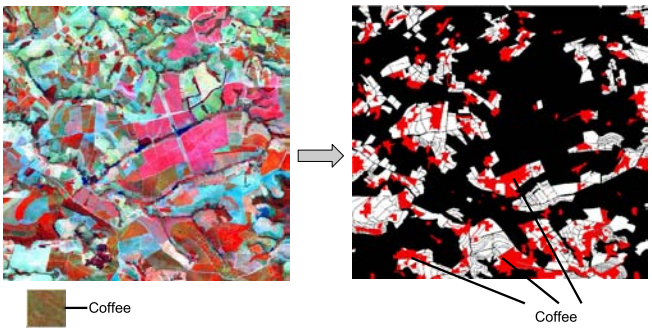

Figure 3: The Image Classification Process

The next workflow activity (see figure 2) is to obtain historical productivity values for this crop and region. The invocation of a service called *ProductivityService* returns information for a given crop, for a specific region and year. The returned information includes productivity values retrieved from another IBGE database, which maintains official information for different crops, grouped by geographic region – macro and micro region, state and county – and by year.

Figure 4 shows an excerpt of this historical information, for several counties. It indicates that Monte Santo de Minas produced 963 kg/ha of coffee for the year of 2005 – the year in which the image was taken.

Producer Counties, Micro and Macro Regions	Area reserved for crop (ha)	Harvested Area (ha)	Yield (t)	Average production (kg/ha)
São Sebastião do Paraíso	87 185	87 185	69 810	800
...				
Jurucaia	5 280	5 280	3 986	754
Monte Belo	4 500	4 500	3 321	738
Monte Santo de Minas	8 960	8 960	8 637	963
Muzambinho	7 000	7 000	4 802	686
...				

Figure 4: Historical Productivity Values for Coffee in Monte Santo de Minas

Once the annotation is validated, the annotation units are transformed into semantic annotation units, using the ontology terms selected during the configuration phase.

Figure 5 presents part of these annotations. This corresponds to the extended information of the schema, considering agriculture issues. For example, the image contains coffee parcels, which is identified by the pair <*crop*>, <*rdf resource*>. For more details on the adopted annotation schema, the annotation units and the ontologies used, see [20].

```
<crop>
<rdf:Bag>
<rdf:li
rdf:resource="http://www.lis.ic.unicamp.br/ont/agricZoning.owl#Coffee"/>
<rdf:li>Coffee</rdf:li>
</rdf:Bag>
</crop>
```

Figure 5: Semantic annotation unit generated for the Monte Santo de Minas remote sensing image

Experts finally have to validate the semantic annotations created. Based on these annotations, a Brazilian government expert may confirm the spatial extent of a crop, and compute 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).

3.3 Decision Support and Annotations

There are several scenarios in which a set of semantically annotated images can be used in decision support. Let us examine two of these scenarios, in agriculture. The starting hypothesis is that there exists a database containing images which have already been annotated by our framework. The first situation concerns comparative analysis of rural areas, e.g., to detect relationships between multiple crops/yields. For instance, experts may want to derive rules concerning coincidence of different types of culture in a given area, such as "*if coffee then sugar cane*" has a low probability for county-based analysis in Brazil but "*if coffee then beans*" has a high probability in Brazil, as a common practice to ensure advantages for coffee crops such as nitrogen fixation and erosion control.

This kind of rule is hard to mine from standard yield statistics, but may be derived in a straight foward way, using basic (value-based)data mining algorithms, from our semantically annotated images. The biggest difficulty in mining tables published by official sources is to derive pattern that depend on geospatial features. Annotated images provide the needed spatial cluster, and thus one has to mine for co-existence of semantic annotation units associated to images that concern the same counties.

Another interesting issue involves evolution of cultures for a given set of regions (spatio-temporal analysis). In this case, experts base their analysis on image time series – i.e., studying changes in crops planted, by considering a large set of images for the same region. This can give margin to several kinds of diagnostics, e.g., identifying economically-based phenomena. For instance, in the 1980's, Brazil launched a strong program to produce fuel from sugar cane (alcohol). This fostered intensive research in this field, and, moreover, completely changed the crop cover in many areas in the country, as farmers took advantage of government subsidies to move from other crops to sugar cane. During the 90's, such subsidies disappeared and so did several sugar cane plantations. With the emergence of renewable bioenergy, many farmers are (re)turning to sugar cane. This evolution in vegetation cover can be observed via image analysis, with drastic changes in a very short period of time. Agricultural census and statistics can also provide such information, but

in a less timely manner. Still another phenomenon that can be detected by temporal analysis is the spread of diseases (many times first identified via image-based patterns).

