Bridging the gap between geospatial resource providers and model developers

Gilberto Z. Pastorello Jr
IC-UNICAMP
P.O. Box 6176, 13084-971
Campinas, SP, Brazil
gilberto@ic.unicamp.br

Rodrigo D. A. Senra
IC-UNICAMP
P.O. Box 6176, 13084-971
Campinas, SP, Brazil
rsenra@acm.org

Claudia B. Medeiros
IC-UNICAMP
P.O. Box 6176, 13084-971
Campinas, SP, Brazil
cmbm@ic.unicamp.br

ABSTRACT

This paper analyzes how interoperability and componentization efforts in the geospatial domain have an underestimated impact on the user perspective, directly 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.

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*

General Terms

Interoperability

Keywords

Geospatial data representation and processing, data and process availability, data publication

1. INTRODUCTION

In geographic information science, interoperability is a key issue, given the wide diversity of available geospatial data and scientific data processing tools. There are many research initiatives to meet this challenge, from data interchange standards and service-oriented architectures (SOA) to user interface design. This paper concentrates on two kinds of interoperability aspects: processes and data. Processes interoperability is related to 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, data consumers must be able to interpret one data set according to the same set of concepts. Both points of view are intimately related, since processes consume and produce data.

We show that efforts towards these directions have a desirable side effect: they are progressively shielding end users from having to deal with low level data management issues. This helps bridging the semantic and operational gap between data providers and scientists whose main interest is to design and test geospatial models, without having to concern themselves with low level implementation details.

The paper presents two main contributions towards helping solve process and data interoperability problems, namely: (i) a conceptual framework that structures those transformation steps into several layers, with clear cut interfaces and responsibilities, thereby helping systems designers; and, (ii) a real case study of this framework showing its advantages on reducing the gap between resource providers and model developers. The framework is being adopted within WebMAPS, a multidisciplinary project involving research in CS and agricultural and environmental sciences.

2. GEOSPATIAL DATA MANAGEMENT

The architecture of interoperable data management systems is often specified following a basic three-layer cycle: providers (data layer), transformers (service layer) and consumers (client layer). An example is the infrastructure provided by INSPIRE (`www.ec-gis.org/inspire/`), an initiative for the creation of a spatial infrastructure for Europe, with a distributed network of databases, linked by common standards and protocols to ensure compatibility and interoperability of data and services. Though useful to understand the functionalities provided, this kind of organization is insufficient for designers of geospatial systems to choose and compose process and data interoperability solutions. In order to meet this challenge, we propose an extended framework which induces a methodology for geospatial data management. This framework, shown in Figure 1, describes a data management cycle for GIS applications – from data acquisition (at the bottom) to data publication (at the top), to be consumed by applications that embed models. This cycle can be repeatedly pipelined: the data publishers of one cycle can become the data providers of the next cycle. The first five layers can be compared to a *Extract-Transform-Load* (ETL) process in data warehouse environments.

Our full data management cycle has seven layers, which alternate between representing either data or processes. Layers 2, 4, and 6 represent data, and boxes with gears (Layers 1, 3, 5, and 7) represent data manipulation operations. The flow is from bottom to top, with the operations being ap-

ACM GIS '08, November 5-7, 2008. Irvine, CA, USA
Copyright 2008 ACM 978-1-60558-323-5/08/11 ...$5.00.

plied to the data on their way up. We point out that not all stages of the cycle are mandatory – e.g., a given intermediate stage may not be needed, or applications may retrieve raw data directly from providers. Furthermore, an entire cycle may be under the control of a single organization (e.g., our case study of Section 3), or distributed on the Web.

