ABSTRACT

In geographic information science, interoperability is a key issue, given the wide diversity of available geospatial data and of scientific data processing tools. There are many research initiatives to meet this challenge, e.g., in data interchange standards, service-oriented architectures (SOA), knowledge representation paradigms and user interface design. In parallel, a trend in systems design and development is to break down GIS functionalities into modules that can be composed in an ad hoc manner. This component-driven approach increases flexibility and extensibility. For scientists whose research requires geospatial analysis, however, all these efforts mean more than interoperability and flexibility. They are progressively shielding these users from having to deal with problems such as data representation formats, communication protocols or pre-processing algorithms. Once these scientists are allowed to abstract from these lower level concerns, they can shift their focus to the design and implementation of the models they are interested in. This paper analyzes how interoperability and componentization efforts have this underestimated impact on the user perspective, thereby affecting model development. This discussion is illustrated by the description of the design and implementation of WebMAPS, a geospatial information system to support agricultural planning and monitoring. By taking advantage of new results in the above areas, the experience with WebMAPS presents a road map to achieve seamless composition of distributed data sources and processing solutions, and to leverage model development.

Categories and Subject Descriptors
H.2.8 [Database Management]: Database Applications—scientific databases, spatial databases and GIS; H.3.5 [Information Storage and Retrieval]: Online Information Services—Data sharing

1. INTRODUCTION

Heterogeneity has been a problem in computational systems since technology allowed for two systems to interact. This problem has taken new dimensions in current scenarios where theoretically any two systems are capable of interaction. However, semantic mismatches on models of the world hampers interoperability both from the software interaction and on data representation points of view. Much progress has been made in providing interoperability for systems and data, mainly based on common agreements such as standards. Nonetheless, the problem is far from having a definitive solution, as requirements interoperability requirements evolve.

This paper discusses recent efforts in interoperability for geospatial systems and data. The discussion is intertwined with the impacts of the interoperability solutions in this area over geospatial model development. Our experience with some aspects of this discussion were tested with the development effort in the WebMAPS, a multidisciplinary project involving computer science and agricultural and environmental sciences aspects.

There are two basic interoperability points of view to be considered: systems and data. Systems interoperability is related on how two (or more) heterogeneous systems can interact. To that end, the systems must have means of determining which operations can/should be invoked from each other’s interface to execute a task. Data interoperability concerns data representation formats and manipulation. To achieve data interoperability both sides must be able to interpret one data set according to the correct concepts represented by it. Both points of view are largely related, since the operations take and generate data as input/output parameters.

Furthermore, we categorize interoperability approaches in three fronts: standards based, ontologies based and services based. Standards concern reaching an agreement on a domain and specifying interfaces, protocols and data representation formats... Ontologies are also related with a domain, but an a priori agreement is not always necessary... Services present a generic way of encapsulating operations from a system...

Systems design trend on GIS: modularization...
2. GEOSPATIAL DATA USAGE SCENARIO

The scenario for geospatial data usage has changed over time. In early applications, the paths which data sets went through were centralised within a single institution. As computer networks became more accessible, geospatial data sharing became more common. This data sharing scenario is shown in Figure 1, which specifies a data management cycle for GIS applications – from data acquisition (at the bottom) to data publication (at the top), to be consumed by applications. The data acquisition stage can rely on another such cycle (i.e., it accesses data that went through the same kind of processing to provide data at the end). The publication stage can also be the basis for the same kind of cycle.

The figure shows a full data management cycle in eight layers, each of which responsible for some kind of data storage of processing. The little boxes with horizontal dotted lines represent data sets and the ones with gears represent a data manipulation operation. The usual flow is from bottom to top, with the operations being applied to the data on their way up.

The bottom layer houses data providers of many kinds, including sets of file, databases, sensors and data services. Sensors can range from small ground-based sensor networks to large satellite embarked multi-spectral electromagnetic sensors.

Sensor-produced data pose several new challenges to geographic applications. These data has a variety of characteristics which impact how they are stored, processed, published and accessed [1, 2]. Among such characteristics we single out: (i) regularity: data may be produced in independent blocks or as continuous streams; (ii) spatial resolution: the spatial frame of a data set; (iii) temporal resolution: the time frame of a data set; (iv) transmission mechanism: e.g., manual readings, wired/wireless transmissions, transmission medium error introduction, and others; (v) position: impacts on the readings of the relative position of the sensor with respect to the observed phenomena; (vi) mobility: does the sensor move with respect to the observed phenomena and does the movement interferes with the readings. Many other characteristics could be considered important, according to the sensor capabilities and the application requirements. The broader is the coverage of these aspects, the larger is the number of consumers to which the data may be adequate.