These kinds of situation are common in image-based decision making in agriculture, and can be accelerated by processing annotations rather than the images themselves. Finally, many procedures that require the study of spatio-temporal evolution of phenomena based on image series can also be made more flexible by taking advantage of annotations.

4. OUR APPROACH FOR GIR

Information retrieval (IR) traditionally is performed through the specification of a set of words which represent the semantics of the desired information. These words are compared to a set of indices – a collection of selected words or concepts used to describe the resources and associated to them – looking for matchings. This approach, called *keyword-based retrieval mechanism*, has been used for several years, and although IR has evolved, indices continue to be essential.

According to [3], the quality of the retrieval task is greatly affected by the user interaction with the system. Thus, the information description and the specification of the data search are very important for the success of IR. When considering geospatial data, this can be a hard task, mainly to: (i) data – satellite images, maps, graphs and others – may contain essential information that can not be described using traditional keyword descriptions; (ii) the georeferenced information usually is not considered during the retrieval operation. Hence, for this kind of data, different IR approaches, such as content-based and semantic retrieval, have been applied. Taking advantage of all of them, we propose a combination of these different techniques, which will be described in the following.

4.1 Content-based Retrieval

Performing automatic recognition of region images, regardless the used method, scarcely generate a complete satisfactory solution. The place or the age of the crops, for example, may hinder the recognition process. In these cases the spectral response and the texture patterns to the same kind of crop can be different. A crop can be planted in different ways and this factor, allied to the different phases of the plants, tends to create a distinction between regions of the same class [5]. In Content-based Image Retrieval (CBIR) some of these problems are similar.

CBIR is centered on the notion of image similarity – given a database with a large number of images, the user wants the most similar images to a query pattern (normally an example image). In general, the retrieval process is based on characterizing visual features (such as color, texture, shape) by using descriptors. A descriptor can be characterized by two functions: *feature vector extraction* and *similarity computation*. The feature vectors encode image properties, like color, texture, and shape. Therefore, the similarity between two images is computed as a function of their feature vectors distance [6].

Different image descriptors encoding different even the same image properties have been proposed to support image retrieval [24, 8]. These descriptors, generally, are combined in order to meet users' perception. In a higher level, descriptors encoding different properties can be combined to support different perception criteria of different users. Many of combinations strategies are based on using weights, which are supposed to asses the importance of a given descriptor. Some approaches applies machine learning techniques such as GA [26], GP [7] and SVM [27] to combine descriptors and improve retrieval results.

Recently, some descriptors for RSI purposes has been proposed. Tusk et. al. [28] presented algorithms that allow for automatic selection of features for region and tile similarity searches applying relevance feedback. Samal et. al. [22] proposed an RSI descriptor, called SIMR (Satellite Image Matching and Retrieval). SIMR computes spectral and spatial attributes of the images using a hierarchical representation. A unique aspect of this descriptor is the couples of second-level spatial autocorrelation with quad tree structure.

4.2 Retrieval Based on Semantic Annotations

Semantic annotations rely on ontology terms, which are referenced via their URIs. This opens new perspectives on retrieval of georeferenced data sources. To start with, our annotation units can be used as a basis for standard keyword-based retrieval techniques (since they are based on filling the schema with natural language expressions). The main difference, in our case, is that annotation generation is guided by workflows,thereby speeding up annotation procedures and decreasing the amount of errors that can occur when users are given the task of manually filling metadata fields (e.g., [12], [29]). The work of [2] is an example of the difficulties of performing search based on categorical metadata. However, once annotation units are transformed into semantic annotation units, one can start taking advantage of advances in the semantic web. For instance, search can be performed not only on the ontology terms themselves, but also on reasoning over these terms, using ontology axioms to derive new information.

Also, one can take advantage of several ontology-based operations to extend search parameters. For example, consider an ontology involving territorial divisions (such as the one used by us considering Brazilian IBGE names of counties, regions, macro-regions, etc). Suppose that some expert has built another geo-ontology of place names for localities of interest, that involve spatial relationships - see [1] for a proposal on constructing such ontologies. This second ontology contains not only place names, but geographic footprints. Then, one can search for data sources that "contain coffee parcels" and "are situated in Brazilian counties that are adjacent to Itamogi". The geo-ontology built by the expert for her own application needs will indicate that "Monte Santo de Minas" is adjacent to "Itamogi". If our IBGE-based ontology is aligned with the expert's geo-ontology, then a new extended ontology can be built in which "Monte Santo de Minas" is found to denote the same concept in both ontologies, and moreover adjacent counties can be derived.