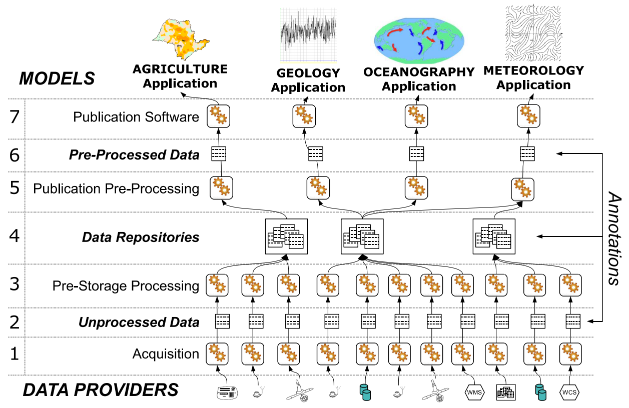


Figure 1: Geospatial data usage scenario

The bottom layer houses *Data Providers* of many kinds, including sets of files, databases, sensors and data services.

Layer 1 (*Acquisition*) hosts data acquisition software, which works as a mediator (or wrapper) to data providers. Layer 2 consists in *unprocessed data*, obtained directly from data providers in a variety of formats. Layer 3 (*Pre-Storage Processing*) represents the processing phase where data is transformed before its storage. Examples include signal processing functions for improving precision, and data cleaning for detecting variations or errors. Layer 4 (*Data Repositories*) 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 store and how to fill in the gaps left by several types of acquisition errors.

Layer 5 (*Publication Pre-Processing*) is responsible for transforming the data, filtering or augmenting it, in order to meet application requirements. Examples of such requirements include adjusting spatio-temporal resolution, access periodicity and specific presentation formats. The execution of operations in Layer 5 are guided by application needs while operations executed in Layer 3 are oriented towards storage requirements. Thus, unless the operation was executed in Layer 3 and the result is already available in the repositories, a request from an application is executed in Layer 5. Layer 6 contains the *Pre-Processed Data* sets, ready to be published to and consumed by models.

Layer 7 (*Publication Software*) represents the software that will make interfaces to operations and data access mechanisms available to applications. The publication software must allow applications to select pre-processing operations, among the ones available, to be applied on the data before transmission.

The upper layer (*Models*) 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 tune and interact with these models. Annotation mechanisms are orthogonal to all layers, using metadata standards or free annotations – see Section 4.

3. PUTTING THE FRAMEWORK TO USE

We illustrate the use of the framework from the previous section with the WebMAPS project, whose goal is to provide a platform based on Web Services to formulate, perform and evaluate policies and activities in agro-environmental planning. The project caters to two kinds of users – farmers, and domain experts, such as agronomers or earth scientists. WebMAPS data repositories include primary raw data (e.g., product classification from Brazilian official sources) and derived data (e.g., composite images). Geospatial data sets include satellite images, and coordinates of county boundaries.

We focus on one product from WebMAPS: NDVI (*Normalized Difference Vegetation Index*), which is a vegetation index correlated to biomass conditions of vegetation. An NDVI graph plots the average NDVI pixel value in a region through time from a temporal series of images – Figure 3 (bottom right). This can be used for crop monitoring and prediction, e.g., in the sugar cane culture, a curve with higher values may indicate a product with better quality.

NDVI graphs require two kinds of data – those acquired periodically (satellite images) and those that, once acquired, are only sporadically updated (e.g., county boundaries). This section describes the management cycle for these data within WebMAPS. We will not enter into details of acquisition periodicity nor procedures to refresh data, but such issues are embedded into constraints treated by our 7-layer framework. Figure 2 shows the main phases of the workflow that specifies the computation of the graph, following the layers of Figure 1.

Data Aquisition. There are many satellite imagery providers. For NDVI analysis, WebMAPS' agro-scientists have chosen to use pre-computed NDVI images provided by NASA from MODIS sensors (e.g., `http://rapidfire.sci.gsfc.nasa.gov/`). Here we faced typical problems of geospatial data acquisition. Each image depicts a geographical region much larger than the ones for which this first version of WebMAPS is being conceived (Brazil's southeast). Moreover, retrieving each image meant browsing the NASA web site to find the download link, which made assembling our image database a time-consuming task. To improve on that, we have developed Paparazzi, a tool to automate the retrieval of remote data sets from specific web sites by means of screen scraping techniques. Paparazzi is worth using whenever the number of files to be retrieved is large, and hyperlinks to target files are not concentrated in a single page, but scattered across several pages, as is the case with NASA MODIS images. If done manually, for each image, the user needs to visit three different web pages prior to starting a 50 Mb file download.