Data services are another kind of data providers, delivering data products provided by organizations which created or enriched a given data set and made it available. These kinds of data also have inherent characteristics which are crucial to subsequent manipulation. Examples include which models were used to generate the software that produced the data, which parameters were applied to calibrate the model, or how reliable was the data over which the processing was done.

In this context, machine processable knowledge representation becomes an important issue. We believe that two approaches play an important role here: standards and semantics. These paradoxically conflicting approaches are further explored in Section 4, where we discuss how they “wrap” the data management cycles.

The second layer, from bottom to top, consists in unprocessed data, which can be, for instance, raw sensor collected data or data from other providers in different formats. In order to be used within the data management cycle, the data must be ingested in order to be represented in a compatible way, i.e., using the chosen characteristics and formats for the cycle. This task is performed in the next layer.

The third layer represents the pre-processing phase that
techniques, however, rigid limits. Annotations, on the other hand, impose spatial resolutions, and testing data format conversions to determine accuracy.

The fourth layer corresponds to the storage facility, often a data repository of some kind, such as a database system. Two of the major issues to be dealt with in this layer are problems on what to be stored and how to fill in the gaps left by several types of acquisition errors. Selecting what is going to be stored is important since the amount of data acquired may be far too large to be stored in full \([\text{?}, \text{?}]\). Gap filling is often hard to be done accurately \([\text{?}]\). Querying these data also brings forth the problem of combining data access mechanisms for both stored and streamed data \([\text{?}]\). Nonetheless, the storage layer may have its role filled by a proxy service with some kind of caching mechanism, which is more natural for many kinds of applications (e.g., real time traffic monitoring).

The fifth layer holds the pre-processing functions that are applied to the data so that these data will fit application requirements. Examples of requirements include spatial and/or temporal resolution, access rate, data format or more specialised manipulations. Instances of operations to fulfill these requirements include composition operations (e.g., fusion of data from different data sources), scaling operations (e.g., changing temporal or spatial resolution), customisation operations or even complex compositions of operations. However, as most of these data are georeferenced, the more traditional GIS operations, e.g., see \([\text{?}]\), are the most common in this phase.

The sixth layer houses the pre-processed data sets. Data format, spatio-temporal resolution, and associated description are the main concern here. The format is specified in the request and is translated as needed. The resolution adaptation may require interpolation algorithms for larger resolutions and summarization algorithms for smaller resolutions. Data description is achieved using either metadata or annotations. Metadata has a predefined structure and often expected ranges to the for filling the fields. This allows, for instance, indexing and retrieval based on the metadata, imposing, however, rigid limits. Annotations, on the other hand, have no structure and do not allow indexing, presenting a challenge for retrieval, often requiring content-based techniques \([\text{?}]\), allowing, however, very flexible descriptions. Section 4 discusses how to improve metadata semantics with standards and providing some structure to annotations without hampering flexibility.

The fifth and sixth layers did not appear in the early non-shared data scenario. Their appearance reveals an interesting issue: as models and algorithms become more stable and accepted within a community, their application starts to be considered basic and taken for granted. This moves the results of such models and algorithms from the application layer to the publication pre-processing layer with impact on interoperability and cooperative scientific work. An example of such migration is the georeferencing of satellite images, once done as a necessary step within geospatial data applications and currently available as a default resource in most of published images.

The seventh layer represents the publication software. This kind of software is responsible for providing data access mechanisms, which must somehow be known by the publication software and the client applications. This is achieve by previous agreement, which restricts interoperability with new systems, or by providing rich enough descriptions to allow the determination of adequacy. Different approaches are used depending on the requirements, e.g., protocols with less overhead for large data sets, or richer protocols for initial stages of communication.

Another important feature that must be available on the publication software is the selection of pre-processing operations to be applied on the data before transmission. As the requirements from the applications vary, many different operations may be needed before the data can be used. It is often undesirable or unfeasible to first transmit the data and then perform these operations on the application \([\text{?}]\).

