Anotação Semântica de Dados Geoespaciais

Este exemplar corresponde à redação final da Tese devidamente corrigida e defendida por Carla Geovana do Nascimento Macário e aprovada pela Banca Examinadora.

Campinas, 16 de dezembro de 2009.

Profa. Dra. Claudia Maria Bauzer Medeiros
IC - UNICAMP (Orientadora)

Tese apresentada ao Instituto de Computação, UNICAMP, como requisito parcial para a obtenção do título de Doutora em Ciência da Computação.

Substitua pela ficha catalográfica

(Esta página deve ser o verso da página anterior mesmo no
caso em que não se imprime frente e verso, i.é., até 100 páginas.)

Substitua pela folha com as assinaturas da banca

Anotação Semântica de Dados Geoespaciais

Carla Geovana do Nascimento Macário

Dezembro de 2009

Banca Examinadora:

- Profa. Dra. Claudia Maria Bauzer Medeiros IC - UNICAMP (Orientadora)

- Profa. Dra. Cristina Dutra de Aguiar Ciferri ICMC - USP

- Prof. Dr. Clodoveu Augusto Davis Júnior DCC - UFMG

- Profa. Dra. Marilde Terezinha Prado Santos DC - UFSCar

- Profa. Dra. Ariadne Maria Brito Rizzoni Carvalho IC - UNICAMP

- Prof. Dr. Cesar A. C Teixeira DC - UFSCar (suplente)

- Prof. Dr. Jansle Vieira Rocha FEAGRI - UNICAMP (suplente)

Dedicatória

Dedico este trabalho ao meu pai e à minha madrinha.
Grandes incentivadores... Grandes perdas sofridas...

Agradecimentos

À Profa. Claudia, por ter me aceitado como sua aluna, por todos os conselhos, oportunidades, ensinamentos, apoio e incentivo, e pela confiança em mim e no meu trabalho. Muito obrigada.

À minha família, que sempre me apoiou e me incentivou em tudo que fiz. Ao meu marido Alexandre, pela compreensão e pelo apoio nos momentos mais difícies. Aos meus filhos Isabella e Eduardo, pela felicidade de ser sua mãe, o que me motiva a nunca desistir. À minha mãe, que mesmo com todas as suas perdas, nunca deixou de me confortar. Ao meu pai, onde quer que ele esteja, sei que está orgulhoso. À minha irmã, por me ouvir; e, à minha sogra, por sempre me ajudar a tornar tudo possível.

À Embrapa, pela oportunidade de desenvolver este trabalho. Em especial agradeço aos colegas, Sílvio, Sílvia, Sérgio, Stanley, Angélica, Goretti, Leila, Kléber, José Ruy, Assad, Álvaro, Jardine e Ricardo, pelo apoio e pela torcida. Também agradeço, ao Rubens, pesquisador do CEPAGRI, cujas sugestões e informações foram essenciais para o desenvolvimento deste trabalho.

A todos os meus colegas do LIS. Um agradecimento especial ao Alan, parceiro em várias atividades e que me ajudou sempre que precisei. Também agradeço ao Gilberto, Senra, Nielsen e Evandro, que estiveram presentes nos momentos difícieis. Aos parceiros de trabalhos, Sidney, Jefersson e Profs. Ricardo e Edmundo, obrigada pelas sugestões e contribuições. Aos demais colegas, Lin, Joana, Celso, Jaudete, Andréia, Rodrigo, Bruno, Fábio, Otávio e todos os outros, obrigada por tudo. Vou sentir muita falta de todos!

Ao meu amigo Marcos Chaim por todo apoio e incentivo antes e durante o meu doutorado.

Ao IC, pela oportunidade. Ao professor Cid, pelo conselho acertado, à professora Ariadne, pela torcida, aos professores Tomaz, Flávio, Dahab, Rodolfo, Islene, Julio, pelas conversas e conselhos úteis. Aos funcionários do IC, aqueles que fazem tudo acontecer. E a todos que de alguma forma fazem parte da Unicamp.

Finalmente, agradeço às agências de fomento à ciência e tecnologia – FAPESP, CNPq e CAPES – e ao Instituto Virtual Microsoft Research - Fapesp (projeto EFarms).

Resumo

Dados geoespaciais constituem a base para sistemas de decisão utilizados em vários domínios, como planejamento de trânsito, fornecimento de serviços ou controle de desastres. Entretanto, para serem usados, estes dados precisam ser analisados e interpretados, atividades muitas vezes trabalhosas e geralmente executadas por especialistas. Apesar disso estas interpretações não são armazenadas e quando o são, geralmente correspondem a alguma informação textual e em linguagem própria, gravadas em arquivos técnicos. A ausência de soluções eficientes para armazenar estas interpretações leva a problemas como retrabalho e dificuldades de compartilhamento de informação. Neste trabalho apresentamos uma solução para estes problemas que baseia-se no uso de *anotações semânticas*, uma abordagem que promove um entendimento comum dos conceitos usados. Para tanto, propomos a adoção de workflows científicos para descrição do processo de anotação dos dados e também de um esquema de metadados e ontologias bem conhecidas, aplicando a solução a problemas em agricultura. As contribuições da tese envolvem: (i) identificação de um conjunto de requisitos para busca semântica a dados geoespaciais; (ii) identificação de características desejáveis para ferramentas de anotação; (iii) proposta e implementação parcial de um framework para a anotação semântica de diferentes tipos de dados geoespaciais; e (iv) identificação dos desafios envolvidos no uso de workflows para descrever o processo de anotação. Este framework foi parcialmente validado, com implementação para aplicações em agricultura.

Abstract

Geospatial data are a basis for decision making in a wide range of domains, such as traffic planning, consumer services disasters controlling. However, to be used, these kind of data have to be analyzed and interpreted, which constitutes a hard task, prone to errors, and usually performed by experts. Although all of these factors, the interpretations are not stored. When this happens, they correspond to descriptive text, which is stored in technical files. The absence of solutions to efficiently store them leads to problems such as rework and difficulties in information sharing. In this work we present a solution for these problems based on *semantic annotations*, an approach for a common understanding of concepts being used. We propose the use of scientific workflows to describe the annotation process for each kind of data, and also the adoption of well known metadata schema and ontologies. The contributions of this thesis involves: (i) identification of requirements for semantic search of geospatial data; (ii) identification of desirable features for annotation tools; (iii) proposal, and partial implementation, of a a framework for semantic annotation of different kinds of geospatial data; and (iv) identification of challenges in adopting scientific workflows for describing the annotation process. This framework was partially validated, through an implementation to produce annotations for applications in agriculture.

Contents

List of Tables

List of Figures

Chapter 1

Introdução

O termo *dado geoespacial* refere-se a todos os tipos de dados sobre objetos e fenômenos do mundo que têm características espaciais e que referenciam alguma localidade na superfície da Terra. Estes dados constituem a base para sistemas de decisão aplicados em vários domínios. Em especial na agricultura são úteis para responder questões como "o que plantar, onde, quando, e como". Sendo a agricultura uma atividade de destaque no Brasil, cujos ganhos correspondem a aproximadamente 25% do Produto Interno Bruto do país, melhorias no acesso e no uso deste tipo de dado permitiriam mais eficiência no planejamento e previsão de culturas. Como consequência, teria-se o aumento nos valores obtidos com as produções.

Para serem usados, estes dados precisam ser analisados e interpretados, atividades muitas vezes trabalhosas e dependentes do contexto e do domínio de uso. Estas interpretações, quando armazenadas, geralmente correspondem a alguma informação textual, em linguagem própria e gravadas em arquivos técnicos. A ausência de soluções eficientes para armazenar estas interpretações leva a problemas como retrabalho e dificuldades de compartilhamento de informação.

Uma solução para estes problemas baseia-se no uso de anotações. No contexto desta tese, anotações são definidas como dados que descrevem dados. Entretanto, a simples adoção de anotações não é suficiente, já que cada especialista ou empresa pode usar linguagem ou métodos de descrição próprios, criando barreiras para o entendimento da informação. Para reduzir este problema de entendimento, uma solução é usar na descrição termos de ontologias, como forma de prover semântica. Isso dá origem ao uso das chamadas *anotações semânticas*, uma abordagem que promove um entendimento comum dos conceitos usados, garantindo a interoperabilidade semântica entre produtores e consumidores da informação.

Assim, o principal objetivo desta tese é prover um mecanismo para anotação semântica de dados geoespaciais, como forma alternativa de armazenamento da interpretação asso-

ciada a dados geoespaciais, aplicando a solução a problemas em agricultura. Com isso, espera-se permitir o seu reuso e também apoiar especialistas na aplicação de métodos de planejamento e previsão de culturas. Para alcançar este objetivo, o desenvolvimento da tese atacou diferentes desafios envolvendo anotações semânticas, projeto de workflows científicos e manipulação de dados geoespaciais. Neste sentido, focou em:

- identificação de requisitos para o mecanismo de anotação;

- identificação de características a serem descritas em cada tipo de dado;

- uso combinado de ferramentas para descrever automaticamente as características identificadas;

- automação, o máximo possível, do processo de anotação;

- descrição de cada passo deste processo;

- armazenamento e gerenciamento de anotações semânticas.

Para atacar estas questões, começamos analisando catálogos e portais de informação geoespacial bem conhecidos, como o da FAO (Food and Agriculture Organization)[1] e o GOS (Geospatial One-Stop)[2]. Desta análise resultou uma lista com as principais características e requisitos para a busca semântica a dados geoespaciais. O passo seguinte foi testar ferramentas de anotação para identificação das principais características a serem consideradas no desenvolvimento de nosso framework de anotação. Ao contrário dessas ferramentas, nosso ambiente é genérico e permite a anotação de diferentes tipos de dados. Por esta razão, adotamos workflows científicos para especificar o processo de anotação de cada tipo de dado considerado. Para execução desses workflows, adotamos o ambiente YAWL [94]. O framework foi parcialmente implementado e validado na anotação de dois tipos de dados: uma série temporal de NDVI [3] e uma imagem de sensoriamento remoto, usada para identificar áreas de culturas em uma dada região geográfica.

As contribuições desta tese são:

1. Identificação de características que catálogos de dados geoespaciais devem apresentar para apoiar a busca semântica de dados. Testamos e analisamos alguns catálogos e portais bem conhecidos, o que nos permitiu identificar questões em aberto a serem respondidas levando em conta aplicações web geoespaciais avançadas;

---

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

[2]gos2.geodata.gov

[3]Indíce de Diferença de Vegetação Normalizada - um valor computado a partir de pixels de imagens de satélites, que indica a quantidade de biomassa de uma dada região

2. Identificação de requisitos para anotação de semântica de dados geoespaciais. Testamos ferramentas de anotação semântica em uso como forma de identificar requisitos para nosso mecanismo de anotação. Com base nestes requisitos, propusemos um framework genérico que provê um mecanismo de anotação semi-automática de dados;

3. Implementação parcial do framework de anotação, para validação da proposta. Nesta atividade, levamos em conta características como generalidade, possibilidade de anotação de diferentes tipos de dados e extensibilidade;

4. Identificação de desafios existentes no uso de workflows científicos para orquestração do processo de anotação de dados geoespaciais, e proposta de solução. Dadas suas características, workflows científicos mostraram ser uma boa opção para a automação do processo de anotação de dados geoespaciais. Entretanto, esta opção nos coloca à frente de novos desafios, decorrentes da complexidade das anotações a serem produzidas.

O texto da tese corresponde a uma coletânea de artigos, estando dividido nos seguintes capítulos:

- **Capítulo 2.** Visão geral dos problemas e desafios de pesquisa associados ao trabalho;

- **Capítulo 3.** Apresentação da análise de catálogos e portais de dados geoespaciais, tendo como resultado uma lista de características desejáveis para permitir a busca semântica de dados;

- **Capítulo 4.** Apresentação do teste de ferramentas existentes para a anotação de dados, tendo como resultado uma lista de requisitos para ferramentas de anotação semântica de dados geoespaciais;

- **Capítulo 5.** Descrição do mecanismo e do framework de anotação, resultados desta tese;

- **Capítulo 6.** Apresentação dos desafios identificados na adoção de workflows científicos como ferramenta para orquestrar o processo de anotação de um dado geoespacial;

- **Capítulo 7.** Conclusões do trabalho e possíveis extensões.

Chapter 2

Problem Overview: The Semantic Annotation of Geospatial Data

2.1 Motivation

Agriculture is an important activity in Brazil. According to CEPEA [12] and IBGE [45], in 2007 approximately 25% of Brazil's GNP of U$ 1,477 billion corresponded to agricultural activities. However, this could even increase, if experts could enhance their use of geospatial data, thus supporting more accurate crop 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. Issues involved require defining what to plant, where, when and how. Such questions are important for planning and definition of public policies concerning agricultural practices, also allowing the environmental control of protected areas. Answers to these questions require reliable access to data that is up-to-date; decisions should be based on data that have been properly filtered or summarized, and whose provenance is known.

To be used, geospatial data have to be analyzed and interpreted. These interpretations are context and domain dependent. Data interpretations usually correspond to descriptive text, which is stored in technical files and often not even recorded. Hence, every time a user wants to use such information, the data have to be interpreted again. The absence of solutions to efficiently store them leads to problems such as rework and difficulties in information sharing.

One approach to alleviate these problems is the use of *annotations*. An annotation, in this work, is defined as data that describe other data and, in this sense, can be used

to store interpretations of geospatial data. However, the simple adoption of annotations is not enough, as each expert or researcher, company or country has its own language and description methods, which can create barriers for understanding the meaning of the description. Hence, semantics are needed. This gave origin to the notion of *semantic annotations*, in which ontologies are used to eliminate ambiguities and promote a common understanding of concepts. This, moreover, promotes semantic interoperability among data producers and consumers.

There are several initiatives based on this approach [99, 82, 69]. However, they focus on offering a methodology for manual annotation of data. This is a hard task, especially considering the volume and variety of data to be processed. It is also prone to errors, when performed manually. There are also tools that perform the annotation in an automatic way, some of them considering the semantic issue. However, most of them do not consider geospatial data [89, 98, 80, 13, 41]. When the spatial component is considered, the tool focus on textual data or the annotation is manually performed [43, 7, 48, 50]. This research goes a step further, specifying and partially implementing a framework for semantically annotating different kinds of geospatial data. This framework is being tested for distinct kinds of data, for agricultural planning.

2.2   Research Aspects

This thesis combines many aspects of computer science research. The main challenges comprise issues in semantic annotations, design of scientific workflows and handling data heterogeneity. These issues are briefly analyzed in the following, with focus on agriculture.

Geospatial process interoperability   Interoperability of geospatial process requires handling data appropriately. Our proposal to tackle this relies on annotation of these data. Production of semantic annotations requires accessing different data sources, e.g. data on temperature, climate and crop productivity. How can experts combine the available data to obtain the desired results? In 1999, Geographic Knowledge Discovery (GKD) specialists conducted a meeting to identify research priorities in this area. One of the issues was the incorporation, to geographic applications, of knowledge discovered through queries based on distributed databases [71].

The Semantic Web for geographic information, called Geospatial Semantic Web by Egenhofer [23], is a way to process requests involving different kinds of geospatial information. According to him, it requires the capture and analysis of such information, and their grouping using a criteria that extrapolate their syntactic context. All of this process requires the development of multiple spatial and domain ontologies, their representation

in a way that computers can understand and process, the processing of queries considering these ontologies and the evaluation of results based on the required semantics.

Unfortunately, the Semantic Web is far from becoming a reality [85]. Although a lot of effort has been developed, there are too many things that need to happen. 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 [11].

Heterogeneity of geospatial data   Geospatial data can be of different kinds. How to specify an annotation mechanism that allows the annotation of these different kinds of data, and which is general enough to deal with heterogeneity questions? We analyzed a set of annotation tools, some of which consider the geospatial component [42, 6, 48, 50]. However, they mainly consider textual data, basing the annotation in machine learning methods. Since the identification of annotations is based on string matching, the use of ontologies is essential for the disambiguation and also to correct identification of spatial evidences. However, if the content is an image or a video, the use of ontologies is not enough for automating annotations: the content has to be manually annotated. Except for [50], the analyzed tools did not consider other kinds of content, like satellite images, maps and graphs.

Unlike the tested tools, we combine several components in our framework to facilitate the annotation process and to foster reuse of annotations. Moreover, our framework is extensible and general purpose and considers different kinds of geospatial data.

Geospatial standards and ontologies   Heterogeneity is often handled through two approaches: metadata standards and ontologies. The growing need for geospatial information led to the development of a number of initiatives to obtain spatial metadata according to a variety of formats within agencies, communities of practice, or groups of countries. This resulted in established standards like the ISO 19115 Metadata Standard [46], the FGDC geospatial metadata standard [27], and the Geography Markup Language (GML) [78]. The objective of these standards is to provide a common set of terminology and definitions for the documentation and exchange of geospatial data. These metadata are usually published by geospatial portals, enabling users to discover and retrieve data [75].

Which annotation schema is best? Which are the requirements to be considered in this choice? Can they be directly adopted in agricultural applications? As will be seen, we adopted and extended the FGDC's standard, adding fields that are common in agricultural analysis.

While standards provide a more structured solution to the heterogeneity problem,

ontologies concern semantics. An ontology is a formal and explicit specification of the elements of a domain and the relationships among them [40]. A common vocabulary defined for a domain facilitates information sharing, search mechanisms and reuse [4]. Spaccapietra et al. [90] consider two main kinds of ontology:

- descriptive ontologies, which are those that enrich the description of the semantics of concepts by associating to each concept a structured description of its properties; these ontologies share with conceptual database schema the effort to model some domains or some activities,

- spatio-temporal ontologies, which are those that take into account the spatial and temporal characteristics of information (spatial elements and relationships, and temporal elements and relationships). A difficulty in their development is the lack of an appropriate model, capable of dealing with space and time at the ontological level, and of a suitable reasoning engine.

Ontologies have been used in computing for heterogeneous database integration and knowledge database organization. However, the development and adoption of common ontologies is still a challenge [85]. Hence, in this proposal we focused on well-known ontologies.

Semantic Annotations   "To annotate" means to add notes, to comment. In computing an *annotation* is used to describe a resource (usually a textual resource) and what it does, by means of formal concepts (e.g., using entities in an ontology) [80]. 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. A way to promote interoperability is to use the entities of a domain ontology as those concepts, as in [92]. For example, an annotation may relate the word *orange* that occurs in a text to an ontology that identifies this word as an abstract concept *fruit* (as opposed to concept *color*). This helps removing ambiguity from its meaning. The increase in quality of the retrieved information and in interoperability are some benefits from the adoption of annotations.

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 the example, an ontology is useful to distinguish orange *fruit* from orange *color*, but it is not enough to describe if a document is about the fruit itself or concerns orange culture management.

In geographic applications, annotations should also consider the spatial component, since geographic information associates objects and events to localities, through a rich vocabulary of places and geographic object names, spatial relationships and standards.

Hence, the geospatial annotation process should be based on geospatial evidence – those that conduct to a geographic locality or phenomenon.

As semantic annotations can be hard to read, we decided to maintain the "natural language" description of the our annotation units. This was also part of our research.

