Data Driven Research at LIS: the Laboratory of Information Systems at UNICAMP

Claudia Bauzer Medeiros, André Santanchè, Edmundo R. M. Madeira, Eliane Martins, Geovane C. Magalhães, Maria Cecilia C. Baranauskas, Neucimar J. Leite, Ricardo da S. Torres

Institute of Computing, University of Campinas, UNICAMP, Brazil
{cmbm, santanche,edmundo,eliane,geovane,cecilia,neucimar,rtorres}@ic.unicamp.br

Abstract. This article presents an overview of the research conducted at the Laboratory of Information Systems (LIS) at the Institute of Computing, UNICAMP. Its creation, in 1994, was motivated by the need to support data-driven research within multidisciplinary projects involving computer scientists and scientists from other fields. Throughout the years, it has housed projects in many domains - in agriculture, biodiversity, medicine, health, bioinformatics, urban planning, telecommunications, and sports - with scientific results in these fields and in Computer Science, with emphasis in data management, integrating research on databases, image processing, human-computer interfaces, software engineering and computer networks. The research produced 14 PhD theses, 70 MSc dissertations, 40+ journal papers and 200+ conference papers, having been assisted by over 80 undergraduate student scholarships. Several of these results were obtained through cooperation with many Brazilian universities and research centers, as well as groups in Canada, USA, France, Germany, the Netherlands and Portugal. The authors of this article are faculty at the Institute whose students developed their MSc or PhD research in the lab. For additional details, online systems, papers and reports, see `http://www.lis.ic.unicamp.br and http://www.lis.ic.unicamp.br/publications`

Categories and Subject Descriptors: H. Information Systems [**H.2. Database Management**]: H.2.8. Database Applications

Keywords: Data-intensive computing; GIS; Multidisciplinary research; Scientific data management

1. OVERVIEW OF LIS - WHAT IS IN A NAME?

There are many meanings and a long explanation behind the name LIS - Laboratory of Information Systems. It was chosen to convey (a) the kind of research conducted (data- and information-driven, motivated by real world applications) and (b) the nature of some of the data managed (LIS being the Portuguese word for "lily", a flower, and thus biodiversity and environmental applications). The name conveys the provenance and provides the appropriate introduction to the laboratory – when it was created, for which purposes, what kind of research has been conducted since, and by whom.

First, there are very many definitions for the term "Information Systems", with distinct views of what such systems should comprise – some researchers concentrate on data and software issues, others add hardware considerations, and an even broader view integrates people, managerial and organizational issues. LIS' "Information Systems" belongs to the first kind of issues, for research that is data-driven, thus requiring state-of-the-art work in data and information management. Its focus is to design and construct models, techniques and software tools to help end users *store, manage, retrieve, analyze, visualize, share, reuse and interact with* their data, through "information systems".

The work described in this article has been funded by Brazilian government research financing agencies FAPESP, CNPq and CAPES, foreign grant agencies NSF, GMD, CNRS, COFECUB, as well as Microsoft Research, HP Research Labs, EMBRAPA, COOXUPE, FAEPEX-UNICAMP, CPqD,and FUNTTEL. Many of the students that developed their theses in the group were faculty from Brazilian or foreign universities, and were supported by them.

The laboratory was born out of a need to support collaborative research among computer scientists of different domains, being initially motivated by challenges posed by handling *geo-referenced data* (i.e., data that are associated with some geographic location on the earth). Even though there had been research on databases conducted at UNICAMP's Computer Science Department[1] since 1985, LIS was created by Claudia Bauzer Medeiros in 1994, as part of a large project funded by CNPq, involving 10 universities and research centers, coordinated at UNICAMP, to design tools and techniques for geo-referenced data, for environmental and telecommunication applications. Large amounts of data had to be handled for a variety of spatial scales and constraints, (e.g., from the design of entire telecommunication networks, down to the location of individual equipment; or from a micro-region to entire biomes). Thus, in order to devise solutions to the data management problems, there was a need to involve researchers from other CS domains. As such, besides theoretical work, there has always been emphasis on prototypes (to meet end-user needs), and consequently research on computer networks, software engineering and human-machine interfaces. Since a wide range of geo-referenced applications require image processing and interpretation, one of the emphases has been the work on images – initially from satellites and radar data sources, and afterwards other kinds of images (photos, fingerprints, images of living beings).