The other two kinds of data used are (i) vector-based coordinates, corresponding to the geographical regions of interest (using either manual region definition or importing shapefiles), and (ii) textual descriptions of crops and their attributes (also using screen scraping techniques).

Unprocessed Data. Satellite images retrieved using Paparazzi and shapefiles are encapsulated in temporary files, for subsequent quality checking. The rest of the data used goes directly to Layer 3 (Pre-Storage Processing). Our multilayer framework allows determining which stages should be followed for each kind of data.

Pre-Storage Processing. There are three main concerns

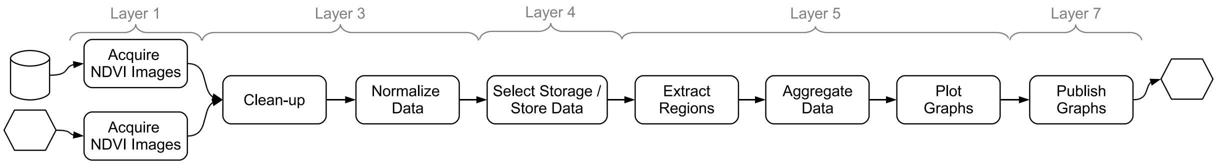


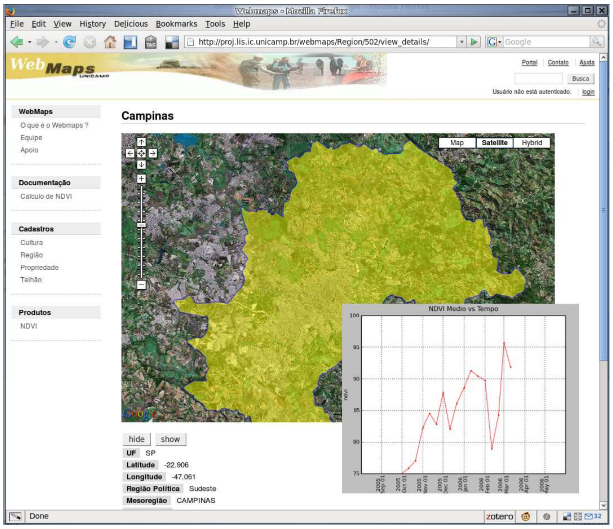
Figure 2: Computation and publication of NDVI series for a region

here: corruption detection, data normalization and assembly of the data sets. Corruption detection is mandatory and is made explicit in our framework, since data providers and data transmission are never 100% reliable. Data normalization is a recommended step to make data processing easier and more efficient. We automatically convert all files to a single and uniform representation format, and all measurement units to the same system. Data set assembly is the last pre-storage processing step, consisting in using coherent spatio-temporal units – in our example, creating a composite NDVI image from a mosaic of acquired NDVI images.

Data Repositories. Once the data are pre-processed, they are ready for storage. We use two types of storage: a relational database and the filesystem. Crop descriptions, geometries, textual properties, and data set descriptions are stored in PostgreSQL/PostGIS. Raster images in GeoTIFF format are stored filesystem partitions. Our preliminary experiments with these data appear in [6].

Publication Pre-processing. This phase concerns transforming information, and ultimately preparing it for user consumption. In our example, this means (i) computing the average NDVI pixel value for the region defined by the user; and, (ii) iterating (i) for the input time span. These steps are performed automatically without user intervention.

Data Publication. This is the last phase in our case study, when NDVI graphs constructed in the previous phase are published as images. WebMAPS innovates allowing data produced in any of the framework phases to be directly accessible in many representations. This is facilitated by isolating the responsibilities of each framework layer.