Automation of Annotations Part of the efficiency in the annotation process comes from the way it is performed. Reeve and Han [84] 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 is concerned with the design and development of algorithms and techniques that allow computers to "learn". These systems utilize 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 extracts 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 [39]. However, most of the proposed mechanisms consider annotations only of textual content, not taking into account other kinds of content, e.g. images or maps. There is a scarcity of mechanisms to annotate these data, motivating our research.

One important challenge faced was the definition of mechanisms to automate annotations for different kinds of data. Related work is scarce. Usually, proposals are geared towards helping manual annotation procedures. Our solution was based in adopting scientific workflows to orchestrate the annotation process. This, in turn, required work concerning workflow design – i.e., how to construct and execute such workflows, how to specialize them, how to deal with the dynamics of data?

2.3 Overview of the Solution

This section gives an overview of the proposed solution for the problem of semantic annotation of geospatial data. First, we define semantic annotations in the context of this thesis, then we describe the annotation process and finally, we present the adopted architecture.

2.3.1   Semantic Annotations

A semantic annotation combines 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 [64]:

Annotation Units. An *annotation unit* **a** is a triple $< s, m, v >$, where s is the subject being described, m is the label of a metadata field and v is its value or description.
Annotation. An *annotation* **A** is a set of one or more annotation units.
Semantic Annotation Units. A *semantic annotation unit* **sa** is a triple $< s, m, o >$, where s is the subject being described, m is the label of a metadata field and o is a term from a domain ontology.
Semantic Annotation. A *semantic annotation* **SA** is a set of one or more semantic annotation units.

As an example, Figures 2.1 and 2.2 (the latter reproduced from chapter 5) respectively show an annotation unit and the corresponding semantic unit. We point out that our solution is based on storing both kinds of annotations, on for user consumption (in natural language) while the semantic annotations are machine processable.

```
- <metadata>
- <extendinfo>
- <location>
   <county>Monte Santo de Minhas</county>
   <state>Minas Gerais</state>
   <country>Brazil</country>
</location>
- <product>
   <name>Arabica Coffee</name>
  </product>
</extendinfo>
</metadata>
```

Figure 2.1: Annotation generated for a remote sensing image

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 friendlier formats.

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

Figure 2.2: Semantic annotation generated for the same remote sensing image

2.3.2 Framework Overview

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

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

Although expert systems are frequently used in annotation systems [52, 84], not all of our annotation processes can be described by decision systems. Moreover, we are dealing with geographic phenomena. Hence, we have decided to use scientific workflows to describe each annotation process [91, 29]. Each workflow contains information on the annotation schema that will be used during the process, the ontologies to describe these data and the operations to perform.

Figure 2.3, reproduced from [63], gives an overview of the annotation process supported by our framework, which has three main steps: selection of annotation workflow, workflow

execution and ontology linkage. The workflow orchestrates the generation of annotation units. In the last step (linkage) each annotation unit is transformed into a semantic unit, replacing the natural language content by a reference to the associated ontology term. Users may intervene to validate the annotations being generated.

Here, one of the main problems is workflow specification so that annotations can proceed. This requires specifying the workflow, as well as its activities – e.g., implementing them as web services. Moreover, the annotation schema for each geospatial source must be defined. This is what we call the *Configuration step* to prepare the framework for annotating some specific kind of geospatial data source.
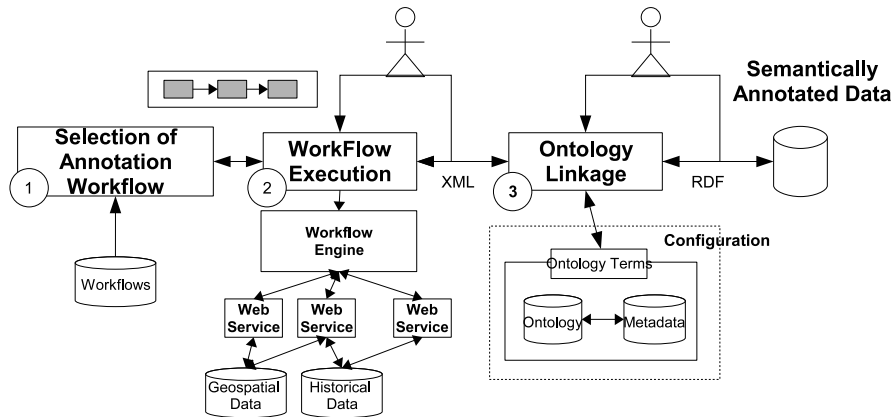
Figure 2.3: The GeoSpatial Data Annotation - Main steps

2.3.3   Architecture of the Framework

The architecture of our framework is divided in two parts: (1) the annotation manager, annotation services and the ontology linker, and (2) the persistence layer, which includes the database manager. This basic architecture was designed taking into account interoperability issues. It is illustrated in Figure 2.4, which is reproduced from [64]. White boxes correspond to external modules invoked by the framework.

The *Annotation Manager* manages the execution of the steps presented on Figure 2.3, working as an event controller. It receives a request for data annotation, identifies the type of the data and makes a request for the retrieval of the corresponding workflow. This workflow will be executed by a Workflow Management System (WfMS) and once the annotation is ready and validated, it is forwarded to the Ontology Linker, for association with ontology terms. *Annotation Services* correspond to the implemented web services that are invoked by an annotation workflow to generate the desired content. The *Database Manager* works as a mediator, providing interoperability for the underlying databases. These databases contain annotation workflows, ontologies, annotated geospatial data and

additional spatial data that is used by the services (e.g., historical information on crop productivity or time series for a given region and phenomena such as rainfall or temperature).

WOODSS [70] provides means to edit workflows that will be executed by a Workflow Management System (WfMS). As it is easy to use, domain experts can describe the annotation process by themselves. As an additional feature, the system provides a set of annotations, enabling the retrieval of the workflows. Aondê [17] is a Web service responsible for handling ontologies. In our framework it provides operations to search, rank, analyze, align and integrate ontologies. This is very useful when the desirable term is not available to be used on the annotation process. In this case it is possible to perform the alignment operation, considering other ontologies besides the one being used.
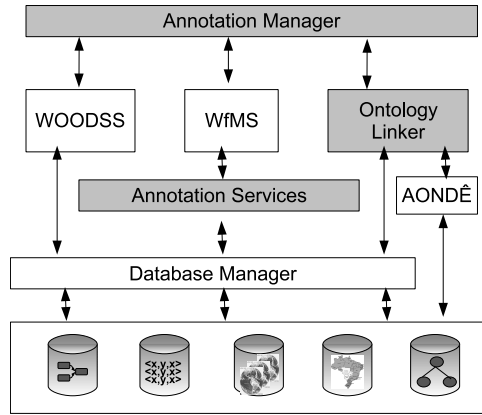
Figure 2.4: The Architecture of the Framework

2.3.4   Implementation Overview

Figure 2.5 gives an overview of the technologies used in implementing our prototype. As described before, we use the WOODSS system to edit the annotation workflows. These workflows are translated to workflows for the YAWL environment [94], which was chosen as the workflow engine to execute the workflow. YAWL was chosen because it is publicly available. More importantly, it directly imports the RDF Schema from the FGDC standard.

Each task in the annotation workflow is responsible for producing one or more annotation units, through the invocation of a web service. The web services were implemented using the Java language and also the framework Axis2 that automatically generates the web service stubs. These services access geospatial and historical information stored in PostgreSQL and PostGIS DBMSs and perform specific operations, such as spatial queries,

to generate/obtain the desired information. The Tomcat server was used as the cointainer for the implemented web services.

The produced annotations are stored as XML files. These annotations will be translated into semantic annotations during the *Ontology Linker* step. The configuration of the framework is performed using the Semantic Annotation Management (SAM) infrastructure [20], which provides means to index and retrieve ontology terms by their semantics. SAM also performs the *Ontology Linkage* step.
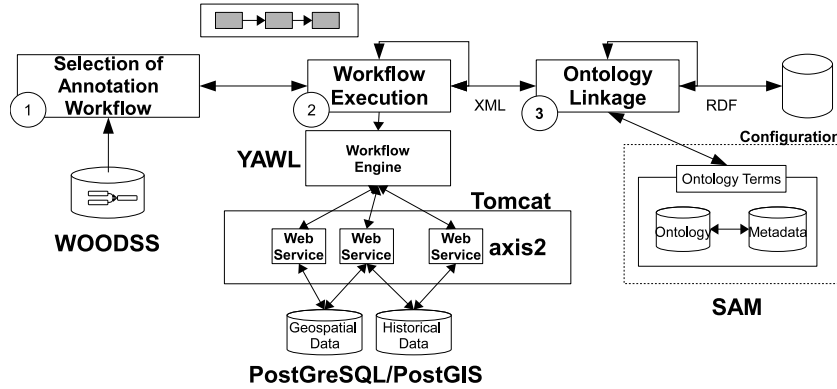
Figure 2.5: Implementation Overview.

2.4   Objectives and Contributions

The main goal of this thesis is to provide an alternative way to store geospatial data interpretation, in agriculture, allowing its reuse and also enhancing the use these data by experts, thus supporting more accurate crop prediction and planning methods. This is done through the use of semantic annotations, in where ontology terms are used to describe the information of each data source. To support such task, this thesis proposes a semi-automatic mechanism for semantic annotation of geospatial data. This mechanism takes into account features of each kind of data being considered and uses computational tools to produce the interpretation.

2.4.1   Work Methodology

To achieve our goal, we focused on advancing the following aspects:

- requirements for the annotation mechanism;

- features to be described for each kind of data source;

- combination of tools in order to automatically describe the identified features;

- automation, as much as possible, of the annotation sources;

- description of each step in such processes;

- storage and management of semantic annotations.

In order to attack these issues, we began by analyzing widely-used portals and catalogs that publish geospatial data, such as FAO[1], for agriculture, and GOS[2]. From this study, we derived the main features and requirements for data access.

The next step was to test a group of well known annotation tools, like the AKTive-Media [13] and CREAM [41], which present methods for semantic annotation of visual resources. This showed what should be supported by an automatic annotation mechanism, how it works and how it should consider spatial information. We realized that some of the tools are not totally automatic. In fact, most of them work as helping tools for the annotation process. During this test, we were also interested on evaluating annotation of non-textual data sources. What we found were tools that annotate textual data. The only exception was [50], that proposes a theoretical method for semantic annotation of maps, based on spatial operations.

The result of this step was the identification of important requirements for our annotation framework, such as the need to establish a metadata schema, annotations format and storage, and annotation methods. This was the basis for the proposal of our mechanism to semantically annotate geospatial data. This led us to the specification of an architecture for a framework to support this mechanism. Since our framework, differently from the tested tools, should consider several kinds of data, we decided to use scientific workflows to specify each annotation process.

To validate this proposal, we started annotating an NDVI [3] graph. Through interviews with experts, we identified the following features for this data source: crop and locality name, expected and historical productivity values, and crop production phases. We also described how to obtain/produce this information. Having this, the next step was to search for data sources and tools to be used. We decided to use the Brazilian Geographic and Statistical Institute - IBGE's [45] productivity and locality data sources, available satellite images from MODIS sensor (*Moderate Resolution Imaging Spectroradiometer*) and a tool to perform queries based on similarity on temporal series [68]. Next, it was necessary to design a workflow describing how to combine the access to the selected data

---

[1]www.fao.org/geonetwork/srv/en/main.home
[2]gos2.geodata.gov
[3]Normalized Difference Vegetation Index – a value computed from pixels of satellites images, which indicates the amount of biomass in a region

sources, in order to produce the desired annotations. Another issue was the definition of the metadata schema to be used. Although we decided for the FGDC proposal [28], we realized that it was not enough to provide information needed in agriculture. Hence, we proposed an extension for the standard, focusing on additional fields. FGDC was chosen because it is an international an open standard.

The next step was the implementation of the proposed architecture. For the execution of annotation workflows we choose YAWL [94], which is a public available workflow engine. This presented some challenges, which led us to some implementation decisions. During this, we also identified the need for a configuration phase, to indicate which ontology terms should be considered during the annotation process. We also had to decide about annotation storage. This was developed within a master dissertation [20].

Finally, the annotation mechanism and the framework had to be validated considering other kinds of data. We chose a remote sensing image, used to identify crop areas in a given geographic region.

2.4.2   Contributions

Considering all this development, the contributions of this thesis are, therefore, the following:

1. Identification of features that geospatial catalogs should have to support semantic search.

   Geospatial information catalogs are complex infrastructures that store and publish geographic information. To be useful, a catalog must efficiently support discovery and retrieval of geospatial information. We identified the main features a geospatial catalog should have to provide semantic search. Considering these features, we selected, on the web some well known catalogs, comparing them by means of these features. This comparison is based on a set of examples which were for all catalogs. Based on this comparison, we identified some open issues that should be addressed considering advanced geospatial applications on the Web. This contribution is detailed on chapter 3. The main contributions are centered on sections 3.3 through 3.6.

2. Identification of requirements for semantic annotation of geospatial data.

   We tested well-known annotation tools, using a basic test case of a web page, as a way to identify requirements for our annotation mechanism. Based on these requirements, we proposed a framework that is generic and provides a semi-automatic annotation mechanism. This contribution is detailed on chapter 4, mainly from section 4.3 to section 4.5.

3. Partial implementation of the annotation framework, to validate our approach, for applications in agriculture. We had to take into account the requirements identified for the annotation process, and also consider other like be general, enable the annotation of different kind of geospatial data and be extensible. This contribution is detailed on chapter 5. The main contributions are centered on sections 5.2.2 through 5.6.

4. Identification of challenges in using scientific workflows to orchestrate the process of data annotation, and how to deal with them. Scientific workflows have emerged as a paradigm for representing and managing complex distributed scientific computations. Such workflows capture the individual data transformations and analysis steps as well as the mechanisms to carry them out in a distributed environment [76]. Workflows proved to be a good choice to help automate the annotation process of geospatial data. However, at the same time, they presented new challenges, given the complexity demanded by these annotations. This contribution is detailed on chapter 6, mainly from section 6.4 to 6.8

A potential gain of our annotations, because of the semantic descriptions, is the increase of the number of relevant documents retrieved in a query operation (the recall factor).

2.5 Thesis Organization

Each chapter of this thesis corresponds to a paper that has been published or submitted to publication.

Chapter 3, *The Geospatial Semantic Web: are GIS Catalogs prepared for this?* corresponds to [66], presented on the 5th International Conference on Web Information Systems and Technologies(Webist 2009). It analyzes some well known geospatial catalogs and their requirements to effectively provide semantic search in the Geospatial Semantic Web. In particular, it discusses some features that GIS catalogs should have, focusing in semantic issues and identifies desirable characteristics.

Chapter 4, *A Framework for Semantic Annotation of Geospatial Data for Agriculture* corresponds to [60], published on the Int. J. Metadata, Semantics and Ontology - Special Issue on "Agricultural Metadata and Semantics", pp. 118–132. The paper presents the main features of semantic annotation tools for geospatial and non-geospatial data. Considering these features, nine annotation tools were tested and analyzed, to identify the main requirements for semantic annotation tools, focusing in agriculture. Having identified these features, we proposed a framework for semantic annotation of geospatial data, presenting a case study for NDVI graphs.

Chapter 5, *Annotating Geospatial Data based on its Semantics* corresponds to [64], presented at 17th ACM SIGSPATIAL GIS Conference. This paper gives an overview of the annotation framework, which is described in detail, as well as the choices made in its design and implementation. To illustrate the use of the framework, the paper presents a case study in agriculture, annotating a remote sensing image.

Chapter 6, *Using Scientific Workflows for Semantic Annotation of Geospatial Data: what are the challenges involved?* corresponds to [63], that was submitted to the Journal of Universal Computer Science (J.UCS). This paper describes the main challenges involved in using scientific workflows to orchestrate the task of semantically annotating geospatial data. It also presents our design and implementation choices to address these challenges, in a prototype developed to validate our ideas.

Chapter 7, *Conclusions and challenges to be met.*

Additional publications associated with work conducted within this research are:

- *"The WebMAPS project: challenges and results (in portuguese)"* [62], presented at the IX Brazilian Symposium on GeoInformatics - Geoinfo 2007, with an overview of the WebMAPS project, its challenges and the obtained results.

- *"Crop monitoring via the web: a successful case in multidisciplinary research"* [65], presented as a poster at the 6th Brazilian Congress of Agroinformatics - SBIAgro 2007. It presents a case study for the WebMAPS project.

- *"An infrastructure for sharing and executing choreographies"* [72], presented at the 4th International Conference of Web Information Systems and Technologies (WE-BIST). The paper proposes the use of semantic annotations to facilitate discovery, sharing and execution of web services choreographies.

- *"Specification of a framework for semantic annotation of geospatial data on the web"* [59], presented at the XXIII Brazilian Symposium on Databases (SBBD 2008) - VII Workshop of Thesis and Dissertations on Databases. In this paper we presented the framework proposal and the expected contributions.

- *"Specification of a framework for semantic annotation of geospatial data on the web"*, that was presented at the 16th ACM SIGSPATIAL GIS 2008 - Ph.D. Showcases. This paper was subsequently published as an article at the ACM SIGSPATIAL Special [61].

Chapter 3

The Geospatial Semantic Web: are GIS Catalogs prepared for this?

3.1 Introduction

The term *geospatial data* refers to all kinds of data on objects and phenomena in the world that are associated with spatial characteristics and that reference some location on the Earth's surface. Examples include information on climate, roads, or soil, but also maps or telecommunication networks. Such data are a basis for decision making in a wide range of domains, ranging from studies on global warming to those on urban planning or consumer services.

For example, geographic applications for consumer services, like those provided by [7] and [47], assign a location to Web pages, based on existing geospatial evidence, such as addresses and phone numbers. This information can be subsequently used, for example, to find consumer services using fuzzy queries and to correlate Web pages spatially. In emergency management, geospatial information can be useful to identify areas prone to disasters [51] or to help in traffic control. In agriculture they are very useful for agroenvironmental planning [62, 60], providing means to enhance agricultural productivity.

The Web plays an important role in this scenario, having become a huge repository of distributed geospatial information. Data are collected and stored by different organizations, which are required to exchange such data. These distributed data may be retrieved and combined in an *ad hoc* way, from any source available in the world, extrapolating their local context. Usually, the search for these data and methods is done by their syntactic content, focusing primarily in keyword matching. This can lead to retrieval of irrelevant data, and to omission of relevant facts. Hence, semantic interoperability is also a key issue in discovery, access and effective search for data in different application contexts. Solutions must take into account the constant modifications in the real world, and the

evolution of our knowledge about the world.

There is a large amount of research on the management of geospatial data, including proposals of models, data structures, exchange standards and querying mechanisms. One area of activity concerns the so-called Geographic Information System (GIS) catalogs. These work as metadata catalogs that can be indexed by various means, such as by geographic location, and provide support for users to search for the data in different GIS data repositories. Catalogs are based on a common set of ideas which do not take semantic interoperability into account. This is a critical function necessary for advanced GIS applications, specially in the context of the Geospatial Semantic Web [23]. In this work we identify important criteria that must be met by catalogs. Based on the results of comparing six widely used catalogs, we point out issues for research and development in the Semantic Web context. This discussion points at directions that must be followed in order to enhance the interoperability of GIS on the Web.