Besides being data-driven, research in LIS has always been motivated by applications – with cooperation with researchers in the life sciences (notably bioinformatics, biodiversity, environmental management, health and medicine), agricultural sciences (mostly crop planning, monitoring and management) and engineering (especially infrastructure networks). The first set of domains further helped define the name of the laboratory – LIS is the Portuguese word for a flower (lily), and thus there was an underlying connotation with living beings and biodiversity issues. The first PhD thesis from LIS [Pires 1997] (the second PhD granted in the Institute of Computing) proposed a computational framework, called `uape`, to design environmental applications. This is a tupi-guarani[2] word meaning *water lily*, and the associated explanation for the name, in the thesis, is that water lilies have a nice user-friendly interface, underneath which flourishes a complex ecossystem of interacting software and modeling tools to handle heterogeneous data.

2. MULTIDISCIPLINARITY IN DATA-DRIVEN PROBLEMS: LIS' DRIVING FORCE

The projects conducted within LIS have always been motivated by problems posed by the management of data, most of the times involving some application domain. As such, research could not be restricted to database experts, and demanded cooperation with computer scientists in other fields, as well as with domain experts.

2.1 Overview - integrating multiple research lines

Figure 1 schematically shows the relationships among some of LIS' main research lines, and how the authors of this article contributed to these results. Data and information management are at the center of the figure, surrounded by some of the main collaborating research lines. Core database research was conducted to solve issues associated with multiple kinds of digital content produced by devices, experiments, and people. Since several of the data sources were distributed, and distinct kinds of processing were required, there was a need to involve computer networks research in load balancing, fault tolerance and Web services (oval at the bottom). Given that a large portion of the data analyzed in LIS are images, and videos, it was necessary to involve researchers in image processing. The interaction with scientists from other domains showed that their work (and requirements) could be supported through the use of workflows, thus prompting research in this field. Most results have

[1] The CS department gave origin to the Institute of Computing.
[2] Belonging to the Tupi group of native South American languages.

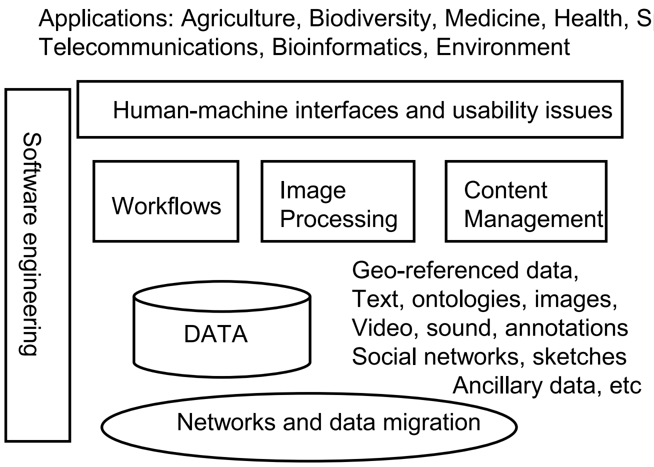

Fig. 1. Overview of interdisciplinarity within LIS, to solve data- and information-driven challenges.

been materialized (and validated by end users) through prototypes, for which research in interfaces and in software engineering was also needed.

The authors of this article are faculty at the Institute whose students developed their MSc or PhD research in LIS. Though all are concerned with data management issues, their additional contribution to the research conducted at LIS, within the scenario depicted in the figure, is as follows. M. C. Baranauskas contributed to interface issues; R. Torres and N. Leite, in image processing; A. Santanchè, in content management; E. Madeira in networks; E. Martins in software engineering; G. Magalhães and C. Medeiros at the database level, the latter also in workflow bases.