5. RELATED WORK

In geographic applications, annotations should consider the spatial component. Hence, the geospatial annotation process should be based on geospatial evidences – those that conduct to a geographic locality or phenomenon. E-Culture [15, 14], OnLocus [4], SPIRIT [16] and Semantic Annotation of Geodata [17] are approaches that consider the spatial component for the annotation of digital content, always having in mind supporting flexible information retrieval.

E-Culture is a project that focuses on semantic annotation and searching of images of paintings, considering spatial properties within an image. Though not related to geographic issues, its approach is interesting insofar as it points out certain needs for identifying spatially related objects within an image, using concepts which can be adapted to georeferenced images. The project adopts an annotation schema based on VRA Metadata with at last 4 terms – agent, action, object and recipient – where each object is associated to terms of WordNet, AAT, ULAN and Iconclass ontologies. For spatial features the project uses the WorldNet ontology, which gives absolute positions of an object in the painting, and the SUMO ontology, for spatial relations. During the search process, concepts like class equivalence and ontology alignment are considered, to increase the searching coverage. The annotation process is manual. Similar to E-Culture, we also take advantage of operations on spatial ontologies to augment annotation capabilities. However, our annotation process is performed in a semi-automatic way, in which the user just has to confirm if the terms provided are correct.

OnLocus consists of a geographic information retrieval approach for recognizing, extracting and geotagging of geospatial evidences of local features such as address, postal codes and phone numbers available on the Web. Through these evidences it is possible to correlate the content of a Web page, or part of it, to an urban geographic location. This approach is supported by the OnLocus urban space ontology. Hence, search machines may use this information to retrieve pages of urban services and activities in a specific locality or near it. Unlike our work, OnLocus is centered on annotating Web pages and is applied to urban applications; similar to our approach, it also relies heavily on ontologies to support retrieval.

Similar to OnLocus, SPIRIT – Spatially-Aware Information Retrieval on the Internet – 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. During the process of adding geographical identification metadata to pages being analyzed (geotagging process), metadata can be associated with Web sites or images, and also with other 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. Different from them, we extend this to other kinds of media. We also use information that cannot be obtained directly from the resources, but are equally important, such as data provided by other (related sources). In our case study, for instance, annotation content is first derived from the coordinates, which are then used to extract additional content from other data sources. If this imposes an additional burden on the annotation process, on the other hand it supports flexibility in annotation, since experts can define workflow tasks according to their needs.

Finally, the work on Semantic Annotation of Geodata proposes 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 approach considers multiple ontologies defined by homogeneous themes (such as hydrology, geology, ecology, transportation planning) [18]. The idea is to generate information using spatial analysis methods – for example, to identify if an area is candidate to flooding. Like this work, we use geographic ontologies, and also some spatial relations, during our annotation process. However, we do not base the whole annotation process on them. Moreover, we also tailor annotations to the kind of content.

Satellite images are a very important (non-structured) source of georeferenced data for decision support - such as the situations presented in the introduction, or in our case study. Hence, our workflows frequently have to deal with image issues, in particular vectorization, segmentation and classification. However, there is still no satisfactory method to classify these images. Traditional classification methods are based on pixel analysis. The one most frequently used, MaxVer [23], is not so effective. Our work in this direction is also promising, and shows that effective information retrieval requires a combination of several kinds of information extraction techniques.

6. CONCLUSIONS

Geospatial data are a basis for several decision making systems. However, these data have to be interpreted to be used. This is a time-consuming task that has to be performed every time the data is required. The absence of approaches to efficiently store these interpretations leads to problems such as rework and difficulties in information sharing. To alleviate this, we proposed an approach to record the geospatial data interpretation based on semantic annotation. This approach is supported by a framework for semi-automatic annotation of geospatial data.

This paper discussed our approach for annotation of geospatial data, which is consonant with efforts on the Geospatial Semantic Web [11]. We described a case study in agriculture, which is being used to annotate remote sensing images relying on official Brazilian government data sources, and uses well known ontologies. As shown in the paper, our work allows the invocation of several kinds of tools to compose a complex annotation, which can be used as a basis for decision support. In particular, several tools are invoked to generate parts of an annotation, taking advantage of spatial coordinates (for instance, to get information on county), spatially-sensitive information (vegetation cover) and image texture and color. Thanks to annotation workflows, annotations can grow in complexity and be tailored to specific user needs.