3.2   Related Concepts

3.2.1   Geospatial Semantic Web

The Semantic Web was initially proposed by Berners-Lee [5] 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. In this scenario, the Semantic Web would enable machines to comprehend semantic documents and data, through: (1) adoption of standardized data element names to describe and exchange the data; (2) description of information in terms that allow common understanding; (3) exposing data to be found and retrieved; (4) designing efficient retrieval mechanisms.

A standard establishes the name of data elements (metadata) and/or groups of these elements, providing a common set of terminology and definitions for the description and exchange of data. The adoption of a common vocabulary in this description ensures that data producer and consumer share the same understanding of data. Hence, in the Semantic Web, the description of the meaning of data using ontology terms, through standardized metadata is a way to provide semantics, increasing interoperability. This description process is called *annotation*.

The Semantic Web for geographic information, called Geospatial Semantic Web by Egenhofer [23], is a way to process requests involving different kinds of geospatial information. This requires the development of multiple spatial and domain ontologies, their representation in a way that computers can understand and process, the processing of queries considering these ontologies and the evaluation of results based on the required semantics. All of this leads to the search for a geospatial information retrieval framework

that relies on ontologies, allowing users to retrieve desired data based on their semantics.

In spite of several efforts, the Semantic Web is far from becoming a reality [85]. Although several standards have been developed and adopted, there are too many variables that need to be considered. The variety of user profiles and needs, and of application domains – and thus of ontologies – are just some of these factors. 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 [60].

3.2.2   Geospatial Catalogs

Catalogs are complex structures that enable data to be found and retrieved, through the publishing of descriptions of these data by metadata, known as annotations [75], and operations on these annotations. Catalogs offer search mechanisms that access them to retrieve the desired data.

A GIS catalog is a Web application to publish descriptions of geospatial data, enabling users to search for the desired data [77]. Because of standardized interface specifications, different users can access them from all kinds of sites to search for the content they need.

The Open Geospatial Consortium, OGC [77] is a non-profit international organization that is leading the development of standards for geospatial and location based services. The consortium aims at interoperability among geospatial systems, making complex spatial information and services accessible and useful to all kinds of applications. It describes three basic operations that a geographic catalog should provide: publication, discovery and retrieval of geospatial metadata.

Geospatial data is described by metadata and these descriptions are published in a catalog to support data discovery. Data discovery can be performed either by browsing the content of the catalog or by choosing certain query terms. Once the desired metadata is found, the referenced data can be retrieved.

3.3   Desirable GIS Catalog Features

In a Web environment, GIS users need to explore available databases to discover the desired information. In order to find the data, the first step is to search for specific GIS catalogs and, once connected to the catalog, look for candidate metadata describing the desirable data. As the needed data is found, the users can download and use it in theirs applications.

However, this is not an easy task to perform. Geospatial data are complex, due to their spatial component and its dynamic characteristics. Besides this, users are hampered in their queries because of the many different concepts and terms used to describe data items.

Catalogs seldom publish semantic annotations. One possible approach for this is the use of terms of an ontology to describe data, helping to remove the ambiguity. The increase in quality of the retrieved information and enhanced interoperability are some benefits from the adoption of semantic descriptions, also known as semantic annotations. Although there is extensive research in geospatial semantics, it is focused mainly in the adoption of standardized data element names and of ontology terms to describe the data. It is not common to find semantic catalogs, which are those that publish semantic annotations and support search on them as a way to enhance the retrieval of information. In this section we describe the main features that a catalog should provide in order to make the Geospatial Semantic Web a reality. These features are based on those presented by [55] and [24], always considering the user viewpoint.

Feature 1: OGC Compliance

One of the many standards proposed by OGC is the Catalog Services Interface Standard (CAT), which supports the ability to efficiently publish and search collections of metadata about geospatial data, services and related resources. Hence, focusing in interoperability, a catalog should be OGC compliant, enabling its use by users and also by other catalogs.

Feature 2: Standards for Metadata

Catalogs should support metadata standards. The growing need for geospatial information led to the development of a number of initiatives to obtain spatial metadata according to a variety of formats within agencies, communities of practice, or groups of countries. This resulted in well established and widely used standards like the ISO 19115 Metadata Standard [46], or the FGDC geospatial metadata standard [28]. The objective of these standards is to provide a common set of terms and definitions for the documentation and exchange of geospatial data.

The ISO 19115 standard [46] is a well known standard for geographic information metadata that defines the schema required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data [86]. The Federal Geographic Data Committee [28] develops geospatial data standards for implementing the USA National Spatial Data Infrastructure. The Content Standard for Digital Geospatial Metadata (CSDGM), which is often referred to as the FGDC Metadata Standard, provides the definition of profiles and extensibility through user defined metadata extensions.

Feature 3: Support Advanced Search

Catalogs should provide different means for users to perform their queries, considering different access levels to each catalog and its contents. Users may perform the search considering specific metadata elements, in a way to refine their query. It is a good choice to provide exploration tools, enabling users to explore the retrieved data to determine

suitability to their applications. Users should be able to select the desired sources and categories and kinds of data to be retrieved. Besides this, it is important that each search option be described, enabling its use by foreign people. In this sense, the adoption of standard interfaces can be very useful. Catalogs should also allow users to view metadata records to determine if the retrieved data is suitable for the intended use.

Feature 4: Save Data Online
Catalogs should allow users to view entire metadata records to determine if the corresponding data is suitable for the intended use. Once the user finds the desired content in a catalog, it is important to have means to save its description or even the content itself. Hence, catalogs should support a range of methods for online data delivery (e.g., live data streaming, commonly used data formats, FTP download, and CD-ROM).

Feature 5: Provide Access to Multiple Servers
A catalog should support search considering other metadata servers, increasing the number of repositories to be searched. It has to be done in a consistent way, enabling users to discover new information repositories. The study presented in [24], shows that most users do not perform distributed search due to problems on catalogs. Instead, they go to specific GIS catalogs and browse them to find relevant data for their projects. The portal should also support a search against a single catalog.

Feature 6: Cater to Geospatial Data Diversity
Geospatial data users are always looking for different kinds of data, and also Web services. Hence, catalogs should provide description of all these kinds of data, allowing access to them. For example, maps should be viewable in the browser or through an appropriate software.

Feature 7: Support Semantic Search
Traditional search mechanisms based on keyword matching are restrictive. More expressive search algorithms, which enhance recall and precision, should be available – e.g., via thesauri, gazetteers and multilingual processing. A more flexible option is the use of ontology terms to describe the data. In this sense, the catalog should enable automatic matching of these terms during the discovery process.

3.4 Comparing GIS Catalogs

3.4.1 Overview of Selected Catalogs

We tested some GIS catalogs, as a means to identify issues for research and development in the Semantic Web context in order to enhance the interoperability of GIS on the Web. Although these catalogs are standardized interface specifications, they are implemented considering different requirements, even for the geographic domain. In this test we con-

sidered the guidelines we stated in section 3.3.

Embrapa Information Agency [89] is a Brazilian Web system to organize, deal with, store, publish and access the technological information generated by Brazilian Agricultural Research Corporation - Embrapa and other agricultural research institutes. Knowledge is organized hierarchically, under the form of a tree. Although directed to agricultural domain, knowledge is described using Dublin Core metadata [100], to allow its retrieval by different user profiles. Only a syntactic search for discovery of the stored resources is available, and search results can be saved in a textual file.

INSPIRE (www.inspire-geoportal.eu) is an European initiative that aims to provide geospatial information to be used to formulate, perform and evaluate european policies. Its objective is to create a spatial information infrastructure to deliver integrated spatial information services. The main users of INSPIRE include policy-makers, planners and managers at European, national and local level as well as the citizens and their organisations.

FAO – The UNO Food and Agriculture Organization leads international efforts to defeat hunger [26]. The FAO catalog aims to share geographically referenced thematic information between different organizations. It was implemented using the GeoNetwork opensource (geonetwork-opensource.org), a standard based, free and open source catalog application to manage spatially referenced resources through the web. It offers metadata editing and search functions, as well as an embedded interactive web map viewer. The catalog provides access to interactive maps, satellite imagery and related spatial databases maintained by FAO and its partners.

IDEE – Spatial Data Infrastructure of Spain (www.idee.es) aims to integrate all data, metadata, services and geographic information produced in Spain. Its goal is to make the location, identification, selection and access of these contents an easier operation to their potential users. The IDEE catalog enables users to search for geographic information – maps, ortophotos, etc – available for an area or a theme, in a specific period of time.

GeoSpatial One Stop - GOS (gos2.geodata.gov) is a public GIS catalog that aims to improve the access to geospatial information and data. The catalog is constructed under the U.S. Geospatial One-Stop E-Government initiative for enhancing government efficiency and improving citizen services. Through the catalog it is possible to find data or map services, make a map, browse community information, cooperate on data acquisition. Information is provided by government agencies, individuals, and companies, or obtained by harvesting the data from geospatial clearinghouses.

3.4.2   Comparison of Catalogs

Table 3.1 shows a comparative analysis of the presented catalog systems, taking into account the features presented on section 3.3.

Table 3.1: Evaluated GIS Catalogs.

Catalog	OGC Compliance	Standard Metadata	Save data	Advanced Search	Multiple servers	Data Diversity	Semantic search
Embrapa Information Agency (Brazil)	no	Dublin Core, in portuguese	Yes, in a textual format (descriptive)	yes	no	Digital and non digital	no
INSPIRE (Europe)	yes	ISO19115	no	yes	yes	Digital, web services and applications	no
FAO Catalog (FAO)	yes	ISO19115 FGDC Dublin Core	Yes, the standard metadata in XML format	yes	yes	Digital and non digital	no
IDEE (Spain)	yes	ISO19115 in different languages	no	no	no	Digital and non digital	no
Geodata.gov (USA)	yes	FGDC ISO19115	yes, in textual format (csv)	yes	yes	Digital and non digital	no

Except for the Embrapa Agency and GOS, all the analyzed tools were implemented considering the specifications provided by OGC. Though GOS is not compliant with OGC, it was implemented according to the National Spatial Data Infrastructure directives provided by FGDC, which also focus on cooperative production and sharing of geographic data. All the catalogs provide data that are described using metadata standards, most of them using FGDC or ISO 19115. This indicates that they all aim to promote the exchange of the data they provide. However, to really support data exchange, it is necessary that these descriptions be supplied in an exchangeable format, like XML or csv. The translation of element names from a standard, or saving data descriptions in a textual format, as Embrapa Agency and IDEE do, restricts this exchanging.

The search for data is provided both in simple and in advanced ways in all tested catalogs, except on IDEE, which offers only the advanced one. A simple search enables the user to look for the keyword occurrence within the entire record. However, this can be a hard operation. Embrapa Agency, though offering both kinds of search, has a limited number of options for the advanced search. The same occurs with IDEE. Only three of the catalogs provide access to multiple GIS catalogs, supporting search in different repositories. Though IDEE has this feature, at present it accesses only the National Geographic Institute data. All catalogs provide digital and non digital data, but INSPIRE also provides search for services and applications, which can improve the interoperability among geographic systems. Finally, none of the analyzed systems enables a search based on the semantics of the data.

3.5   Open Research Topics

This section summarizes some open research issues that we have identified as a result of the comparison presented on subsection 3.4.2. This reflects what we expect to be the most important features to be supported by catalogs, towards making the Geospatial Semantic Web a reality.

- *Search on Multiple Servers*: We identify this as a challenge because of the following: (1) some catalogs presented bad performance, thus motivating the need to develop or adopt better algorithms; (2) some results were very difficult to interpret because of the language they use, making the data useless. Hence, content description has to be also in a well-known language; (3) some results were dependent on available services. As many catalog or data providers were offline, it was impossible to get the data.

- *Semantic Search*: This is a central issue to be considered. The available catalogs do not provide this kind of search, in spite of its usefulness when it comes to geospatial data. A good survey of semantic search approaches can be seen in [67].

- *Query Modification*: Although this is part of the previous item, it is also an important issue to be considered by itself. Query modification in catalog search can help disambiguate search expressions and enhance semantics.

- *Adoption of Standards*: This is a large ongoing effort, focusing on interoperability of geospatial data. The FAO Catalog and GOS are good examples for this issue. However, each one is based on a different, but well known, geospatial standard. Hence, if their contents are to be combined, one must develop translators from one to the other. Common standards would avoid this kind of problem.

- *Standard Interfaces*: Once a user wants to search for data in different catalogs, she has to identify the available search options and what each field means. We identify the design of common interfaces as a promising research area. The development of standardized services can also enhance the use of the available catalogs.

3.6   Conclusions

Geospatial data available on the Web are very useful to answer important questions for various domains, such as emergency management, services and agroenvironmental planning. Geographic catalogs are organized as descriptive lists of metadata, which describe existing geospatial data. Through the publishing of these metadata, users are allowed to

search for the desired information to be used in their systems. However, this search is not a trivial task, subject to a wide range of problems. In particular, in the context of the Geospatial Semantic Web, there are two main issues to be addressed: (1) how to perform semantic search, seen as a means to reduce the ambiguity of terms? (2) what should be done in order to have a huge semantic geospatial data network?

This work discussed features that GIS catalogs should present, focusing in the Geospatial Semantic Web. These features are based on interoperability issues, from the user viewpoint. We tested some existing and well known GIS catalogs, comparing them by means of these criteria. Furthermore, we identified research and development issues that are not addressed by the tested catalogs, and that are very important for advanced Geospatial applications. Although many of the existing catalogs are good, they are far from what is needed to support Semantic Networks. Much effort has to be directed to the use of ontologies on search operations. Distributed search also represents a challenge, as this is not a controlled operation. Finally, the adoption of standard interfaces could facilitate the search for data. Initiatives such as OGC are doing a good work in this direction. However there are still gaps to be filled.

Chapter 4

A Framework for Semantic Annotation of Geospatial Data for Agriculture

4.1 Introduction

Agriculture is an important activity all over the world. According to the Brazilian Geographic Institute [44], in 2007 approximately 25% of Brazilian GNP of U\$ 1,477 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 organizations. 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, [67]. 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 [40].

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 centered on a framework to support:(1) creation, validation and management of semantic annotations of geospatial data on the Web, for agricultural planning; and consequently (2) 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 [62].

The rest of this paper is organized as follows. Section 4.2 introduces concepts used. Section 4.3 presents our semantic annotation framework, and its role within WebMAPS. Section 4.4 contrasts our proposal with related work. Section 4.5 describes conclusions and ongoing work.

4.2   Related Concepts

4.2.1   Geospatial Semantic Web

The Semantic Web was initially proposed by [5] 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: (1) to adopt standardized metadata to describe and exchange the data; (2) to describe information in terms that allow common understanding (e.g. ontologies); (3) to expose data so that they can be found and retrieved; and (4) to design efficient retrieval mechanisms.

The Semantic Web for geographic information, called Geospatial Semantic Web by [23], is a way to process requests involving different kinds of geospatial information. This process requires multiple spatial and domain ontologies, to be used in semantic query processing. This leads to the search for a geospatial information retrieval framework that relies on ontologies.

In spite of extensive research, the Semantic Web is far from becoming a reality [85]. 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.

4.2.2   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) [79]. An annotation is represented by a set of metadata that provides 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 evidence – those that conduct to a

geographic locality or phenomenon.

Reeve and Han [84] 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 utilize 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 [39]. 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.

4.2.3   Overview of the WebMAPS Project

WebMAPS [62] 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: (1) 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); (2) the suitability to the Brazilian geographical context; (3) the real time exploration of image content; (4) 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 behavior – see some of these tools at http://www.lis.ic.unicamp.br/projects/webmaps.

Figure 4.1 gives an overview of WebMAPS' 3-layer architecture, part of which is al-

ready 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.

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 [70, 54] 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, visualized as graphs.

Ontology management is performed by Aondê – [17] – a Web service responsible for handling ontologies. It provides a wide range of operations to store, manage, search, rank, analyze 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 [18]. 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 – [75], 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.

4.3 The Annotation Service

4.3.1 Overview

The goal of the annotation service is to semantically annotate different kinds of geospatial data, such as satellite images and maps. According to [1], an annotation model should be as uniform as possible, considering all kinds of content, but also flexible, making it possible
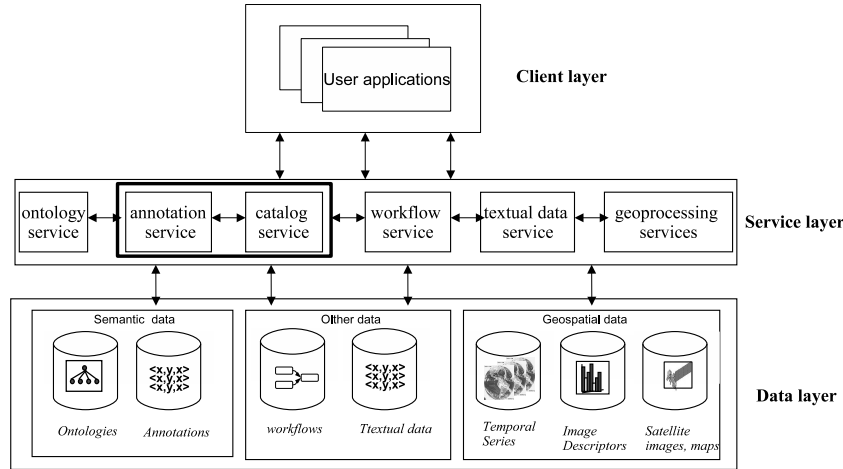
Figure 4.1: WebMAPS 3-layer Architecture

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: (a) an *annotation schema* of metadata labels; and (b) an *annotation content* – ontology terms from official Brazilian sources. The backbone for the annotation schema uses FGDC's [28] 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 [52, 84], 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 [91, 29]. 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 Figure 4.4, 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 4.3), 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 4.2 gives an overview of the annotation service, which comprises 3 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 4.3.2. Step 3 enables discovery, and requires extending the catalog service (see section 4.2.3).

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 [17] 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 is already being used by WebMAPS (e.g., on produce and on regional and ecological characterizations in Brazil). We are extending them with terms from FAO (Food and Agriculture Organization 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.
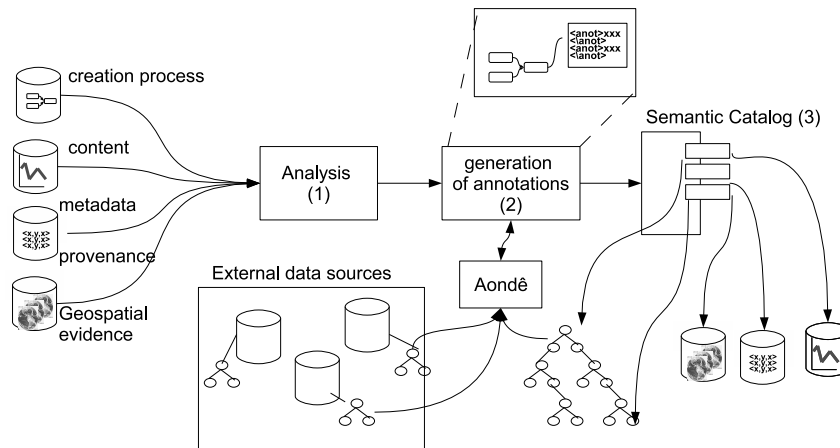
Figure 4.2: WebMAPS' annotation service

4.3.2   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 [57], 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 normalized 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 through a temporal series of images. This can be used for crop monitoring and prediction. For example, in the sugar cane culture, a curve with higher values may indicate a product with better quality. Curves can be compared and analyzed for yield forecast 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 4.3 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 analyze erosion trends, or to define areas suitable for a given crop [29]. 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 [70]. The annotation service takes advantage of this workflow base.