Data heterogeneity has been a constant problem in most of the applications at LIS, as shown by the partial enumeration of data types in Figure 1. There has always been a great variety of data types to be processed, indexed and combined in queries, and, for each kind of data, a wide range of spatial and temporal scales and units, bringing up challenges in capture, storage, integration, retrieval, analysis, visualization. Major applications domains have been medicine, health, bioinformatics, agriculture, biodiversity, urban planning, telecommunications, and sports. Examples of data sources handled include (not exclusively):

—Satellite images – time series of satellite images, covering decades in time, in which each series concerns a given geographic region, and a set of satellite sensors – for applications in agriculture, biodiversity and environmental management;

—Real-time streams and historical data from sensor networks and weather stations – again covering decades, data on rainfall, humidity, light, wind, and others, collected at distinct time intervals by a large variety of sensors and sensor types, with varying callibration and sensitivity characteristics - for applications in agriculture;

—Photos of living beings – images of animals, either taken in their natural habitats or in a laboratory (e.g., museum environment), using distinct devices. In some cases, a set of images is necessary to help animal identification (e.g., multiple distinct views) – for a variety of biodiversity applications;

—Data on urban environments and urban networks, for telecommunication applications;

—Medical images – mostly from microscopes, for medical studies.

—Videos – of humans in sports activities, and advertisements, for health and sports studies.

—Recordings of animal sounds – mostly captured in nature, over 10 years, using a wide variety of devices, for biodiversity studies.

—Textual data – ontologies (both provided by users and constructed in the lab), annotations, manuals, scientific texts and ancillary data in general.

—Other kinds of data to be combined with the previous sources – descriptive data, maps (in vectorial format), statistical databases, workflow repositories, fingerprint images and others.

2.2 Data-driven research and database issues

This section discusses some of the research results involving more traditional database topics. Briefly, work developed within core database research concerned a wide range of problems – performance and benchmarks (e.g., [Souza 1995; Gatto 2000]), indexing and data structures (e.g., [Costenaro 1997; Carneiro 1998]), query processing (e.g., [Rocha 2003; Daltio et al. 2008; Vilar et al. 2009; Fedel 2011], ontologies and digital content management (e.g., [da S. Torres et al. 2006; Santanchè and Medeiros 2007b; Daltio and Medeiros 2008; Santanchè and Medeiros 2007a]), access control (e.g., [Sasaoka and Medeiros 2006]), models and metadata (e.g., [Faria et al. 1998; Fagundes and Medeiros 2000; Pastorello Jr et al. 2005]), work on digital libraries (e.g., [Goncalves and Medeiros 1998; Gonçalves and Medeiros 1999; Oliveira et al. 1999; Torres 2004; da S. Torres et al. 2006; Pedronette and da S. Torres 2010], text processing and annotations (e.g., [Macário and Medeiros 2009; Pastorello Jr et al. 2010]), sensor data management and data mining (e.g., [Pastorello et al. 2008; Koga et al. 2011; Mariotte et al. 2011]).

2.2.1 *Handling heterogeneous data sources.* One of the earliest research lines at LIS on geo-referenced data was concerned with data structures to store and access spatial data for network and urban infrastructure application databases, with performance studies based on models and benchmarks. Analytical and experimental studies on factors that influence the performance produced recommendations on how to tune spatial database structures based on data distribution present in real world databases. With the advent of object oriented databases the use of these results and benchmarks helped to compare object oriented with relational databases wrt performance for the same databases [Vasconcelos 1996]. An associated line of research, to be applied in planning and pre-design engineering tools, was the use of spatial predicates and data analysis methods to allow data mining on spatial databases with conventional mining tools available for relational databases [Silva 2003; 2004]. Moreover, Web services for location based services using OGC standard specifications were prototyped to analyze their feasibility and performance in mobile access over smart phones. The main application domain in all performance studies concerned design and monitoring of telecommunication networks. The underlying databases stored information on multiple urban scales, concerning all aspects necessary to design and maintain such networks, from the consumer to the centrals: circuits and components, cables and transformers, streets, city sections and entire cities. Handling data entities at these multiple levels was an additional challenge, requiring studies in active databases to maintain spatial constraints (e.g., [Medeiros and Cilia 1995]).

Agriculture was another early major application domain in LIS' research, also involving geo-referenced data. Figure 2 shows an example of a screen copy of a Web tool developed within a project motivated by the needs of crop planning and monitoring. It further illustrates the cooperation among some of the authors. It shows an NDVI (Normalized Difference Vegetation Index) curve, for a given region, obtained through processing of the content of a time series of 100 satellite images stored in an image repository.

For researchers in agricultural sciences, the curve indicates the behavior of a given crop (here, sugarcane) in a region, for a time period. This can serve as the basis for several decision actions: monitor the behavior of that crop for that region, comparing it with the behavior of the same kind of crop for other periods and/or regions; forecast crop yield; detect problems and plan activities to improve or correct the situation (e.g., using fungicides).