Ongoing work includes several directions. We are designing annotation workflows for several kinds of data sources. Also, we intend to refine the annotation process so that parts of a data source can be annotated (e.g., polygons in an image, or subsets of a time series). An important step in our research is to set up the validation of our proposal. Since this work is being conducted within a large project in agricultural data management, we are designing the experiments for validation together with domain experts, who are providing geospatial data sources to be annotated. We are furthermore continuing our design and development of new services to be invoked by annotation workflow tasks. Finally, we are also extending our image processing tools, e.g., to include other image descriptors.

7. ACKNOWLEDGEMENTS

This work was partially funded by the Fapesp-Microsoft Research Virtual Institute (eFarms project), and by CNPTIA - Embrapa. Additional support was provided by the Brazilian funding agencies Fapesp, CNPq and CAPES.

8. REFERENCES

[1] A. I. Abdelmoty, P. Smart, and C. B. Jones. Building place ontologies for the semantic web:: issues and approaches. In *GIR '07: Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 7–12. ACM, 2007.

[2] R. Albertoni, A. Bertone, and M. D. Martino. Semantic analysis of categorical metadata to search for geographic information. In *Database and Expert Systems Applications, International Workshop on*, pages 453–457, Los Alamitos, CA, USA, 2005. IEEE Computer Society.

[3] R. Baeza-yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[4] 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.

[5] S. Cordero-Sancho and S. A. Sader. Spectral analysis and classification accuracy of coffee crops using landsat and a topographic-environmental model. *Int. J. Remote Sens.*, 28(7):1577–1593, 2007.

[6] R. da S. Torres and A. X. Falcão. Content-Based Image Retrieval: Theory and Applications. *Revista de Informática Teórica e Aplicada*, 13(2):161–185, 2006.

[7] R. da S. Torres, A. X. Falcão, M. A. Gonçalves, J. P. Papa, B. Zhang, W. Fan, and E. A. Fox. A genetic programming framework for content-based image retrieval. *Pattern Recognition*, 42(2):283–292, 2009.

[8] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):1–60, April 2008.

[9] S. R. de Sousa. A semantic approach to describe geospatial resources. In *3rd International Workshop on Semantic and Conceptual Issues in GIS (SeCoGIS 2009)*, volume LNCS 5833, page 327Ű336, November 2009.

[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] M. J. Egenhofer. Toward the semantic geospatial web. In *Proc. of the ACM GISt'02*, pages 1–4, 2002.

[12] FAO. FAO - GeoNetwork, 2008. http://www.fao.org/geonetwork/srv/en/main.home.

[13] 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.

[14] L. Hollink. *Semantic Annotation for Retrieval of Visual Resources*. PhD thesis, Vrije Universiteit Amsterdam, 2006.

[15] 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.

[16] 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.

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

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

[19] 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.

[20] C. G. N. Macário, S. R. Sousa, and C. B. Medeiros. Annotating geospatial data based on its semantics, 2009. Accepted for publication. 17th ACM SIGSPATIAL Conference, 2009. Seattle.

[21] 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.

[22] A. Samal, S. Bhatia, P. Vadlamani, and D. Marx. Searching satellite imagery with integrated measures. *Pattern Recogn.*, 42(11):2502–2513, 2009.

[23] R. Showengerdt. *Techniques for Image Processing and Classification in Remote Sensing*. Academic Press, New York, 1983.

[24] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

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

[26] Z. Stejic, Y. Takama, and K. Hirota. Relevance feedback-based image retrieval interface incorporating region and feature saliency patterns as visualizable image similarity criteria. *IEEE Transactions on Industrial Electronics*, 50(5):839–852, Oct. 2003.

[27] S. Tong and E. Y. Chang. Support vector machine active learning for image retrieval. In *Proceedings of 9th ACM international conference on Multimedia*, pages 107–118, New York, NY, USA, 2001.

[28] C. Tusk, K. Koperski, S. Aksoy, and G. Marchisio. Automated feature selection through relevance feedback. In *Geoscience and Remote Sensing Symposium, 2003. IGARSS '03. Proceedings. 2003 IEEE International*, volume 6, pages 3691–3693 vol.6, July 2003.

[29] USA.Gov. Geodata.gov - US Maps & Data, 2009. http://gos2.geodata.gov/wps/portal/gos.

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