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 4.4 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 organization ontology [29] and on production statistics provided by the Brazilian Agriculture Ministry (www.ibge.org.br/concla).
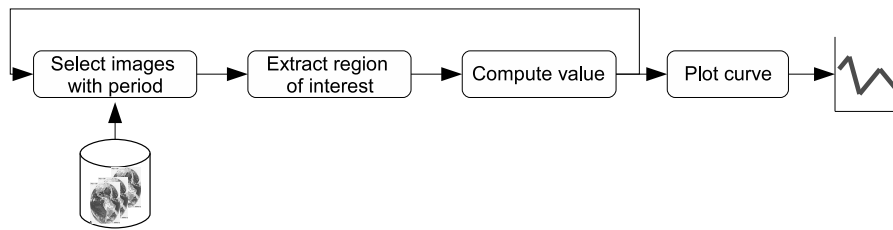
Figure 4.3: Scientific workflow used to generate a set of NDVI graphs

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 [68]) 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 4.4.1), 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.
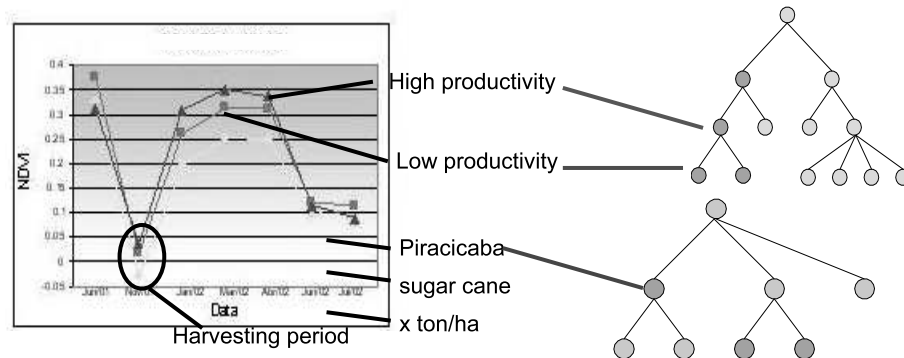
Figure 4.4: NDVI graph with possible semantic annotations

4.3.3 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: (1) enter the information on the farm in the WebMAPS database, including its geometry (see screen copy of data entry on Figure 4.5); (2) generate the NDVI series for the region of the farm – see Figure 4.6, showing the NDVI graph dynamically generated by WebMAPS for that farm, for a given period; (3) use tools that mine time series to retrieve other NDVI series with similar behavior – see Figure 4.7, a screen copy of our series mining tool; (4) analyze the annotations for these series, looking for information on the *yield* ontology term (Figure 4.8).

Figure 4.9 shows the workflow we implemented with help of expert users, to generate semantic annotations for an NDVI graph. At the moment, these workflows are being designed using the YAWL Workflow management system [94], 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.

Figure 4.5: Insertion of a farm in WebMAPS

Figure 4.6: An NDVI graph dynamically generated by WebMAPS for the farm of fig 4.5

Figure 4.7: Retrieval of similar NDVI series

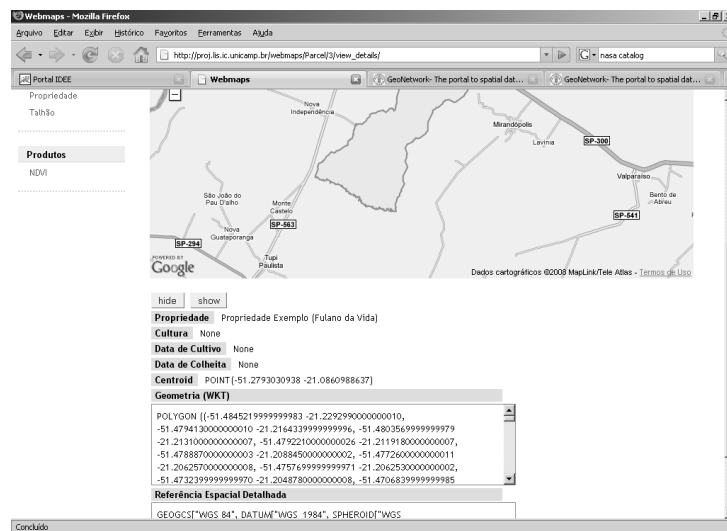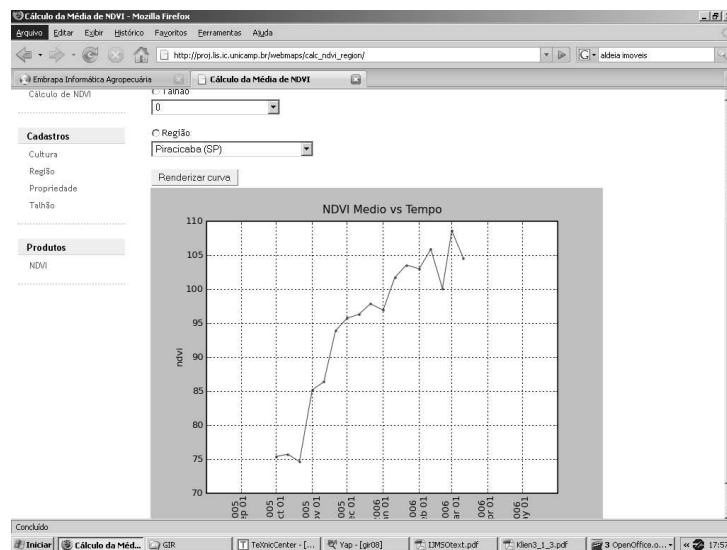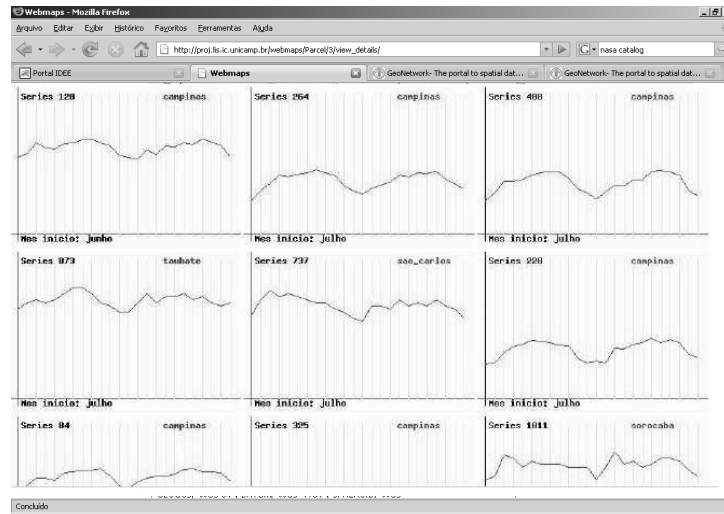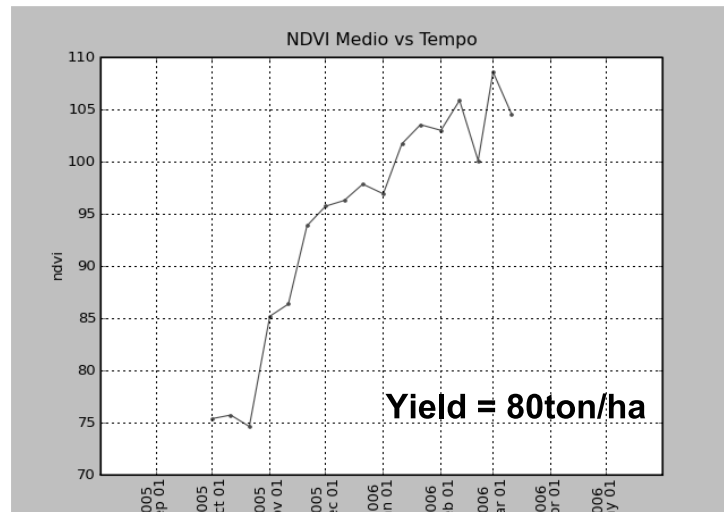Figure 4.8: The desired answer

Get Similar Curves uses our tool for time series mining [68]. 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.
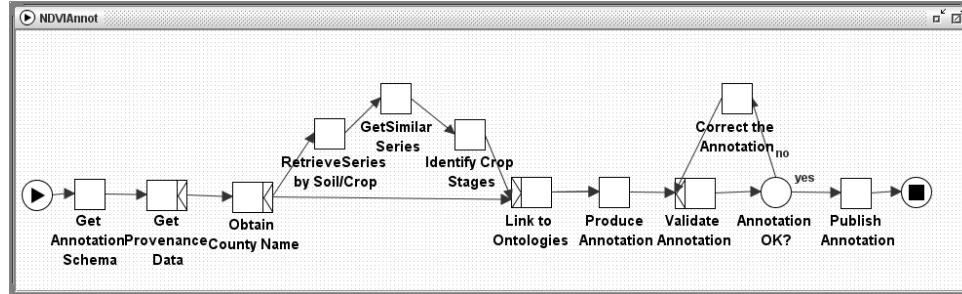
Figure 4.9: The workflow to annotate an NDVI graph

Figure 4.10 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_Organization_Information*, uses IBGE ontology terms. We extended the FGDC standard to include other annotation fields, such as *productivity*, *crop* 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 4.9, the branch (*Get Provenance - Obtain County Name*) is automated, resulting in data shown in Figure 4.10. Similarity of curves is obtained by manually invoking our similarity tool [68]. 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 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.

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.

Figure 4.10: Part of an annotation produced for a geospatial time series

4.4 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.

4.4.1 Non Spatial Annotation Mechanisms

Embrapa Information Agency [89], Amaya [98], KIM [79] 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 [13] and CREAM [41] present methods for semantic annotation of visual resources.

Embrapa Information Agency   [89] is a Web system to organize, deal with, store, publish and access the technological information generated by Embrapa and other agricultural research institutes. Information is organized through a tree branched structure named *knowledge tree*, in which knowledge is organized hierarchically. Each information node can be complemented by information resources (papers, books, image and sound files, etc.) The system uses Dublin Core metadata [100] 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   [98] 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*) [79] 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 analyze texts, in a manual or automatic way, to recognize 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 4.11 shows the Kim Annotation Plug-in. In this example, the

WebMAPS home page was analyzed using the KIM ontology (on the left side). Five entities of class *GeneralTerm* were automatically recognized: *analysis*, *data* (datum), *factors*, *region* and *project*. The plug-in highlighted the annotated entities with the same color of the related ontology term.
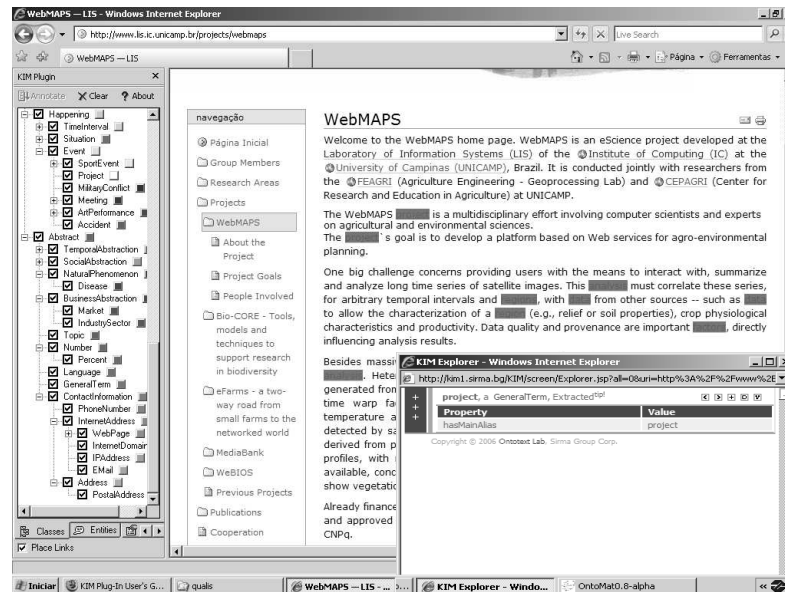
Figure 4.11: Annotation the WebMAPS main page using the Kim Annotation Plug-in

AKTive Media [13] 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 4.12 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 *NonProfitOrganization*; *Institute of Computing*, *University of Campinas* and *FEAGRI* as *EducationalOrganization*, 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 [41] – is a framework that allows the creation of metadata that instantiate interrelated definitions of classes in a
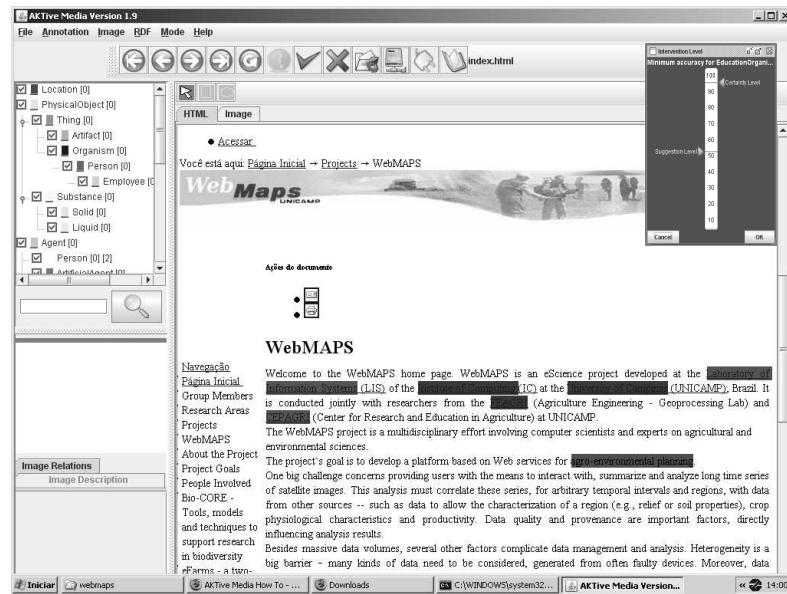
Figure 4.12: Annotation the WebMAPS main page using the AKTive Media

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, organization, location. This schema can be modified to cover the desired annotations. Automatic annotations are created using the processing 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 4.13 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, CEPA-GRI and CNPq were annotated as instances of class *Organization*, the last one as a *research-funding organization*. 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.

4.4.2 Spatial Annotation Mechanisms

The traditional systems described in section 4.4.1 are not able to mine for information based on spatial components, mainly because their search mechanisms do not have features
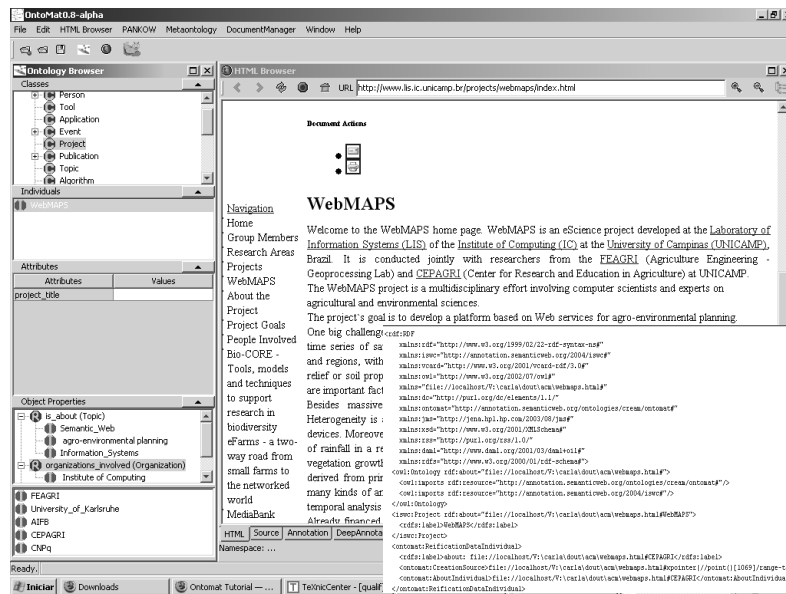
Figure 4.13: Annotation the WebMAPS main page using the OntoMat tool

to deal with spatial relationships. We now present some approaches that consider the spatial component.

E-Culture [43, 42] 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 [97], an extension of Dublin Core [100] for images. This schema has at last 4 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 [6] consists of a geographic information retrieval approach supported by the OnLocus ontology for recognizing, 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 conceptualize 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 4 million Web pages. Like our proposal, it is based on ontological spatial knowledge. Unlike ours, it is centered on annotating Web pages and is applied to urban applications.

SPIRIT – Spatially-Aware Information Retrieval on the Internet [48] – 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 recognize 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 analyzed (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 [50, 52] 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) [58]. 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 analyzed according to different aspects, such as topography, hydrology and geology. Figure 4.14 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.
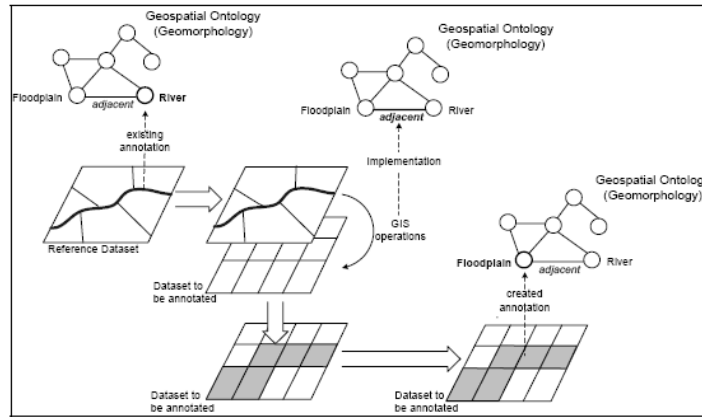
Figure 4.14: Procedure for (semi-)automated annotation of geodata from [52]

4.4.3   Analysis of the Presented Tools

Table 1 shows a comparative analysis of the presented tools, taking into account the requirements pointed by [84] and [92] 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*

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	geoespatial evidences (addresses)	Textual Web pages	yes
SPIRIT		yes		yes	geoespatial evidences	Textual Web pages	yes
Geodata Annotation	XML, using ISO 19115 metadata	yes		yes	Spatial methods, string matching	Geographic data	yes

Table 4.1: Summarization of the analyzed annotation tools

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 analyzed focus on annotation of textual 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.

4.5 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 4.4, present annotation mechanisms are centered 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 4 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 [17] was developed to meet some of these challenges, but much remains to be done. For more information on open problems, the reader is referred to [25].

Chapter 5

Annotating Geospatial Data based on its Semantics

5.1 Introduction

The term *geospatial data* refers to all kinds of data on objects and phenomena in the world that are associated with spatial characteristics and that reference some location on the Earth's surface. Examples include information on climate, roads, or soil, but also maps or telecommunication networks. According to [87], this kind of data corresponds to about 80% of the available data. Therefore, geospatial data contribute significantly to human knowledge. They constitute a basis for decision making in a wide range of domains, from studies on global warming to those on urban planning or consumer services.

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

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

51

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

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

The rest of this paper is organized as follows. Section 5.2 presents our semantic annotation framework, giving details of its architecture. Section 5.3 discusses implementation aspects. Section 5.4 presents a case study in agriculture. Section 5.5 contrasts our proposal with related work. Section 5.6 describes conclusions and ongoing work.