UF SP
Latitude -22.906
Longitude -47.061
Região Política Sudeste
Mesoregião CAMPINAS

Figure 3: WebMAPS embedding map generated by Google Maps

WebMAPS can act as a data provider and mediator by re-distributing geometries acquired from an authoritative source (Brazilian Geographic Institute), e.g., transforming them into a suitable format to feed Google Maps. In Figure 3, we depict WebMAPS acting as a client of Google's map rendering service. The map rendered by Google Maps (using geometry data from WebMAPS) is mashed-up with results from a user query, composing the web page shown in Figure 3. The query results comprehend textual metadata and a NDVI graph for the given region and time frame.

This interaction pattern between WebMAPS and Google Maps is a combination of resource-oriented (from WebMAPS) and service-oriented (from Google Maps) paradigms – see Section 4. In this example, using KML (`www.opengeospatial.org/standards/kml/`) and WKT (`www.opengeospatial.org/standards/sfa`) enabled us to rapidly build a prototype for cartographic visualization, including satellite image overlays provided by Google Maps. End users are rapidly able to visually assess the quality of the data, and test the outcomes of different analyses. Hence, standards offer much more than interoperability. Their use has sped up the validation of user requirements in terms of interaction needs. More importantly, it has leveraged model development, so that users can start testing their ideas much sooner, while we work on other system issues. We are also experimenting with other kinds of Web service-based solutions (see [3] for our use of GeoServer to publish GML data for biodiversity systems).

4. INTEROPERABILITY APPROACHES

From the data interoperability perspective, standards deal with representation and formatting issues, e.g., OGC's GML. From the process interoperability perspective, standards are used in the specification of protocols, interfaces and descriptions of processes. Examples include OpenDap *Open-source Project for a Network Data Access Protocol* and OGC standards, such as WFS for vector data access and the more recent *Web Processing Service* (WPS) for processes. Although WPS does not describe the specific behavior of an operation, it provides general description mechanisms, such as *Profiles* and *ProcessDescriptions*. This, however, still leaves room for semantic mismatches.

Standards 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 publication are the two points where these solutions are most useful: WebMAPS can be seen as a client application and a data provider to client applications. As a server, WebMAPS strives to adhere to standards, to enable interoperation with other systems. As a client, taking advantage of standard interfaces is important, however, being able to handle involuntary, non-standardized, access mechanisms might be equally important. As part of those efforts, its development is adopting Web services and SOAP protocols, OpenDAP, Microformats and KML.

In our framework, OpenDAP is used as a means to retrieve and publish data, in layers 1 and 7. For instance, images are acquired and served by WebMAPS using OpenDAP. In the first case, it is at the receiving end (Layers 1 and 2), while in the second case it is at the top of the cycle. As exemplified by [7], this allows scientists to exchange and visualize results of complex models. Microformats (e.g., *Geo* for geographical coordinates) is a web-based data formatting approach to re-use existing content as metadata, through standard annotations conveyed by XHTML (or HTML) classes and attributes. Their use has direct impact in the representation of data in Layer 6, after being generated in Layer 5 along with other transformation processes. KML is an XML-based language schema for expressing geographic annotation and visualization for 2D and 3D Earth browsers (e.g., Google Earth). Our geometry files are represented in KML, in which case we are acting as data providers.

In the services interoperability, there are two paradigms competing in the Web: Service-oriented architectures (SOA) and Resource-oriented architectures (ROA). SOA is a direct evolution of concepts born from distributed computing and modular programming practices. It is an architecture where functionality is grouped around processes and packaged as interoperable RPC-style services, loosely coupled with operating systems or programming languages. On the other hand, ROA is intimately related to the Web. It rescues the principle of Representational State Transfer (REST). REST outlines how resources are defined, addressed and accessed through simple interfaces, where domain-specific data is transmitted over HTTP without any additional messaging layer or session tracking mechanisms. ROA is more scalable than SOA, and easier to implement due to its uniform interface and adherence to Web model and standards. SOA and ROA are complementary paradigms, together they maximize interoperability – as is the case with WebMAPS: using SOA when accessing Google, and ROA when serving it.