From a data-driven research viewpoint, on the other hand, producing this screen required results

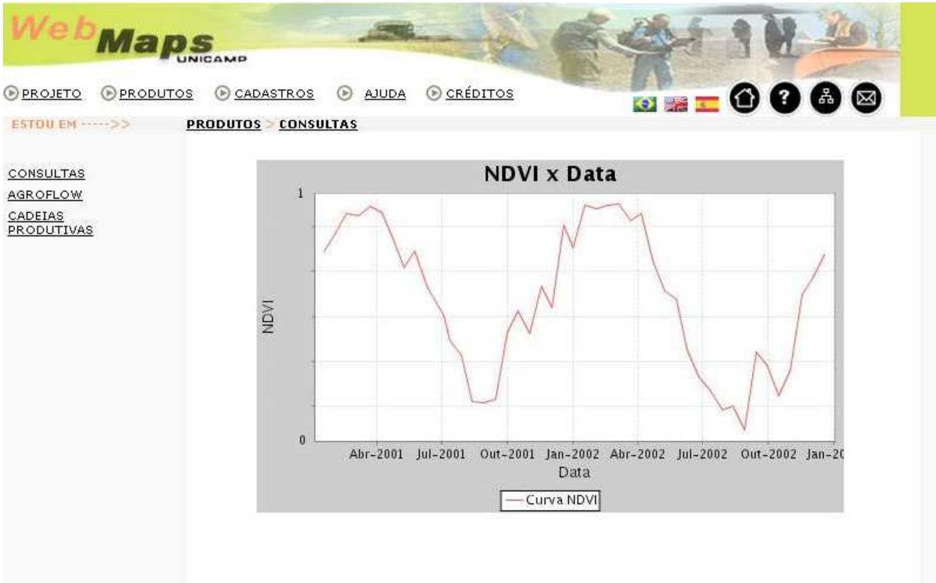

Fig. 2. Screen copy - WebMaps project in agriculture.

in issues such as: creating annotations for heterogeneous data sources (e.g., [Macário and Medeiros 2009], integration of image processing algorithms and geographic databases, for large data volumes [dos Santos et al. 2008; dos Santos et al. 2011]; design and implementation of new methods to test Web software, with emphasis in reliability and fault tolerance [Escalona et al. 2008]; design of interfaces for geographic Web software, using research on semiotics (e.g., [Schimiguel et al. 2005b; Schimiguel et al. 2006]); adoption of Semantic Web models and standards to improve data interoperability (e.g., [Fileto et al. 2003; Pastorello Jr et al. 2010]). The same data sources, combined with real time data from sensor networks are behind research in time series mining [Mariotte et al. 2011; Koga et al. 2011].

Agriculture poses a wide spectrum of challenges for data-driven research. One particular issue is the sensitiveness to phenomena that vary in time and space – and thus the need for work in spatio-temporal databases. Time series management is an example in which LIS has dedicated much effort – from storage and indexing, to mining algorithms. Besides investigation on the more "traditional" point-based time series (e.g., data streams captured from sensors or weather stations on temperature, rainfall and such [Koga et al. 2011]), there has also been intensive work on image time series (e.g., to monitor a crop's behavior along time through its spectral response in images). This required, among others, designing new image processing algorithms (to identify crop patches in a given image – see section 2.3). First, each image has to be pre-processed to extract the desired information, and identify the pixels of interest, which correspond to a given region [dos Santos et al. 2008]. Next, this region is "followed" through time, processing the associated pixels for each image, resulting in a time series with points $p_i = < V, t_i >$. Each point p_i is a descriptor of that region, for an image taken at time t_i [3] and where V, the value at t_i, can vary from a single real value to a vector of values, depending on the descriptor chosen. Our new mining algorithms support processing multiple heterogeneous series simultaneously [Mariotte et al. 2011] – the emphasis is on the co-evolution of many series, and not on a single series. TIDES – the new time series descriptor developed – concentrates on the oscillation patterns of the series, instead of on the variation of values, which is the usual way of processing time series. TIDES is able to identify recurring patterns of co-evolving phenomena (e.g., rainfall and temperature), thereby helping monitor conditions that affect a crop.