5.2   The Annotation Framework

5.2.1   Semantic Annotations

This work combines characteristics of metadata and annotations into semantic annotations: metadata fields are filled with ontology terms, which are used to describe these fields. Based on this, and following [81], we define semantic annotations as follows.

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

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

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

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

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

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

5.2.2   Framework Overview

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

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

Although expert systems are frequently used in annotation systems [52, 84], not all of our annotation processes can be described by decision systems. Moreover, we are dealing with geographic phenomena. Hence, we have decided to use scientific workflows to describe each annotation process [91, 29]. Each workflow contains information on the annotation schema that will be used during the process, the ontologies to describe these data, the operations to perform and how to store the generated annotations.

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

Figure 5.1 gives an overview of the annotation process supported by our framework, which has three main steps: selection of annotation workflow, workflow execution and ontology linkage. The workflow orchestrates the generation of annotation units. In the last step (linkage) each annotation unit is transformed into a semantic unit, replacing the natural language content by a reference to the associated ontology term. Users may intervene to validate the annotations being generated.

In more detail, the framework receives as input a geospatial data file to be annotated

---

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

[2]gos2.geodata.gov

and also some provenance data. The type of data is identified and a specific workflow is selected to be executed. This workflow indicates the annotation schema, and the operations to be performed to produce annotation content. During this process, the annotation units are presented for user validation, usually a domain expert, who may choose another workflow or define a new one. In the third step, appropriate ontology terms are chosen to assemble the semantic annotations (linking annotation units to ontology terms). The semantic annotations are stored as RDF triples in a XML database, where they can be used for information retrieval, e.g. using XQuery statements.
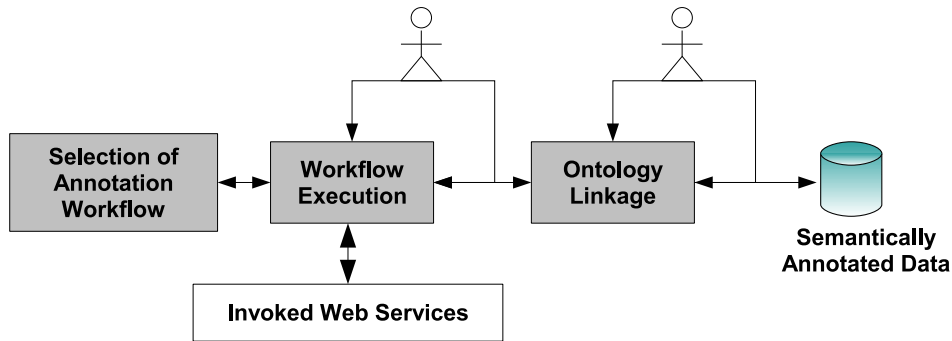
Figure 5.1: The GeoSpatial Data Annotation - Main steps

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

Configuration also involves selection of ontologies, and their terms, to be used for content description. They have to be well-known, consensual, ontologies and adherent to the domain. Good examples are POESIA [29] (for agricultural zoning) and SWEET [73] (for various domains such as geography, physics, and chemistry).

5.2.3   Architecture of the Framework

The architecture of our framework is divided in two parts:(1) the annotation manager, annotation services and the ontology linker, and (2) persistence layer, which includes the database manager. Figure 5.2 presents this basic architecture, which was designed taking

into account interoperability issues. White boxes correspond to external modules invoked by the framework.

The *Annotation Manager* is responsible for managing the execution of the steps presented on Figure 5.1, working as an event controller. It receives a request for data annotation, identifies the type of the data and makes a request for the retrieval of the corresponding workflow. This workflow will be executed by a Workflow Management System (WfMS) and once the annotation is ready and validated, it is forwarded to the Ontology Linker, for association with ontology terms. *Annotation Services* are responsible for implementing the services that are invoked by an annotation workflow to generate the desired content. The *Database Manager* works as a mediator, providing interoperability for the underlying databases. These databases contain annotation workflows, ontologies, annotated geospatial data and additional spatial data that is used by the services (e.g., historical information on crop productivity or time series for given region and phenomenon such as rainfall or temperature).
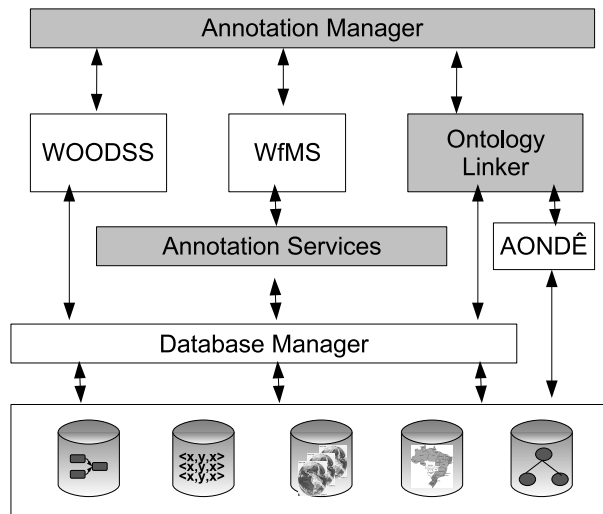
Figure 5.2: The Architecture of the Framework

Workflow Selection - WOODSS

An annotation workflow specifies the process of producing annotations tailored to each kind of geospatial content, for a given use context. These workflows are specified using WOODSS, a workflow tool [70] that provides means to edit and manage scientific workflows. All workflows are stored in a specific repository. Figure 5.3 illustrates a workflow specified using WOODSS, which is used for annotating NDVI time series with county, crop, production, etc.

One can see, for instance, that the generation of annotations begins by retrieving the schema for the particular data source. Once the county name is obtained (e.g., from coordinates) the next step retrieves a set of NDVI series from the same region, which are already annotated and similar to the input series. Each retrieved series is associated with a given crop. Crop names are presented to the user, as annotation suggestions. If there is more than one crop name, the user can choose the most appropriate one. Productivity is next estimated from the similar series.
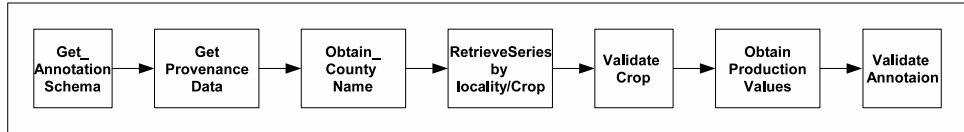
Figure 5.3: A workflow in WOODSS for semantic annotation of a NDVI time series

Workflow Execution – Annotation Units

The WfMS is responsible for executing the selected workflow, through the use of a WfMS, such as the YAWL environment [94].

During this execution, the annotation schema to be filled is retrieved. The schema indicates which metadata elements should be used for each kind of geospatial data file. Workflow execution will produce information to fill each one of these fields. This schema is based on FGDC's [28] geospatial metadata standard, a general purpose and open standard. However, a full description using all fields from this standard may be too long. Hence, for a core geospatial annotation, we identified the most relevant parts of the schema, taking into account the metadata usually provided by some well known Geographic Portals, such as INSPIRE[3], IDEE[4], FAO[5] and GOS[6]. We also realized that the FGDC standard needs to be extended for some special domains, like agriculture. Thus, for the kinds of data we are working with, in our testbed, we have provided additional schema fields, to account for domain requirements.

Our annotation schema is divided into two parts: Identification and Extended Information. Figure 5.4 illustrates this schema. Section idinfo corresponds to Identification information from the FGDC standard, including citation (*citation*), description (*descript*), period that the data comprehends (*timePerd*), status of data (*status*), information of locality (*SpDom*) and keywords (*keywords*). The second part (*extendinfo*) is used to describe the information resulting from data interpretation and can vary according to the kind of

---

[3]www.inspire-geoportal.eu

[4]www.idee.es

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

[6]gos2.geodata.gov

data being annotated, domain being considered or usage context. In the example, for agricultural issues, it includes information on location (*location*) and on crop production (*product*).

Figure 5.4: The adopted Annotation Schema

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

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

Ontology Linker

This module is responsible for linking each annotation unit to a term in an ontology. In other words, an annotation unit *<resource identification> <metadata schema label>*

<natural language content> will be transformed into a semantic annotation unit by linking the content to an ontology term. The module thus deals with our second challenge: automatic identification of the ontology terms to be used. Existing tools for semantic annotation, such as [80], [13] and [41], yield this responsibility to the user performing the annotation task.

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

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

Despite the structure and semantics that metadata can provide, the content of the fields may not be able to avoid this and other kinds of problems [52]. The use of ontology terms guarantees unique meaning, associating annotation units to concepts that semantically represent their content. Ontologies also provide a hierarchical structure that helps to understand their concepts. Figure 5.5 shows the solution for this example, using terms of POESIA Agricultural Zoning ontology [29]. It indicates that *University of Campinas* is a public university and furthermore it is an organization categorized as a public institution. Here, an annotation unit might be *<resource_id><origin><UNICAMP>* while its semantic interpretation is *<resource_id><fgdc:origin class= "http:// www.lis.ic.unicamp.br/ poesia#PublicUniversity"><'University of Campinas'>*.

The Aondê ontology Web service [17] plays an important role in the linkage process, looking for and querying appropriate ontology terms, or aligning ontologies available within the framework to those used by external sources. For instance, suppose the annotation field *origin* is filled with "State University of Campinas". However, this is not a term on the used ontologies. Hence, using AONDÊ alignment services, it is possible to look for synonyms or the correct term – in fact just *UNICAMP*. Alignment involves identifying term and structure similarities between ontologies, and in our case is ensured by Aondê.

Given the country's context and our domain context, our primary ontological sources come from the Brazilian Agriculture Ministry – e.g., on soil, live animals, vegetation, agro-ecological relief and other agriculture-related issues. Information on other geospatial
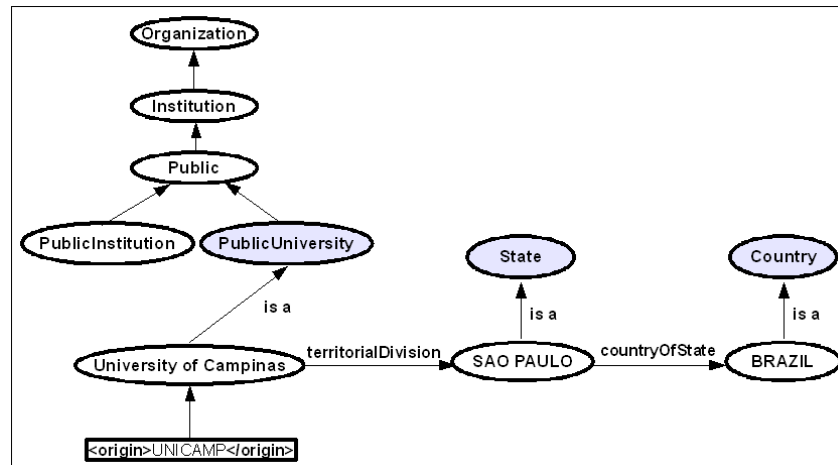
Figure 5.5: Associating an ontology term to an annotation field

features, including an ontology with over 16,000 terms concerning Brazil's spatial unit names and relationships, was taken from IBGE[7] – Brazil's National Geographic Institute.

5.3 Implementation Aspects

We are implementing a framework that supports the whole annotation process to validate our proposal. Its design and construction followed the main principles of adoption of standards and ontologies to provide interoperability. The framework is being implemented in JAVA, since it provides several APIs that can facilitate our work. It also is centered on XML files, which facilitates data exchanging. Since WOODSS does not have a native execution engine, we adopted YAWL for this task [94]. Each activity in the workflow is linked to a Java annotation service.

5.3.1 Configuring the Framework

Editing Workflows

We use WOODSS [70] to edit the workflows, since this is an environment easy to use and it supports annotations of workflows and their storage in a database. In WOODSS, workflows (which are themselves annotated to allow their reuse) are stored in the PostGreSQL DBMS. This allows the automatic selection of the appropriate workflow to execute, which can be retrieved according to the annotations attached to it (e.g., indicating that it is a workflow that orchestrates the annotation of a satellite image, for crop identification in

---

[7]www.ibge.gov.br

agriculture). WOODSS does not have a native execution engine, and its workflows have to be exported for execution.

Choosing Ontology Classes

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

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

Ontology selection is performed by an expert, using a Web interface. Figure 5.6 illustrates this process, which has three main steps: selection of ontologies, selection of ontology terms and their association to annotation fields and storage of this information. In the first step, the user types the URL of some ontology of interest to be used for the annotations. The module loads this ontology and extracts all the URI's of the ontology terms, using the Jena Ontology API[8]. Having all these URI's, the user is asked to indicate which term can be used to fill each annotation field. Note that one term may be associated to one or more annotation fields. At the end, the module stores the URI of the chosen terms, and the label of associated annotation fields in a database.

At this part of the framework, the expert has to indicate the ontology classes to be used in each annotation field, for a semantic description. As most of these classes have associated instances, the name of these instances will work as a controlled vocabulary of natural language terms to be used during the generation of the annotation units. However, in case of absence of appropriate instances, classes can be used to characterize the content. Another option is the usage of AONDÊ, for ontology alignment. Considering the example of Figure 5.5, "*University of Campinas*" is a natural language description for *origin*, whose semantic description is *"http:// www.lis.ic.unicamp.br/poesia#PublicUniversity"*.

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

5.3.2   Creating Annotation Units

During the annotation process, the annotation units are stored in XML files. We used the Java Architecture for XML Binding (JAXB), a java API that easily maps Java classes to

---

[8]http://jena.sourceforge.net/ontology/index.html. Accessed in June 15th, 2009.
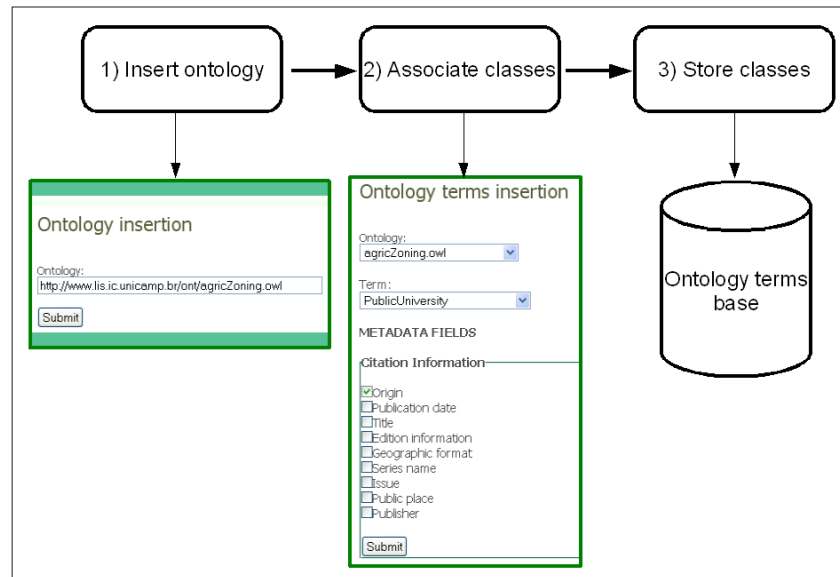
Figure 5.6: Process of association of ontology terms to annotation fields

XML representations. Through JAXB, we just had to define a XML schema (XSD file) for the adopted annotation schema and the API generates java classes to read and write an XML file in accordance with the given XSD file. Since FGDC provides the corresponding XSD files for their geospatial metadata standard, we just had to adapt theses files to our needs.

Figure 5.7 presents part of the XML Schema for our annotation schema presented in section 5.2.3. For example, the annotation schema in XML to be generated is composed of a field *metadata*, which has two kinds of metadata: *idinfo* and *extedinfo*. Field *idinfo* is of *idinfoType*, which indicates that it composed by other six metadata fields: *citation, descript, timeperd, status, spdom* and *keywords*.

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

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

When one of these services is executed, it produces some kind of description in natural

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

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

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

Figure 5.7: Partial XML Schema – FGDC

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

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

5.3.3   Creating Semantic Annotation Units

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

Figure 5.8 illustrates an annotation unit of a remote sensing image, considering the

---

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

schema presented on Figure 5.7. The *rdf:Description* element indicates a description of some resource. The *rdf:about* attribute identifies the resource by its URI. Next, come the annotations units fields, using the following rule: if an element is composed of one or more elements, it must have a *rdf:parse Type="Resource"* attribute indicating that it contains other elements.

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
            xmlns:fgdc="http://www.fgdc.gov/metadata/fgdc-std-001-
1998.xsd#"
            xmlns="http://www.lis.ic.unicamp.br">