5. RELATED WORK

There are many studies concerning use of standards, usually restricted to just one of our layers. Aim4GDI [1], uses OGC standards for accessing distributed data sources and creating composite results. The work presented in [4] considers the use of standards (based on the ISO19100 series) for both data and process interoperability, for distributed sources. However, limiting the standards considered for interoperability into a single domain hampers the construction of multi-disciplinary models and applications, preventing their evolution. This is remarked by [5], which discusses the evolution of the GML standard and the importance of integrating it with standards from other application areas.

Interoperability through services is also common. The work of [2], for instance, describes initiatives towards combining communication and access standards, e.g., providing common grounds for WFS and WCS to work side by side with OpenDAP to access oceanographic data. Like us, their effort shows that combining different standards into systems design is a way of leveraging interoperability.

Our main concern, however, is to provide adequate support to flexible model development. From this point of view, the motivation of GeoModeler [7] is the closest to ours, making geospatial resources more accessible to model developers. GeoModeler is a software framework that combines software components from a GIS with modeling and simulation software, ultimately allowing various forms of analysis and visualization of oceanographic data. Their approach, however, deals with construction of centralized systems and software components interoperability in such systems. It does not consider, for instance, data acquisition or publication.

6. CONCLUDING REMARKS

This paper presented a framework that analyzes the management of geospatial data from a life cycle perspective. This framework is being validated in the design and development of the WebMAPS project.

By isolating each layer in the cycle, with clear interfaces and tasks, the framework induces a methodology to design and develop interoperable geographic applications. Whereas related research concentrates on providing standards or services for one given data transformation stage, we show how these efforts can be seamlessly interconnected. This allows users to shift their focus from the technology being used to the models being constructed.

Future work involves extending the WebMAPS project to comply to more access standards, from both the communication and data representation points of view. Another research issue involves the use of ontology-based techniques to speed up query processing and annotate data and processes – see our work in [3].

7. ACKNOWLEDGMENTS

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

8. REFERENCES

[1] T. Aditya and M.-J. Kraak. Aim4GDI: Facilitating the Synthesis of GDI Resources through Mapping and Superimpositions of Metadata Summaries. *Geoinformatica*, 11(4):459–478, 2007.

[2] P. Cornillon, J. Caron, T. Burk, and D. Holloway. Data access interoperability within IOOS. In *Proc. of MTS/IEEE OCEANS*, pages 1790–1792, 2005.

[3] J. Daltio, C. B. Medeiros, L. Gomes Jr, and T. M. Lewinsohn. A framework to process complex biodiversity queries. In *Proc. of the ACM Symposium on Applied Computing*, pages 2293–2297, 2008.

[4] S.-G. Jang and T. J. Kim. Modeling an Interoperable Multimodal Travel Guide System using the ISO 19100 Series of International Standards. In *Proc. of the 14th Annual ACM Int. Symposium on Advances in Geographic Information Systems*, pages 115–122, 2006.

[5] C.-T. Lu, R. F. Santos Jr, L. N. Sripada, and Y. Kou. Advances in GML for Geospatial Applications. *Geoinformatica*, 11(1):131–157, 2007.

[6] G. Z. Pastorello Jr, C. B. Medeiros, and A. Santanchè. Accessing and Processing Sensing Data. In *Proc. 11th IEEE Int. Conf. on Computational Science and Engineering*, pages 353–360, 2008.

[7] T. C. Vance, N. Merati, S. M. Mesick, C. W. Moore, and D. J. Wright. GeoModeler: tightly linking spatially-explicit models and data with a GIS for analysis and geovisualization. In *Proc. of the 15th Annual ACM Int. Symposium on Advances in Geographic Information Systems*, 2007.