The eighth and last layer is where the applications lie and where end users are able to interact with all the infrastructure on the layers below. Applications embed model execution, hence allowing scientists to visualize results, and to interact with and tune these models.

Although uncommon, it is possible for an application to access the data in any layer of the figure, for instance skipping the pre-processing or storage parts. This occurs mainly when response time issues are at stake, as each layer imply an overhead that may be too expensive.

In this scenario, all the described interoperability problems must be dealt with. Section 3 discusses approaches and solutions to these issues. Section 4 explores how standards and semantics are being employed to enhance existing interoperability solutions.

3. THE WebMAPS PROJECT

The WebMAPS is a multidisciplinary effort involving computer scientists and experts on agricultural and environmental sciences to develop a web-based platform for agro-environmental planning and modeling.

In this section, we describe one of the devised WebMAPS' products that is partially implemented and adheres to the layering described in section 2: computing historical NDVI profiles for a given region and period. NDVI is a widespread vegetation index that holds strong correlation with biomass, therefore NDVI profiles are very useful to monitor crops health and estimate crop yields.

The workflow to compute NDVI profiles can be resumed by the following steps: (i) acquisition phase: acquire satellite-based remotely sensed data, establish crop taxonomy for annotation, define geographical regions of interest; (ii) pre-processing phase: compute NDVI index, normalize data representation; (iii) storage phase: define datasets and use ap-
propriate media storage for each data type: file system, spatial database and relational database; (iv) knowledge management phase: extract regions of interest from raw datasets, reduce data using statistical aggregates, clean-up results; (v) publication phase: plot scatter graphs with historical profiles, serve core data and attributes through standard protocols and open interfaces.

In the acquisition phase, we must handle three different data types: satellite-based imagery represented as large raster files, vector-based geographical regions represented as sets of coordinates, and descriptions of crops and their attributes represented as structured textual records. All of them can be obtained from the Web.

There are many satellite imagery providers, our strategy was to identify those whose data products were closer to satisfy our needs. We found pre-computed NDVI images captured by MODIS sensors and provided by NASA. However, the distribution format was not adequate. Each image depicted a geographical region much larger than the ones we were interested in. Moreover, retrieving a single image meant browsing the web site to find the download link. Thus, assembling a local image database became a laborious, tedious and time-consuming task. To improve on that, we have developed Paparazzi.

Paparazzi is both a library and commandline tool to automate the acquisition of satellite imagery by means of screen scraping techniques. In the following example we are requesting all images (-r) from the first 4 months in 2007 (-b/-e) relative to the geographical region name Brazil4 (-s), each image has spatial resolution (-p) of 250 meters per pixel and represents NDVI measures (-m).

```
paparazzi.py -b 2007-01-01 -e 2007-05-01 -s Brazil4 -p 250m -m 2 -r
```

Our research about geoprocessing, databases, scientific workflows and semantic annotation converges in the WebMAPS project.

4. USING STANDARDS AND SEMANTICS

A standard is...

Standard as a standard form: protocol, representation, procedure

By semantics we mean...

Semantics materialised as ontology construction... need for common point for starting alignment (etc.) implies standard form...

Standards and semantics must be present at least in the frontiers of our data manipulation cycle, “wrapping” it (see Section 2). The communication interfaces for data acquisition and data publication are the two points were these solutions most useful.

OCG standards... in particular WPS...

Currently, the use of semantics is materialised in ontologies...

5. RELATED WORK

Standards x Ontologies Standards: pre-defined consensus Ontologies: pos-defined consensus Both: implicit semantics on the consensus

SOA x ROA...

ROA assumes/makes available CRUD operations – GET, PUT, POST, DELETE – (with pre-defined materialization)... CRUD operations in SOA have no predefined materialisation... CRUD have associated semantics by definition...

ROA — RESOURCES (our case DATA)... SOA — SERVICE RESOURCES (our case PROCESSING SOLUTIONS)...

Cache: HTTP... sem definições...

Geospatial data acquisition, processing and publication...

6. CONCLUDING REMARKS

This paper presented a view on how geospatial data and software is being used and how computer science research can provide means to empower the end users of such resources...

Future work involves extending the WebMAPS project to comply to more access standards, both from the communication and data representation points of view...

7. ACKNOWLEDGMENTS

This work was supported by FAPESP (grant 2004/14052-3) and partially financed by CNPq.