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

                </fgdc:serinfo>
                <fgdc:pubinfo rdf:parseType="Resource">
                    <fgdc:pubplace>Campinas - SP</fgdc:pubplace>
                    <fgdc:publish>LIS, IC-UNICAMP</fgdc:publish>
                </fgdc:pubinfo>
            </fgdc:citeinfo>
    </rdf:Description>
</rdf:RDF>
```

Figure 5.8: RDF annotation of a remote sensing image

In order to link annotation content to ontologies, we use the ontology instances of the annotation units to identify the ontology terms that will be used on the mapping to the semantic annotation units. As these instances are related to ontology classes, it is quite simple to provide the semantic description for the annotation units. As we want to maintain the "natural language" description of the annotation units, we use the predicate *rdfs:comment* from RDF Schema[10] (RDFS), which represents a human-readable description. Hence, a semantic annotation unit is a triple, using the property *rdf:type* to specify that the content of the semantic annotation unit is an individual of an ontology class. In the example of Figure 5.5, the field *origin* contains a human readable description (content of *rdfs:comment*), which says that the resource was originated by "University of Campinas", and a reference to the class *PublicUniversity* (*rdf:resource=http://www.lis.ic.unicamp.br/poesia# PublicUniversity"*), specifying that

---

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

the originator of the resource is an instance of this class (via *rdf:type*). Thus, we want to say that "the resource was originated by UNICAMP, which is a public university".

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

Figure 5.9: Referencing an ontology term to *fgcd:origin* element.

5.3.4   Storing Semantic Annotations in RDF

Another issue we faced was to choose how to store annotations. RDF can be represented by various languages, the RDF/XML language is the most common. One of the essential characteristics of a good quality geographic metadata standard is that it should be XML compatible. Both FGDC Metadata and ISO 19115 have this feature, as well as metadata standards from other domains such as e-GMS [2]. These facts made us choose a XML database to store RDF/XML semantic annotations.

An XML database is a data persistence software that allows storage of data in XML format, mapping these data from XML to some storage format, which can be a relational database or even other XML documents [103]. Queries over a XML database are generally executed using XPath or XQuery statements. It is possible to retrieve RDF/XML data using XQuery.

XPath and XQuery allow retrieval of full XML-based documents or subtrees thereof, using their DOM trees[11]. If we know the schema of an annotation that we want to retrieve, we can retrieve the full annotation or a part of interest. For example, if someone wanted to know who originated the remote sensing image of the example from Figure 5.8, he could retrieve this information using the XPath statement (*/rdf:RDF/rdf:Description/fgdc:citeinfo/ fgdc:origin*).

5.4   Case Study - Agricultural Planning in Brazil

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

---

[11]The XML DOM (Document Object Model) defines a standard way for accessing and manipulating documents compatible to XML, presenting them as a tree structure where elements, attributes, and text are nodes.

which information – prediction, estimates and zoning – are the basis for Brazilian government policies to finance agricultural activities. Besides this, at reaping time, the follow up of this information ensures the payment of insurance, when needed, and allows new financings.

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

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

Figure 5.10 presents a remote sensing image of Monte Alto county, located in one of the Brazilian regions with the highest coffee productivity index. Annotations that are result of the our process are, for instance, the county name, and production and climate factors.

Figure 5.10: Remote sensing image for arabica coffee in Monte Santo county
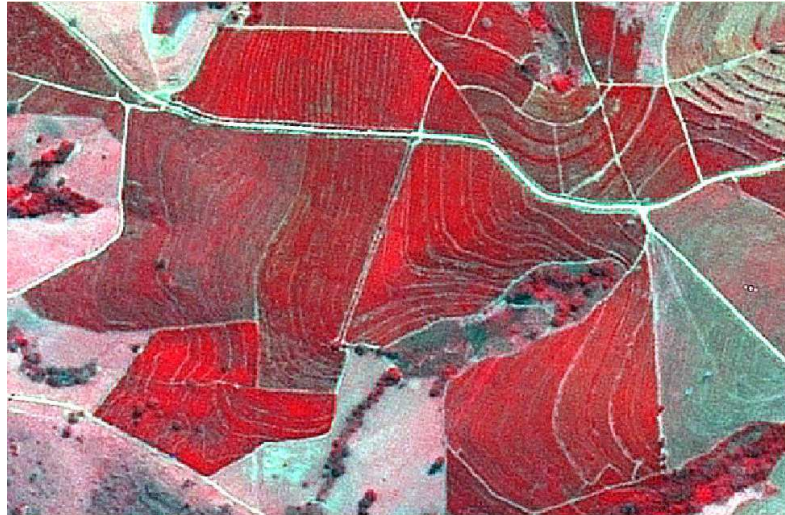
Figure 5.11 presents the workflow for annotation of a remote sensing image. After the selection of the schema, an image classification tool is invoked. This tool [22] uses image processing techniques, and based on spatial and texture information, provides vegetation cover identification (here, crop name). If the user validates the crop, historical productiv-

ity values are obtained for this crop in the same region. These values are obtained from IBGE database, which maintains information of productivity for different crops, grouped by geographic region – macro and micro region, state and county – and by year.
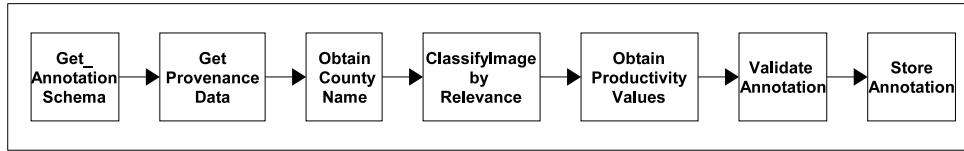
Figure 5.11: The core workflow for annotation of Remote sensing images

Figure 6.4 presents part of these annotations. This corresponds to the Extended Information of the schema. For example, the image is related to arabica coffee crop (Crop Identification), the pair $<crop>$, $<rdf:li$ $rdf:resource="http:// www.lis.ic.unicamp.br/ont/ agricZoning.owl\#Arabica"/>$.

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

Figure 5.12: Semantic annotation generated for a remote sensing image

Figure 5.13 shows a table that explains the terms used in the semantic annotation of this image. The first column shows the annotation fields used. Each field shown in the table is composed by other specific fields, which were abstracted in the table. The second column has a brief description of each element. The third column shows its short name, defined in their respective XML Schema. The fourth column indicates from which metadata standard the field belongs. The fifth column specifies whether the presence of the element is mandatory or not. The last column indicates the ontologies used to describe each annotation field.

The experts just have to validate the created semantic annotations. Using them, a Brazilian government expert may confirm the extension of a crop, producing correct

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

Figure 5.13: Composition of a semantic annotation of a Remote Sensing Image

productivity values. Another important use is the identification of diseases, impacting insurance. As an additional gain, our annotations, because of the semantic descriptions, can enhance the number of relevant documents retrieved in a query operation (the recall factor).

5.5 Related Work

Our paper concerns semantic annotations of geospatial data, including tools and to generate and manage these annotations. This section presents related work concerning these issues, which comprises semantic annotation tools, the use of semantic annotations to record interpretations and representation and sharing of meta-information.

5.5.1 Existing Annotation Tools

Annotation of digital content, due to the volume of available information, is not an easy task, always subject to errors. This led to the development of tools, which aim to facilitate the annotation process. We have tested some of them, taking into account the requirements pointed by [84] and [92]. Embrapa Information Agency [89], Amaya [98], KIM [80] 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 [13] and CREAM [41] present methods for semantic annotation of visual resources.

In geographic applications, annotations should also consider the spatial component, since geographic information associates objects and events to localities, through a rich

vocabulary of places and geographic object names, spatial relationships and standards. Hence, the geospatial annotation process should be based on geospatial evidences – those that conduct to a geographic locality or phenomenon, e.g. see [7, 47]. E-Culture [43], OnLocus [7], SPIRIT [48] and Semantic Annotation of Geodata [50] are approaches that consider the spatial component for the annotation of digital contents.

Except for the SPIRIT project, all the analyzed tools use a *standard format*, like XML, OWL or RDF to save their annotations. Among them, [89],[41] and [50] also adopt standardized metadata (Dublin Core, VRA and ISO 19115), which increases the probability of the annotated content to be found. On the other hand, annotations which are saved on RDF or OWL enable the annotated content to be found during a semantic search, through the use of ontologies. During this comparative study of annotations tools, reported in [60], we also observed that when the data to be annotated are mainly textual, without taking the spatial component into account, the annotation method is based on machine learning. In this case, since the identification of annotations is based on string matching, the use of an ontology is essential for the disambiguation. The same occurs when the spatial component is taken into account: if the process is automated, the use of ontologies is a key factor for the correct identification of spatial evidences. However, if the content is an image or a video, it has to be manually annotated. The analyzed tools do not consider other kinds of content, like maps and graphs, for annotation.

Tools have also to be compared considering storage features, since the efficiency of the annotation process is measured by the results of a content search. Annotations stored in an annotation server, like a catalog – as in [80] and [41] – facilitate content discovery, different from those stored in local files [13]. On the other hand, annotations stored in a relational database, as in [89], will not enable content discovery, unless they are also published in another media, like web pages.

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

5.5.2   Using Annotations to Record Interpreted Information

There are several initiatives that use annotations to store data interpretation. Wang et al. [99] present a framework to annotate medical images, as a way to promote information sharing, in a collaborative annotation process. The annotations can be textual or multimedia. The former ones are based on a limited group of metadata and are used to describe regions of interest on the image. The latter are used to enrich existing information. Unlike us, they do not consider semantic issues.

Rainaud et al. and Mastella et al. [82, 69] deal with recording of interpretations of

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

5.5.3   Management of Metadata

Use of ontologies to deal with interoperability problems in the geospatial domain is discussed in [95, 30, 31, 50], but not focusing on the use of geographic metadata, while [74, 14] discuss interoperability among geographic metadata standards.

Another trend is the representation of geographic meta- information, in which RDF is being widely used. In [16], RDF is used to define a catalog of geographic resources from various Web sites. Córcoles and Gonzáles [15] propose an approach for providing queries over spatial XML resources with different schemas using a unique interface, where the resources are integrated using RDF. Although these works concern aspects like integration and interoperability, they do not explore the use of ontologies.

Our framework uses XML databases to store metadata in RDF/XML, due to the conventional use of XML to share and store meta-information. There are some works that also use XML databases to store other kinds of metadata. In [3], a XML database is used to store metadata in a prototype of a digital library system, which provides queries over metadata from art pieces. The use of XML databases for the management of metadata in the MPEG-7[12] format is discussed in [101], with a survey concerning XML database solutions for this issue. A schema-independent XML database used to store metadata about scientific resources is presented in [49].

Another solution for storing and querying RDF is to use some framework for these purposes, like Sesame [8] and Jena [102]. These frameworks play the role of a layer that manages persistent storage of RDF in files or relational databases and provide queries over RDF in SPARQL or in other specific languages. Moreover, such frameworks provide reading and writing of RDF in different notation languages. We intend to use a framework like these in the future and so compare this approach to the storage in XML databases.

---

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

5.6   Conclusions and Future Work

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

This paper presented and discussed an approach for alleviating this problem based on semi-automatic annotation of geospatial data. This approach was outlined in [60] and this paper discusses architectural and implementation issues. Our proposal, which is being validated in the domain of agricultural planning and monitoring, presents the following characteristics: it is compliant to Semantic Web standards; the descriptions are free of ambiguities in their understanding; and it promotes interoperability.

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

---

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

Chapter 6

Using Scientific Workflows for Semantic Annotation of Geospatial Data:what are the challenges involved?

6.1 Introduction

Geospatial data, in this paper, are all kinds of data sources that have explicit or implicit connection with some location on the Earth surface – i.e., contain information on geographic coordinates, or some means to derive such coordinates (e.g., by providing a place name). Examples include information on climate, roads, vegetation, telecommunication networks, or demography. According to [87], this kind of data corresponds to about 80% of the available data on the Web. Because of the spatial dimension, such data constitute a basis for decision making in a wide range of domains, from studies on global warming to those on urban planning or consumer services.

However, to be used, these data have to be analyzed and interpreted. These interpretations are context and domain dependent and performed every time the data are needed. For instance, a given satellite image will go through distinct analysis processes depending on whether it is to contribute to studies on water pollution, urban occupation, or agricultural practices. Such interpretations produce new information, which may be stored in technical files, but is often never recorded. Hence, every time a user wants to take advantage of such information, the data have to be interpreted again.

To enhance information sharing, some researchers adopt *semantic annotations* to store these interpretations (roughly, a combination of metadata labels and ontology items). However, to annotate data is a hard task. Annotation of geospatial data, in particular,

requires collaboration of multiple experts, and is time consuming, which can be a bottleneck in the process. In order to alleviate this task, we designed a framework that supports semi-automatic annotation of geospatial data [64, 60]. One of the requirements was to provide some kind of technological infrastructure to help the annotation process, making it reproducible. Considering this, we decided to adopt scientific workflows as a means to implement this process.

"Scientific workflows have emerged as a paradigm for representing and managing complex distributed scientific computations. Such workflows capture the individual data transformations and analysis steps as well as the mechanisms to carry them out in a distributed environment" [76]. Workflows proved to be a good choice to help automate the annotation process of geospatial data. However, at the same time, they presented new challenges, given the complexity demanded by these annotations.

This paper discusses these challenges, focusing on how we use these workflows to semantically annotate geospatial data. We also give a brief overview of our annotation process, explaining the implementation choices made in order to address these challenges.

The main contributions of our work are therefore: (1) we show how scientific workflows can be used to orchestrate annotation of geospatial data sources, thereby alleviating experts' tasks; (2) we discuss the challenges we faced, establishing their connection with those defined by an NSF[1] workshop [76]; and (3) we present implementation solutions to some of these challenges, which involve, among others, appropriate specification of interfaces and Web Services.

The rest of this paper is organized as follows. Section 6.2 defines our semantic annotations and briefly presents the semantic annotation framework. In section 6.3, we explain annotation workflows and how they are specified. Section 6.4 presents the challenges identified by NSF and how they are reflected in our annotation workflows. The choices made to solve these challenges are discussed in section 6.5 and the implemented prototype in section 6.6. Section 6.7 describes some related work. Finally, section 6.8 presents our conclusions and ongoing work.

6.2 The Semantic Annotation Framework

6.2.1 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 [64]:

---

[1] USA National Science Foundation, www.nsf.gov

Annotation Units. An *annotation unit* **a** is a triple $< s, m, v >$, where s is the subject being described, m is the label of a metadata field and v is its value or description.
Annotation. An *annotation* **A** is a set of one or more annotation units.
Semantic Annotation Units. A *semantic annotation unit* **sa** is a triple $< s, m, o >$, where s is the subject being described, m is the label of a metadata field and o is a term from a domain ontology.
Semantic Annotation. A *semantic annotation* **SA** is a set of one or more semantic annotation units.
Annotation Schema and Content. An annotation/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.

6.2.2 Framework Overview

Geospatial data are essential to different science domains, being processed to provide information for a multitude of purposes and contexts. For instance, the same (temporal) series of satellite images can be used: (1) by agricultural experts to examine the extent of a crop; (2) by environmental researchers to check the effects of that crop on a preserved vegetation area; (3) by sociologists, to analyze human occupation in the area, as witnessed by crop spatial evolution. Hence, the same geospatial data source can have several kinds of annotations, for the different domains being considered. This was a challenge for our work. How to provide a framework to enable the semantic annotation of different kinds of geospatial data, for different domains?

Our steps to produce annotations closely follow procedures of manual annotation available in Geographic Portals, such as FAO [26] and GOS [93], in which specific annotation schemas are defined to describe the data. Experts will subsequently fill the schema, for each data source, with appropriate information. The result is very much like our sets of annotation units.

A basic premise of our work is that the geospatial component, e.g. coordinates, is free from ambiguity, and can therefore be used to speed up the annotation process, alleviating the task of expert analysis. The second basic premise is that, 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. Having this in mind, we decided to use scientific workflows

to store these core procedures, thereby enabling their repeatability, sharing, reuse and adaptation to new contexts.

Given these premises, our annotation scenario is the following. First, experts need to predefine core annotation procedures for each kind of geospatial data source (e.g., thematic maps, satellite images, sensor time series). Then, every time a given data source needs to be annotated, the corresponding workflow is executed, generating a basic annotation, which may be subsequently validated by experts. Moreover, such workflows can be specialized for particular needs (e.g., considering a given crop in agriculture, or a given habitat characterization in biodiversity). Each workflow contains information on the annotation schema and ontologies to be used, the operations to perform and how to store the generated annotations.

Figure 6.1 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 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. During this process, the annotation units are presented for user validation, usually a domain expert, who may choose another workflow, adapt the workflow or define a new one. In the third step, appropriate ontology terms are chosen to assemble the semantic annotations (linking annotation units to ontology terms). The semantic annotations are stored as RDF[2] triples in an XML database, where they can be used for information retrieval, e.g. using XQuery statements [88].

6.2.3   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

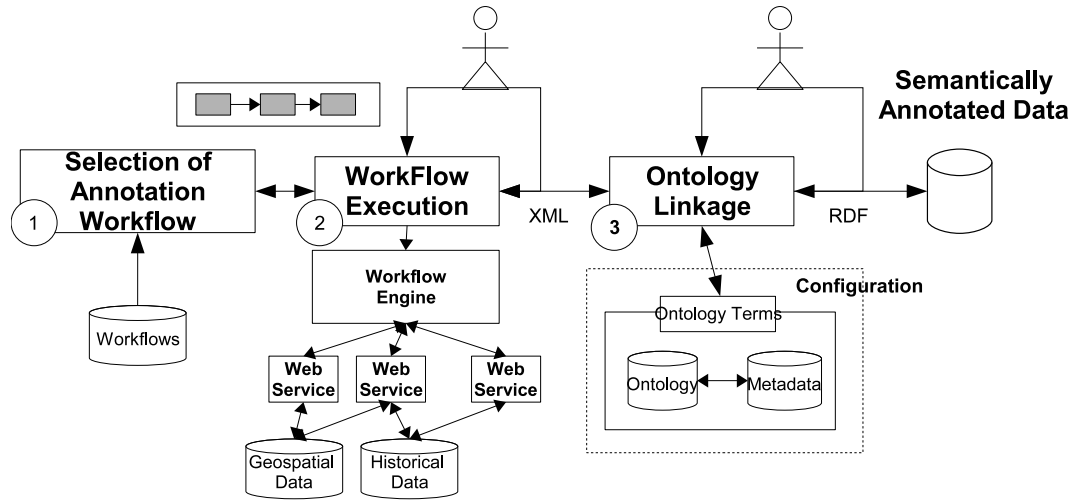---

[2]Resource Description Framework

Figure 6.1: GeoSpatial Data Annotation - Main steps

support.

6.3 Annotation Workflows

6.3.1 Specification of the Workflows

Our framework supports the semi-automatic annotation of different kinds of geospatial data sources. There is a different scientific workflow to describe the annotation process for each kind of data. The annotation workflow has to identify features to be considered for each kind of source data, the metadata schema to be used, and also the computations (tasks) involved in the production of annotation units.

The creation of annotation workflows involves many steps. All of them have to be performed by computer scientists in cooperation with domain experts. The former know how to design a workflow. The latter know which information is important for each source, which services to invoke to obtain this information, the auxiliary data sources to be accessed and the ontology terms to be used.

The first step is the choice of metadata schema to be adopted in annotations. This choice will depend on the domain of the data to be annotated. We adopted the FGDC[3] geospatial metadata standard [28], which is general purpose and open. Once this is done, the specification of the workflows starts, being dependent on the kind of data source, and its context of use. Finally, it is necessary to designate (or to implement) the modules/services that are invoked by workflow activities to produce the desired annotation

---

[3]Federal Geographic Data Committee

units.

6.3.2 Annotating a Geospatial Data Source

Figure 6.2 illustrates a graph plotting NDVI[4] values for sugar cane. Such graphs are used in crop productivity estimates, in agriculture. Such a time series is obtained from a sequence of satellite images, which have been preprocessed, so that their pixels contain NDVI values. Let r be a specific region selected in all images, and $p(r,t)$ denote the average pixel value within region r, for the satellite image with timestamp t. The (NDVI) series for n images is given by $<p(r,t_1), p(r,t_2)..., p(r,t_n)>$. The geospatial component is obtained from r, defined as a list of coordinates.
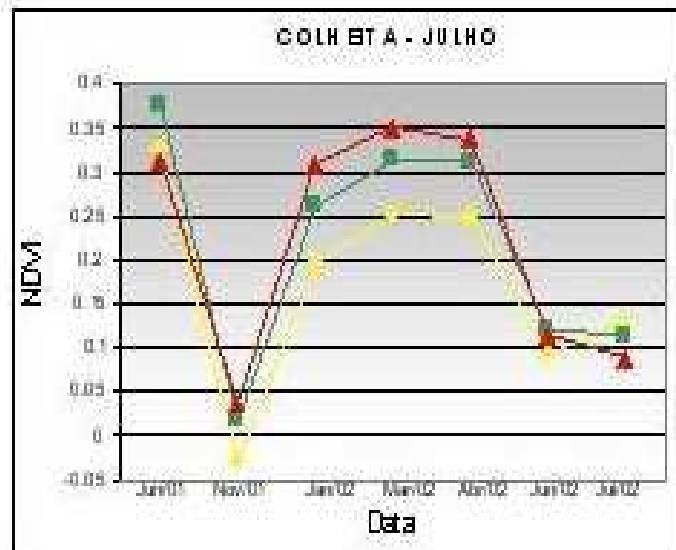
Figure 6.2: An NDVI time series for sugar cane

Figure 6.3, from [64], illustrates a workflow for semantic annotation of NDVI time series, specified using WOODSS, a workflow tool [70] that provides means to edit and manage scientific workflows. The first task in the workflow of Figure 6.3 is responsible for retrieving the metadata schema to be used to generate the annotation. Then, it is necessary to get the provenance data needed to produce the related annotation content – e.g., information on the satellite images used to generate the series (such as sensors used), and region coordinates. From the coordinates, it is possible to obtain the county name; and using a specific tool that compares temporal series [68], it is possible to retrieve

---

[4]Normalized Difference Vegetation Index – a value computed from pixels of satellites images, which indicates the amount of biomass in a region

similar series, which were previously annotated. From these annotations, it is possible to suggest additional annotation content – e.g., crop name. The possible crop names are presented to the user (usually an expert) to indicate the correct one. Considering this name, the time period and the county name, historical data are analyzed to obtain production values. These facts, and others like climate information, are then stored as annotations, to be transformed into semantic annotations.
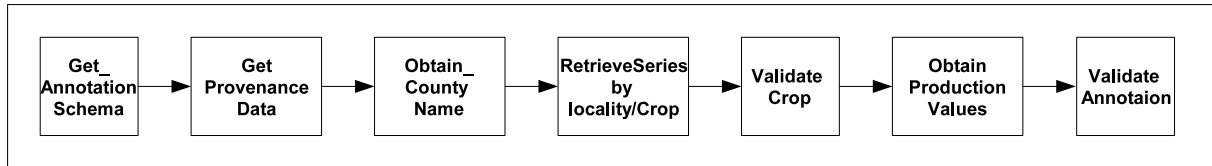
Figure 6.3: A workflow for semantic annotation of an NDVI time series

Figure 6.4 presents part of the semantic annotations produced for an NDVI series. The schema of this annotation is based on the FGDC metadata standard [28], which was extended for agriculture domain. Filed contents are based on ontologies, such as POESIA Agricultural Zoning ontology [29] and SWEET [73]. For example, the Figure shows that the series is related to sugar cane, which is identified by the pair $<crop>$, $<rdf:\ resource>$, and includes the harvesting period – $<harvestingbegdate>$, $<rdf:resource>$ – started at February, 21th, 2006. For more details on the adopted annotation schema, the annotation units and the ontologies used, see [64].