[3]This is a simplification of the problem, since a "satellite image" corresponds in fact to a composition of images and, in agriculture, this composition may vary over time.

2.2.2 *Content management.* The work on images and image databases was extended to encompass other kinds of digital content, including software. This was directed towards the needs of design, development and reuse of digital content, with emphasis on content authoring. The starting point of the research in content management was the Digital Content Component (DCC) [Santanchè and Medeiros 2007b], comprising a component model and infrastructure that implements it. A DCC has a self-describable capsule model, whose internal structure is organized as a complex object and whose external structure is built as a software component. This internal object has a hierarchical containment structure, able to store and relate any kind of digital content, including executable software. Externally any DCC acts as a software component, declaring required and provided interfaces. It can encapsulate data, but also software, and thus be used to design and contruct data-driven applications in which a DCC that encapsulates data can be composed with one that encapsulates the software that will access that data. The DCC framework and infrastructure are being used to attack problems in workflow design, management of multimedia data and interoperability issues – see section 2.3.

2.2.3 *Scientific workflows and workflow bases.* Applications in agriculture, environmental management and bioinformatics provided real case studies for work on scientific workflows. Workflows were initially used to document the interaction of users with decision support for geographic systems, giving origin to work on workflows and geographic information systems [Weske et al. 1998] and the WOODSS system (WOrkflOw-based Decision Support System) [Seffino et al. 1999; Medeiros et al. 2005]. In this context, a workflow is considered to be a description of an experiment or a scientific procedure. The main goal of this research line has been to support scientific work (with re-execution of experiments, or construction of new ones), documentation (whereby workflows indicate which processes and data are used in a specific situation), reuse and provenance. This resulted in the development of a model and data structures to store, index and annotate workflows, an on-line framework for workflow construction and annotation [Medeiros et al. 2005] and an environment to help the design of workflows using planning algorithms from artificial intelligence (e.g., [Digiampietri 2007]). WOODSS was one of the first proposals in the literature to introduce the notion of a database (or, more appropriately, a "workflow-base") to store workflows, to be retrieved and compared with each other, identifying common procedures followed to solve a given set of problems [Kaster et al. 2005]. WOODSS' model and implemented framework support workflow design from the notion of an *abstract workflow* (where activities and flow are stated in abstract terms – e.g., "execute an overlay operation") to a *concrete workflow* (with all workflow elements instantiated, including the code to "execute the overlay operation"). One executable workflow corresponds to a set of stored workflows, representing a top-down process of workflow design, from an abstract to a concrete representation. This also required, among others, investigating alternative workflow storage models, retrieval and composition mechanisms and workflow versioning.

While scientific workflows are primarily used to drive execution of experiments, in LIS there was also work on using them for traceability [Kondo et al. 2007] and management of supply chains in agriculture [Baccarin 2009]. Research on workflow mechanisms and workflow storage systems was combined with work on web services and choreographies, to dynamically construct and execute workflows, as well as on mechanisms to dynamically negotiate multiparty contracts using distributed workflows, for agriculture applications [Baccarin et al. 2011]. Here, rather than drive the execution of experiments, workflows were used to drive the negotiation of parameters in constructing a contract in supply chains. Again using a real life case study, when a farm cooperative negotiates delivery of milk to a dairy processing chain, the cooperative and the dairy chain are global entities that hide multiple partners, each of which may intervene in the negotiation (e.g., negotiating the amount of milk to be delivered by each farm). This work introduced two novelties, thanks to the choice of workflows to support supply chains: the possibility of dynamically adjusting contracts, when external conditions hamper contract enforcement (e.g., in real life, droughts may keep cows from producing milk); and supporting multiple kinds of protocols for negotiation of contract clauses (e.g., dutch auction, direct negotiation). Since most research in e-contracts concentrates in supporting one single protocol, this enabled a more flexible environment to monitor supply chains.

2.3 Data-driven research in collaboration with other CS domains

2.3.1 *Research on image processing to enable queries.* A major emphasis in the laboratory concerns aspects associated with the content of the data being processed, thus requiring algorithms for extracting information from the content and encapsulating it for subsequent retrieval and composition. Related research involved work on implementation of algorithms and tools aiming to support two kinds of content management issues: (a) search for images based on image content (visual properties such as color, texture, and shape), in issues such as mathematical morphology, image segmentation, and classification; and (b) design and development of the DCC model and infrastructure to support semantic access and management for all kinds of digital content, already addressed in 2.2.

The contributions on mathematical morphology are represented by a set of new transformations used to simplify the image content and improve the extraction of regions of interest. Based also on scale-space theory, these transformations consider the problem of monotonically reducing the number of the image extrema (regional maxima and minima) in a more supervised way [Leite and Teixeira 2000]. This feature is very important in applications involving, for example, image segmentation, representation and description, and extraction of descriptors.

Content-based image retrieval relies on the use of *descriptors.* A descriptor [da S. Torres and Falcão 2006] extracts feature vectors that represent image visual properties and defines a distance function that is used to determine how similar two image are, given the distance of their feature vectors. Given a query pattern (usually a query image), descriptors can be used to rank collections of images according to their similarity to the query pattern.

New image descriptors aiming to characterize different visual properties were proposed and evaluated (e.g., color [Penatti and Torres 2008], texture [Montoya-Zegarra et al. 2007; Penatti and Torres 2008] in either rotation-invariant and scale-invariant manners [Zegarra et al. 2008], shape [Andaló et al. 2010], and spatial relationships). Additional work on content-based image retrieval was defined to cope with the problem of looking for a given remote sensing image in a large remote sensing database. Different models for this kind of images were developed, based on the use of multiple content representation models [Cura et al. 2000] and relevance feedback [dos Santos et al. 2011].

Further, a new directional operator was developed to estimate the orientation field of 2D components. This operator, together with some morphological transformations, was used to enhance the quality of fingerprint images and reconnect the broken ridges commonly found in this type of data [de A. Oliveira and Leite 2008], being used in security applications for fingerprint databases.

The work on descriptors was combined with database retrieval mechanisms, for content based systems. Here, artificial intelligence algorithms were combined with image processing methods to improve image indexing and recovery. In particular, machine learning approaches were adopted, such as genetic programming and association rules [Faria 2010; da S. Torres et al. 2009] for combining distance scores defined by different descriptors. Another venue consisted in applying genetic programming to implement relevance feedback methods [Ferreira 2007; dos Santos et al. 2010], where a user's evaluations of image relevance are adopted to improve query effectiveness. Image indexing and clustering approaches were also investigated to speed up the search time.

It is also worth mentioning the research on medical imaging that considers both mathematical morphology and texture information. This work deals with the management of medical images, in cancer research, and required segmentation and classification of chromatin texture of cell nuclei. Motivated by cancer studies, this resulted in a new diagnostic software, based on the notion of granulometric residues, to be used in cytology (e.g., [Ferreira et al. 2006]).

2.3.2 *Video tracking.* Another line of data-intensive research developed at LIS concerns the definition of algorithms for the tracking of objects in videos, allowing an automatic determination of their 2D and 3D coordinates, as well as the quantification of the kinematical variables of the human motion. This work concerned two applications: the clinical analysis of human motion and athletes performance evaluation. In the first case, algorithms were defined for the tracking of markers fixed to the human body in the regions of interest (e.g., in the articulations). In the second case, the problem was tracking soccer players during a match whose video images can include changes of illumination and players' occlusion [Rivero et al. 2003; Leite et al. 1999; Barros et al. 2010; Rivero et al. 2006].

2.3.3 *User interfaces, usability and software engineering.* Since a large portion of the research conducted in LIS is motivated by real applications, design and development activities required research in man-machine interfaces – from their flexible construction (e.g., [Oliveira et al. 1997]) to usability issues. Usability is, for instance, of particular interest in agricultural information systems, some of which target farmers as end-users. Moreover, our work has been always developed together with research in software engineering, in cooperation with researchers of the distributed systems lab.

The work on interfaces was initially concerned with the interaction with digital maps, and subsequently with the design and development of geographic applications on the Web, having cartographic representations as one of the main concerns [Carvalho et al. 2010]. Understood as an artifact through which geographical realities are communicated, the map reveals its semiotic nature. From its origin, having table clothes (which were called mappas) to draw on routes, paths and places, to our times having the computer as a medium, map construction can be understood as a process of communicating our geographical reality. The quality of the user interface has great impact in the process of creating applications upon Geographical Information Systems, as well as in the final product - the geographic representations - to be interpreted by end users.