```
<!-- Crop Identification -->

<crop>
<rdf:Bag>
<rdf:li rdf:resource="http://www.owl-
ontologies.com/unnamed.owl#SugarCane"/>
<rdf:li>Sugar Cane</rdf:li>
</rdf:Bag>
</crop>
<!-- Harvesting Period -->

<harvestingbegdate>

<rdf:Bag>
<rdf:li
rdf:resource="http://sweet.jpl.nasa.gov/ontology/time.owl#Start"/>
<rdf:li>20060221</rdf:li>
</rdf:Bag>
</harvestingbegdate>
```

Figure 6.4: Semantic annotation units generated for an NDVI time series

6.4   Challenges Involved

6.4.1   NSF Challenges and Workflows

In 2006, the National Science Foundation (NSF) organized a workshop to discuss the requirements of future scientific applications and the challenges they represent to existing workflow technologies. This workshop brought together domain, computer and social scientists, which grouped the identified challenges in four main topics: application requirements, workflow representations, dynamic workflows and system-related issues[76]. This identification of challenges resulted in a set of recommendations involving representation, collaboration and computer science issues, as a way to foster scientific workflow usage. All of them are described in [76, 34]. These challenges are briefly presented in the following.

Applications and requirements

"Given the growth of computations, sensors, database and others, why the growth of scientific data analysis and understanding is not proportional?"[76]. First, workflows must enable collaborative research, combining distributed data, computations, models and instruments. They orchestrate the steps of scientific discovery and collaborations, alleviating researchers from manual tasks. Second, workflows have to provide means to reproduce scientific analysis and processes. This involves the access to distributed and heterogeneous provenance data. These data have to be stored, as well as metadata, to enable the reproducibility and the discovery of the workflows and applications used to create the experiments. Finally, the environments have to be flexible to support different users and analysis.

Data and workflow descriptions

"Given the existing practices and benefits of sharing instruments, data and computing, why don't researchers capture and share scientific computations and processes as well?" [76]. Workflows have appeared as a good means for sharing information and processes. However, they should be described using a representation commonly accepted by the scientific community. This representation should also accommodate scientific process descriptions at multiple levels and enable workflow variants, incorporating information about analysis processes to support their discovery, creation, merging, and execution.

Dynamic workflows and user steering

"How can workflows support both the exploratory nature of science and the dynamic process involved in scientific analysis?"[76]. It is necessary to develop mechanisms to support

dynamic workflows that evolve over time. These mechanisms should respond appropriately to external events, such as different kinds of data being processed. Workflow systems should provide a user interface that allows scientists to explore the processes involving dynamic workflows, such as querying. All these can facilitate the task of development of workflow patterns, which can lead to reuse, continuous improvement, and sharing.

System-level management

" *Given the continuous evolution of infrastructure and technology, how can we ensure reproducibility of computational analysis over a long period of time?*"[76]. Scientific workflows should ensure reproducibility, with equivalent results for the same initial data. This requires knowledge of data manipulations, software being used and execution environment, but also stable systems.

6.4.2 Challenges in Annotation Workflows

We now revisit our annotation workflows, sharing how they present all of the NSF challenges.

Applications and requirements

Sharing and reproducibility of annotation workflows   This is a very difficult task, common to all scientific workflows. A given geospatial data source cannot originate diverging annotations, if the context is defined. Hence, the annotations produced have to be equivalent whenever the workflow is reexecuted for the same entries. One problem is that users are allowed to intervene at workflow execution, which may produce different entries for a given geospatial source. Another difficulty is workflow sharing. Some workflows may be reused in different domains, but how to share them? How to describe the goal of each workflow, using a language free of ambiguity, providing a common understanding and allowing their retrieval? (i.e., how to annotate the annotation workflows?).

Management of different kinds of data   The production of annotations involves the interpretation and combination of different kinds of data, such as spatial information, and provenance. This could be a problem, since the transformation process has to know how these data are stored and in which database(s) or repositories.

Linkage to ontology terms   The generation of semantic annotations requires the selection of ontologies, and their terms, to be linked to the annotations. They have to be

well-known, consensual and adherent to the domain. The automatic specification of this linkage process using workflows constitutes one of the main challenges of our work.

Data and workflow descriptions

The definition of each annotation workflow directly involves the collaboration of multidisciplinary research groups. These workflows have to be described using a representation that can be understood by all these scientists, considering different levels of abstractions.

Dynamic workflows and user steering

Human validation of a workflow's result.   A recurrent problem in annotation workflows is the need for intervention of the domain expert, during workflow execution, in order to validate the generated annotation. The workflow engine to be adopted has to allow this.

Evolution of workflows.   An annotation process (and thus the workflow) can evolve by many reasons: addition of new features or tasks, changing of implemented services and technology evolution. Workflow systems have to be prepared for this, enabling and facilitating this evolution.

System-level management: binding of a workflow's specification and executable services

One of the main problems related to the execution of our semantic annotation workflows is the binding between their specification and the services that execute their tasks. This problem can be divided into two subproblems: the discovery of services, and the syntactic matching between tasks and services' interfaces. Besides these problems, users may have to deal with specific peculiarities of each workflow engine. When using YAWL [94], for example, users have to write Java code that wraps Web services, in order to allow the engine to invoke them.

   The next section presents the choices made to address these challenges. We recall that tasks in our annotation workflows are executed by invocation of Web services, to enhance interoperability, and meet some of the challenges discussed here.

6.5   Addressing the Annotation Workflow Challenges

6.5.1   Applications and requirements

Sharing and reproducibility

In WOODSS[5], workflows are annotated to allow their sharing and reuse [96]. These workflows, and their annotations, are stored in the PostGreSQL DBMS. This allows the automatic selection of the appropriate workflow to be performed – see step 1 of Figure 6.1, which can be retrieved according to the annotations attached to it. This also helps the sharing of workflows.

Reproducibility is a difficult issue to address, mainly because we need provenance information to annotate geospatial data sources. This involves different issues related to propagation of annotations, as those posed by [9].

Following the recommendations of [19], we store the provenance data used. Besides this, we are using only data from curated databases sources, which ensures a certain level of data quality.

Management of different kinds of data

During the annotation process, services invoked by workflow tasks have to access different kinds of data. This could be a problem, as the service has to know how such data are stored and where. To facilitate this, our framework contains a special layer between annotation services and data sources. This layer works as a mediator, being responsible for accessing distinct DBMSs, such as PostGreSQL for relational data and workflows, PostGIS for spatial data, and XML databases. Through the methods provided by this layer, the access to the data is performed in a transparent way, regardless of how the data are stored.

Linkage to ontology terms

Our semantic annotations are generated from annotations, using ontology terms – classes and their instances. To automate this transformation, our framework requires a configuration phase, in which the ontologies to be used are selected by an expert. When an ontology is selected for use, the framework extracts all its terms. Then, the user is asked to indicate which term(s) can be related to each annotation schema field. At the end, the configuration phase stores in a database the identification of the chosen terms, and the label of associated annotation fields. This is used subsequently for providing the semi-automatic linkage to ontology terms.

---

[5]see http://woodss.lis.ic.unicamp.br/woodss/

6.5.2 Data and workflow descriptions: collaboration among multidisciplinary research groups

The generation of semantic annotations for geospatial data may involve research groups from different areas, such as biologists and ecologists (domain experts), and computer scientists, who provide computational solutions.

We use WOODSS to edit the workflows, since this is an environment easy to use, with a clear representation language, and it supports annotations of workflows and their storage in a database. It provides means for domain experts to express and share their annotation workflows in a friendly way. Once the workflows are specified, computer scientists can develop the computational solutions needed for the workflows' execution. WOODSS does not have a native execution engine, and its workflows have to be exported for execution. We are now using the YAWL environment [94] to execute the specified workflows.

6.5.3 Dynamic workflows and user steering

Human validation of a workflow's results.

During the execution of an annotation workflow, users are prompted to validate the produced annotations. The procedures to support the validation dialog were implemented as an external application, called *GeoNotes Dialog*. This module (see Figure 6.5) allows the binding of a workflow's task to forms with annotation units to be presented to the user for validation or completion. This facilitates the interaction between the workflow engine and domain experts. The module works as an interface between the implemented services and the user. Hence, every time the user has to interact with the workflow system – providing annotation content, for example – the YAWL engine invokes *GeoNotes Dialog* as a standard web service. The Figure shows that the user is being prompted to validate annotation units whose labels are locality, crop and production.

Evolution of workflows

At present, we are considering different versions of a workflow as different workflows. This may be extended to consider a versioning mechanism as a means to support workflow solution.

6.5.4 System-level management: binding of a workflow's specification and executable services

There is considerable research being developed to answer this challenge, such as automatic discovery and composition of services. However, this field remains still open. We adopted
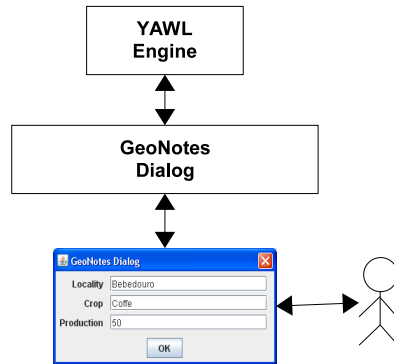
Figure 6.5: GeoNotes Dialog - a service invoked by YAWL to support user interaction with annotations.

a solution in which researchers publish and share their Web services in a collaborative way. In our system, users can publish their Web services in a service repository, which we call *Service Adapter* (Figure 6.6).

In order to bind a Web service to a YAWL workflow task, the user needs to encapsulate the Web service stub using a specific servlet interface provided by the YAWL API [53]. Hence, every time the user wants to adopt a new Web service, this interface has to be implemented. We implemented the *Service Adapter* to facilitate this task; it is a generic service wrapper that enables YAWL to use new services. The user uploads the WSDL file of the Web service to the Service Adapter, and the Adapter generates the Java stub that allows the engine to invoke that service. Thus, users just have to edit the workflow specification and bind its tasks to the Service Adapter, which mediates the Web service invocations.
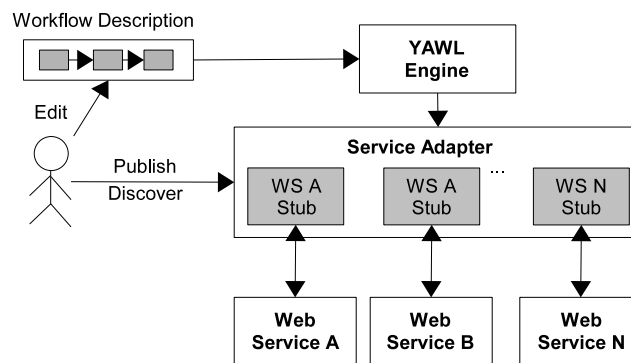

Figure 6.6: Service Adapter.

Once a Web service stub is added to the *Service Adapter*, any workflow is able to invoke this service through the Adapter. If a user wants to reuse an annotation workflow

specified by others, he/she can access WOODSS to discover workflows of interest, and then obtain the executable version of the workflow. Next, in order to execute the workflow, the user can: (i) add the required Web services to its own *Service Adapter* or (ii) register the *Service Adapter* used by the creator of the workflow on his/her engine.

Hence, the Service Adapter also facilitates the discovery of services that process or generate a given type of data.

6.6   Prototype Implementation

We have implemented a prototype as a proof of concept of how we met annotation workflow challenges. Figure 6.7 illustrates the architecture of the prototype. As mentioned before, we adopted the YAWL workflow engine for workflow execution. The Figure highlights structures of the execution stage, where the workflow (shown in the middle of the Figure) has been already retrieved from WOODSS' database. The Java language was adopted to implement web services and ontology linker. The latter is described in [20].
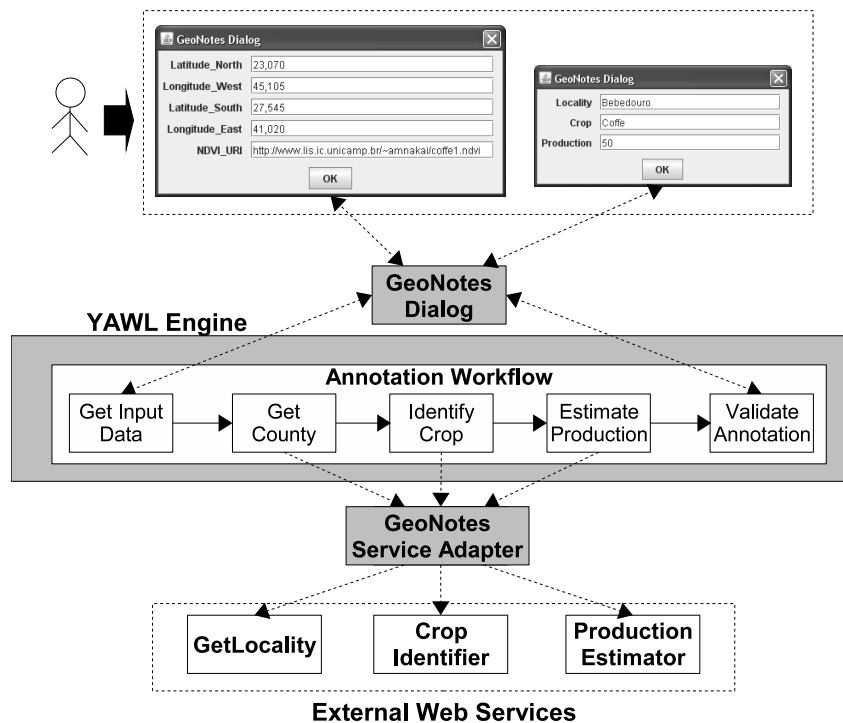

Figure 6.7: Prototype Architecture - Execution.

In this prototype, we focused on the functionalities needed for workflow execution. The *GeoNotes Service Adapter* was partially implemented. Using the current implementation,

a user is able to publish a service which can then be accessed by the YAWL engine. The GeoNotes Service Adapter uses Axis2 tools for automatic generation of Web service stubs. These stubs are wrapped with a specific Servlet interface (*InterfaceBWebsideController*) [53], which is provided by the YAWL API. This functionality fosters the execution of the workflows, because users do not need to write manually the Java code that wraps the service.

GeoNotes Dialog was developed as an external application that receives invocations from the YAWL engine through a component that is also wrapped by the *InterfaceB-WebsideController* interface. When GeoNotes Dialog receives an invocation, it generates a dialog form that requests inputs from the user. In the current implementation, only textual fields can be requested from the users – such as county and crop names, or yield values.

We also implemented a set of Web services to invoke some typical activities of annotation workflows, for researchers in agriculture. These Web services are implemented using the Axis2 framework and are executed on a Tomcat server. The interfaces of three services already implemented were:

- GetLocality service: returns the name of a given county, given its coordinates;

- CropIdentifier service: compares an input NDVI time series with other already annotated series; the most similar series are retrieved, and through their annotations, it is possible to suggest names for the kind of crop related to the input time series;

- Production Estimator: estimates the yield of a crop in a given locality, based on the kind of crop and historical yield information for that locality.

In our experiments, services do not invoke external modules. Rather, they respond invocations with synthetic data. The goal was to validate the architecture. The next step is to implement the services themselves.

6.7 Related Work

Our paper concerns the use of scientific workflows to orchestrate the semantic annotation of different kinds of geospatial data. This section presents related work concerning these issues, which comprises semantic annotation tools and semantics in scientific workflows.

6.7.1 Annotation Tools

Annotation of digital content is not an easy task, always subject to errors. This led to the development of tools that aim to facilitate the annotation process. We have tested and

compared some of them, taking into account the requirements pointed by [84] and [92]. Embrapa Information Agency [89], Amaya [98], KIM [80] 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 [13] and CREAM [41] also present methods for semantic annotation of visual resources.

In geographic applications, annotations should also consider the spatial component, since geographic information associates objects and events to localities, through a rich vocabulary of places and geographic object names, spatial relationships and standards. Hence, the geospatial annotation process should be based on geospatial evidences – those that conduct to a geographic locality or phenomenon, e.g. see [7, 47]. E-Culture [43], OnLocus [7], SPIRIT [48] and Semantic Annotation of Geodata [50] are approaches that consider the spatial component for the annotation of digital contents. However, they do not consider other kinds of content, like maps or graphs for annotation.

Like these tools, we use ontologies for annotation. Unlike them, we combine several components in our framework to facilitate the annotation process and to enhance the reuse of annotations. Moreover, our framework is extensible and general purpose. This generality requires a tool to orchestrate the execution of these components for the different kinds of data being considered.

6.7.2   Semantics in Scientific Workflows

Scientific workflows have appeared as a good answer to the challenge of data discovery and integration. They provide a way to perform data analysis so that results can be reproduced and the method can be reviewed, validated, repeated, and adapted [37]. Considering this, we use them to orchestrate the semantic annotation of geospatial data. In spite of intensive research on scientific workflows, we found no discussion of their use in performing annotations.

We identified the same classes of challenges, although in some cases involving new issues. Gil et.al [34] discuss the challenges for scientific workflows presented on the NSF workshop, pointing out that reproducibility is probably the main one. Provenance is considered an essential component for reproducibility, sharing, and knowledge re-use in the scientific community. Davidson and Freire [19] provide an overview of research issues in provenance for scientific workflows, classifying the kinds of provenance data, describing existing research on provenance for scientific workflows and outlining open problems and new directions for database-related research. These are also our main problems. We have to deal moreover with semantic issues.

Fox and Hendler [32] point out that all capabilities needed by eScience – including data

integration, fusion, mining, workflow development orchestration and execution; capture of provenance, lineage, and data quality – require semantic representation and mediation. Following this, semantics have also been a concern in workflow-related research, involving two main issues, both of which considered to be hard problems: the automatic discovery and composition of services; and the automatic discovery of workflows. Rao and Su [83] give a good survey of existing automated methods for Web services composition. Cardoso and Sheth [10] propose an approach based on the use of ontologies to describe workflow tasks and Web service interfaces, as a way to enable their discovery and integration during workflow execution. Lopes et al. [56] present an approach based on workflows and on Web 2.0 principles to integrate and coordinate services and data sources, aiming to simplify knowledge extraction in different scientific areas. Dong and Wild [21] describe a tool for generation of workflows for automatic discovery of drugs. This tool uses semantic Web service technologies, extending standard services with semantic annotations, for semantic interoperability. More recently, Fujii and Suda [33] propose a framework for dynamic composition of services based on the semantics of components and contexts of users.

The use of semantic representations for discovery and retrieval of workflows has also been addressed, but at a lower level. As pointed by [38], much has been written of web service discovery, but descriptions existing on services and captured in workflows are ignored. Following this line, Gil et al. [36] propose the use of semantic representations to compactly describe complex scientific applications. Based on these descriptions, workflows are automatically generated and the computations are mapped to available computing resources. These descriptions enable search for previous similar workflows. Part of this work was extended [35] to propose the augmenting of workflow descriptions with constraints derived from workflow components and the used data, as a way to enhance workflow discovery.

6.8   Conclusions and Ongoing work

Geospatial data are a basis for decision making systems. However, these data have to be interpreted to be used. Even when recorded, this interpretation is hard to understand; this increases the cost of decisions made on such data. The absence of approaches to efficiently store these interpretations leads to problems such as rework and difficulties in information sharing. To alleviate the task of data annotation, we proposed a framework for semi-automatic semantic annotation of geospatial data. This approach was outlined in [60]. Our approach involves a group of challenges, such as the automatic linkage of annotation units to ontology terms and the annotation of different kinds of geospatial data. As an answer to the last one, we adopted scientific workflows to orchestrate the annotation process.

This paper presented and discussed the challenges for semantic annotation of geospa-

tial data using scientific workflows, and the choices made to address them. A prototype was implemented to validate our choices. Ongoing work includes several directions, such as to consider other features of geospatial data being tested and to improve the annotation validation through YAWL.

Capítulo 7

Conclusões

Esta tese combina diversos aspectos de pesquisa, incluindo anotações semânticas de dados geoespaciais, automação do processo de anotação, interoperabilidade e heterogeneidade de dados geoespaciais, produção de informação, *design* de workflows científicos, esquema de metadados geográficos e uso de ontologias para descrição de informação geoespacial. O resultado é a especificação de um mecanismo para anotação semântica de dados geoespaciais, com foco na agricultura. O mecanismo prevê, dentre outros recursos, a anotação de diferentes tipos de dados, numa maneira semi-automática. A pesquisa especificou e implementou parcialmente um *framework* para apoio ao processo de anotação, que teve como principais requisitos ser genérico e extensível. Alguns experimentos mostraram a utilidade do framework para anotar imagens de sensoriamento remoto e séries temporais de NDVI, validando o processo de anotação para demandas e aplicação em agricultura.

7.1 Contribuições

Esta tese tem como principais contribuições:

1. Análise de catálogos e portais de dados geográficos, resultando na identificação dos requisitos para que permitam a busca semântica aos dados que eles provêem (Capítulo 3). Alguns desses catálogos, inclusive, apresentaram problemas de acesso, por queda de servidores.

2. Análise de ferramentas existentes para anotação de dados e identifição de requisitos para anotação semântica de dados geoespaciais, sejam eles textuais ou não (Capítulo 4).

3. Especificação e implementação parcial de um *framework* para anotação semântica de dados geoespaciais. Isso requer a especificação de workflows de anotação e a

89

definição das anotações básicas para cada fonte de dados considerados. Para isso, é necessário: a indicação do esquema de metadados adotado, bem como sua extensão para acomodar informações agrícolas; etapa de configuração do ambiente, incluindo a seleção de ontologias a serem utilizadas; e indicação do formato usado para armazenar as anotações (Capítulo 5).

4. Uso de workflows científicos para orquestrar o processo de anotação de cada tipo de dado sendo considerado. Isso envolve um conjunto de desafios, que são discutidos e cotejados com desafios propostos em um workshop da NSF. Para cada um dos desafios, esta tese apresenta uma solução de implementação (Capítulo 6).

   Estas contribuições abordam os seguintes aspectos: Levantamento de requisitos para acesso a dados geoespaciais (Capítulo 3); Anotação tradicional destes dados (Capítulo 4); Abordagem alternativa para anotação destes dados (Capítulo 5); e, Desafios na descrição do processo de anotação (Capítulo 6).

Um dos aspectos abordados foi a extensão do FGDC para aplicações em agricultura. Os campos adicionais identificados foram cultura, localização da cultura, produtividade média, fases da produção (incluindo o período de cada uma) e tipo de solo.

7.2   Extensões

Esta tese apresenta muitas possibilidades de extensões, abrangendo desde o mecanismo de anotações proposto e o esquema de metadados utilizado, até o framework implementado. Dentre elas destacam-se:

- **Outras extensões do esquema de metadados adotado.** O esquema básico de metadados adotado é bem amplo, sendo genérico para dados geográficos. Assim, visando uma melhor descrição dos dados sendo anotados, decidimos estendê-lo, como apresentado no capítulo 5. Entretanto, é necessário um estudo com mais tipos de dados, de maneira a identificar todas as possibilidades de informação necessária. Além disso, focando em interoperabilidade, os campos desta extensão devem ser descritos usando termos de ontologias, de maneira a permitir seu amplo uso.

- **Adaptação do *framework* para outros domínios**. O framework foi desenvolvido e parcialmente testado para o domínio agrícola. Entretanto, imagina-se que ele possa ser facilmente estendido para outros domínios. Algumas questões envolvidas: adaptação do esquema de metadados para o domínio em questão; seleção de tipos de dados a serem anotados e identificação de suas características; seleção de ferramentas a serem usadas na produção da informação. Espera-se que alguns dos serviços já em uso possam ser reusados em outros domínios.

- **Tradução automática de um Workflow especificado no WOODSS para o YAWL.** O framework de anotação prevê a especificação inicial do workflow de anotação usando o ambiente WOODSS. Isso deve-se principalmente à sua facilidade de uso e descrição, permitindo que um especialista do domínio possa especificar um workflow com uma descrição inicial do processo de anotação de um tipo específico de dado. Entretanto, para ser executado, este workflow precisa ser traduzido para um workflow no ambiente YAWL. Neste sentido, seria muito útil um tradutor automático WOODSS-YAWL.

- **Validação da anotação durante a execução do Workflow.** O uso de workflows científicos para descrever o processo de anotação apresenta muitos desafios, como descrito no capítulo 6. Dentre eles, destaca-se a validação da anotação gerada. Durante a execução do workflow, a anotação gerada (parcial ou não) pode ser apresentada ao usuário para sua validação. Atualmente apenas a anotação textual pode ser apresentada. Entretanto, como o ambiente pode ser estendido para acomodar outros tipos de anotações, como imagens ou sons, seria importante mostrá-las ao usuário.

- **Anotação de partes do dado.** O mecanismo proposto prevê a anotação de um arquivo como um todo (imagem, série temporal, mapa). Entretanto, pode ser útil anotar também partes do arquivo. Por exemplo, na anotação da imagem de sensoriamento remoto apresentada no capítulo 5, o resultado do processo de análise da imagem é um conjunto de polígonos, cada um deles relacionado a uma cultura. Assim, cada um destes polígonos poderia ter sua anotação específica. Da mesma forma, partes de uma série histórica poderiam ser parcialmente anotadas.

- **Implementação e Integração.** O framework foi apenas parcialmente implementado. Partes do processo de anotação ainda são realizadas manualmente e os serviços não foram implementados (apenas suas interfaces, para validar a invocação). Em particular é necessário um maior esforço em: (i) associar os termos de ontologias a metadados – apesar do mecanismo de indexação/extração de termos da ontologias estar pronto, ainda precisa ser refinado para associar estes termos às unidades de anotação; (ii) refinar os serviços web disponíveis para anotação de dados, tornando-os mais simples e genéricos; (iii) integrar os serviços e ferramentas em uso.

- **Versionamento de workflows.** Anotações podem evoluir ao longo do tempo, o que exigirá dispor de um sistema de versionamento de workflows. Workflows podem também ser especializados ou generalizados, o que se aplica geralmente à questão de versionamento.

- **Metodologia para configuração de uma anotação.** A especificação de um workflow de anotações, das ontologias a serem usadas e dos serviços a implementar são atividades demoradas que exigem grande investimento de tempo de especialistas. Uma outra direção de pesquisa, portanto, é definir uma metodologia nesta etapa, para facilitar o trabalho dos especialistas do domínio e da computação na tarefa de configurar workflows.

Referências Bibliográficas

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

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

[3] Chaitanya Baru, Vincent Chu, Amarnath Gupta, Bertram Ludäscher, Richard Marciano, Yannis Papakonstantinou, and Pavel Velikhov. XML-based information mediation for digital libraries. In *DL '99: Proceedings of the fourth ACM conference on Digital libraries*, pages 214–215. ACM, 1999.

[4] H. Beck and H. S. Pinto. *Overview of Approach, Methodologies, Standards, and Tools for Ontologies (DRAFT)*. FAO - Agricultural Ontology Service Project, april 2002. Disponível em: <http://www.fao.org/agris/aos/Documents/BackgroundPaper.pdf>.

[5] T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. *Scientific American*, pages 34–43, May 2001.

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

[7] K. A. V. Borges, A. H. F. Laender, C. B. Medeiros, and Jr. 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.

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

[9] P. Buneman, J. Cheney, W. Tan, and S. Vansummeren. Curated databases. In *PODS '08: Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART*

symposium on Principles of database systems, pages 1–12, New York, NY, USA, 2008. ACM.

[10] J. Cardoso and A. Sheth. Semantic e-workflow composition. *Journal of Intelligent Information Systems*, 21(3):191–225, 2003.

[11] A. C. P. L. F. Carvalho et al. Grand challenges in computer science research in brazil – 2006 – 2016 workshop report. Technical report, Brazilian Computer Society, May 2006.

[12] CEPEA. *Center of Advanced Studies in Applied Economics (CEPEA)*. ESALQ/USP, 2008. <http://www.cepea.esalq.usp.br/pib/>. Accessed in: 25 march 2008.

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

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

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

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

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

[18] J. Daltio, C. B. Medeiros, L. C. Gomes Jr, and T. Lewinsohn. A Framework to Process Complex Biodiversity Queries. In *Proc. ACM Symposium on Applied Computing (ACM SAC)*, March 2008.

[19] S. B. Davidson and J. Freire. Provenance and scientific workflows: challenges and opportunities. In *SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1345–1350, New York, NY, USA, 2008. ACM.

[20] S. R. de Sousa. Management of semantic annotations of web data for agricultural applications (to be presented). Master's thesis, Institute of Computing - University of Campinas, Brazil, 2010.

[21] X. Dong and D. Wild. An automatic drug discovery workflow generation tool using semantic web technologies. In *Fourth IEEE International Conference on eScience*, 2008.

[22] 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.

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

[24] ESRI. Implementing a metadata catalog portal in a GIS network. Technical report, ESRI, march 2003.

[25] J. Euzenat and P. Shvaiko. *Ontology Matching*. Springer-Verlag New York, Inc., 2007.

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

[27] FGDC. FGDC Geospatial Metadata, 1998.

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

[29] 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.

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

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

[32] P. Fox and J. Hendler. *The Fourth Data-Intensive Scientific Discovery*, chapter Scientific Infrastructure – Semantic eScience: Enconding Meaning in Next-Generation Digitally Enhanced Science, pages 147–152. Microsoft Research, 2009.

[33] K. Fujii and T. Suda. Semantics-based context-aware dynamic service composition. *ACM Trans. Auton. Adapt. Syst.*, 4(2):1–31, 2009.

[34] Y. Gil, E. Deelman, M. Ellisman, T. F. G. Fox, D. Gannon, C. Goble, M.Livny, L. Moreau, and J. Myers. Examining the challenges of scientific workflows. *Computer*, 40(12):24–32, 2007.

[35] Y. Gil, J. Kim, G. Florez, and V. Ratnakara dn P. A. González-Calero. Workflow matching using semantic metadata. In *Fifth International Conference on Knowledge Capture (K-CAP)*, 2009.

[36] Y. Gil, V. Ratnakar, E. Deelman, and G. Mehtaa nd J. Kim. Wings for pegasus: Creating large-scale scientific applications using semantic representations of computational workflows. In *19th Annual Conference on Innovative Applications of Artificial Intelligence (IAAI)*, Cananda, 2007.

[37] C. Goble and D. D. Roure. *The Fourth Data-Intensive Scientific Discovery*, chapter Scientific Infrastructure – The Impact of Workflow Tools on Data Centric Research, pages 137–145. Microsoft Research, 2009.

[38] A. Goderis, P. Li, and C. Goble. Workflow discovery: the problem, a case study from e-science and a graph-based solution. In *IEEE International Conference on Web Services (ICWS '06)*, pages 312 – 319, 2006.

[39] J. Greenberg, K. Spurgin, and A. Crystal. Functionalities for automatic metadata generation applications: a survey of metadata experts' opinions. *Int. J. Metadata, Semantics and Ontologies*, 1(1):3–20, 2006.

[40] T. R. Gruber. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2):199–220, 1993.

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

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

[43] 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.

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

[45] IBGE. *Geographic and Statistical Brazilian Institute (IBGE)*. IBGE/USP, 2009. <http://www.ibge.gov.br/english/>.

[46] ISO. *ISO 19115:2003 Geographic information – Metadata*. ISO, 2008. Available on:<http://www.iso.org/iso/home.htm>.

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

[48] C.B. Jones, A.I. 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.

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

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

[51] E. Klien, U. Einspanier, M. Lutz, and S. Hübner. An architecture for ontology-based discovery and retrieval of geographic information. In *7th AGILE Conference on Geographic Information Science*, Heraklion, Greece, 2004. Parallel Session 3.1–Semantics I.

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

[53] S. Knelpp, l. Bradford, J. Prestdge, M. L. Rosa, and M. Adams. *Editor 2.0 User Manual*. The YAWL Foundation, 2008.

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

[55] J. Larson, M. Siliceo, M. Silva, E. Klien, and S. Schade. Are geospatial catalogues reaching their goals? In *9th AGILE Conference on Geographic Information Science - Poster*, Visegrád, Hungria, 2006.

[56] P. Lopes, J. Arrais, and J. L. Oliveira. Dynamic service integration using web-based workflows. In *10th Int. Conf. on Information Integration and Web-based Applications & Services (iiWAS '08)*, pages 622–625, USA, 2008. ACM.

[57] R. Lunetta, D. Johnson, J. Lyon, and Crotwell J. Impacts of imagery temporal frequency on landcover change detection monitoring. *Remote Sensing and Environment*, 89(4):444–454, fev 2003.

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

[59] C. G. N. Macário and C. B. Medeiros. Specification of a framework for semantic annotation of geospatial data on the web. In *XXIII Brazilian Symposium on Databases (SBBD 2008) - VII Workshop Thesis and Dissertations on Databases*, pages 1 – 8, Campinas, Brazil, October 2008.

[60] 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.

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

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

[63] C. G. N. Macário, A. M. Nakai, C. B. Medeiros, and E. Madeira. Using scientific workflows for semantic annotation of geospatial data: what are the challenges involved? Submitted for J.UCS - Journal of Universal Computer Science.

[64] 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.

[65] C. G. N. Macário et al. Crop monitoring via web: a successful case in multidisciplinary research. In *6o. Brazillian Congress of Agroinformatics - SBIAgro 2007*, 2007. (in portuguese).

[66] C. G. N. Macário and C. B. Medeiros. The geospatial semantic web: are gis catalogs prepared for this? In *5th International Conference on Web Information Systems and Technologies (Webist 2009)*, pages 335–340, Lisbon, Portugal, March 2009.

[67] C. Mangold. A survey and classification of semantic search approaches. *Int. J. Metadata, Semantics and Ontology*, 2:23–34, 2007.

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

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

[70] C. B. Medeiros, J. Pérez-Alcazar, L. Digiampietri, G. Z. Pastorello 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.

[71] H. J. Miller and J.i Han. Discovering geographic knowledge in data rich environments: a report on a specialist meeting. *SIGKDD Explor. Newsl.*, 1(2):105–107, 2000.

[72] A. M. Nakai, C. G. N. Macário, E. Madeira, and C. B. Medeiros. An infrastructure for sharing and executing choreographies. In *Proceedings of the 4th International Conference of Web Informations Systems and Technologies*, Portugal, May 2008.

[73] NASA. Semantic web for earth and environmental terminology (sweet), 2009.

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

[75] J. Nogueras-Iso, F.J. Zarazaga-Soria, R. Béjar, P.J. Á lvarez, and P.R. Muro-Med. OGC catalog services: a key element for the development of spatial data infrastructure. *Computers & Geosciences*, 31:199–209, 2005.

[76] NSF. Workshop on the challenges of scientific workflows , May 2006.

[77] OCG. CSW 2.0 FGDC application profile. Technical Report OGC 06-129r1, 2006.

[78] OGC. *Geography Markup Language.* The Open Geospatial Consortium, 2007. <http://www.opengeospatial.org/standards/gml>.

[79] Ontotext Lab. *The KIM Platform: Knowledge & Information Management.* 2007. <http://www.ontotext.com/kim/index.html>.

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

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

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

[83] J. Rao and X. Su. A survey of automated web service composition methods. In *Semantic Web Services and Web Process Composition*, number 3387 in LNCS, pages 43–54, 2005.

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

[85] N. Shadbolt, T. Berners-Lee, and W. Hall. The semantic web revisited. *IEEE Intelligent Systems*, 21(3):96–101, 2006.

[86] H. Silva. *MIG - Metadata for Geographic Information –Introduction to ISO 19115 standard.* Portuguese Geographic Institute, Portugal, February 2008. (in portuguese).

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

[88] S. R. Sousa and C. B. Medeiros. Management of semantic annotations of data on web for agricultural applications. In *VIII WTDBD - Workshop de Teses e Dissertações em Bancos de Dados*, Fortaleza,Brazil, October 2009.

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

[90] S. Spaccapietra, N. Cullot, C. Parent, and C. Vangenot. On spatial ontologies. In *Brazilian Symposium on GeoInformatics - GEOINFO*, Campos do Jordão, November 2004.

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

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

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

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

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

[96] A. F. Vitaliano F. Mechanisms for semantic annotation of scientific workflows. Master's thesis, Institute of Computing - Unicamp, July 2009. in portuguese.

[97] VRADSC. VRA Core 4.0, 2007. <http://www.vraweb.org/index.html>.

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

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

[100] S. Weibel, J. Godby, E. Miller, and R. Daniel. OCLC/NCSA Metadata Workshop Report. Web site http://www.oclc.org:5046/oclc/research/conferences/metadata/dublin_core_report.html, 1995.

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

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

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