Computer-based technologies, besides enhancing the possibilities of expressing our geographical reality, can reach and benefit a broader range of people, within a myriad of applications. Although functionally powerful, such systems should not presuppose that users have knowledge of specific aspects of the system technology. Potential users are not necessarily used to the inherent complexity of those systems, and should not be diverted from concentrating in making sense of geographical data.

Considering this evolutionary understanding of expression and interpretation of geographical information by experts and non-experts alike, the work developed in LIS has investigated formalisms and methods inspired in Semiotics and Participatory Design for the design and evaluation of systems in the domain of geographical information. The studies have involved from the evaluation of the expressive power of signs present in GIS and Web GIS to frameworks for design and evaluation of such systems (e.g., [Prado et al. 2000; Schimiguel et al. 2005; Schimiguel et al. 2005b; Schimiguel et al. 2006]). Additional work along this line was concerned with the visualization of queries on image data, with the proposal of new visual structures used to display query results (e.g., [da S. Torres et al. 2003]).

Software engineering work conducted within the laboratory concerned the design and development of test methods for the algorithms on data retrieval, and development and performance testing processes for Web GIS applications. The work concentrated on NDT (Navigational Development Techniques) combined with interface research on Organizational Semiotics. Testing conditions considered real time user interaction with great volumes of data and varying the load of simultaneous users, supplying a set of well detailed test activities. These activities have been integrated into a robust process, whose starting point was the identification of the critical use cases to be tested. This work was concerned with agricultural applications [Escalona et al. 2008]. Ongoing work involves fault tolerance studies concerning Web data access for biodiversity applications.

2.3.4 *Computer networks and load balancing to support data migration and distributed management.* Work on computer networks comprised two directions – the management of workflows [Nakai et al. 2008], including dynamic composition of tasks and their orchestration, and optimization of data traffic on the Web, to process distributed queries. Work on workflow systems was described previously. Load balancing and optimization of traffic management is part of ongoing research, being described in the next section.

3. ONGOING WORK AND DATA-DRIVEN MULTIDISCIPLINARY RESEARCH

Ongoing research and development directions roughly concern four main issues: (a) data-driven research in eScience (including work in workflows, annotation mechanisms, sensor data management, multimodal query processing mechanisms and the associated information systems); (b) content management and authoring; (c) load balancing and network issues as the basis for a Web-based data management infrastructure; and (d) research in Web Science, since LIS is one of the laboratories associated with the INCT in Web Science. These directions are motivated by three kinds of user requirements – agricultural planning, species interactions in biodiversity, and social networks. The first three directions are commented next. Work on Web Science permeates the other directions, and cannot be presented in isolation. It is centered on work in social networks, query processing on the Web, and the Semantic Web.

3.1 eScience and data-driven research

A typical data capture and processing cycle in eScience involves: (1) selection of relevant data sources for a given goal (e.g., sensor data, images, sounds); (2) data pre-processing (e.g., algorithms to extract descriptors from images or sounds); (3) data insertion and indexing; (4) specification and development of test procedures; (5) design and development of mining, retrieval and query functions; (6) analysis of the data retrieved (using user-defined models, algorithms and simulations), and (7) visualization of the results. Scientists can interrupt and interact with any of these stages – e.g., tuning parameters, providing relevance feedback. The previous sections described how we are attacking some of these problems, and we will continue along the same lines. Recent work includes annotation management (e.g., [Pastorello Jr et al. 2010]), support for the dynamic organization of digital archives, investigation of spatio-temporal interactions and work on animal sounds [Cugler et al. 2011], to provide support to biodiversity and habitat studies, and species interaction.

Several of the challenges being faced concern data heterogeneity. For instance, one specific research line involves the integration of time series of satellite images (Terabytes per year) with ground sensor time series. Each kind of data varies along distinct spatio-temporal scales. Challenges start from the capture (which data to use, how to collect them), pre-processing and storage (involving data curation and quality assessment), filtering, fusion, combination with textual data (e.g., annotations), analysis (with data- and domain-specific algorithms)

Data heterogeneity presents obstacles inherent to the distinct spatial and temporal scales, user requirements, and devices (with very different precision and quality parameters). Spatial and temporal gaps in data capture must also be taken into consideration, and have to be filled computationally by simulations or interpolation methods, or even by mining other data sources. User requirements also direct the kinds of data management issues to be considered. For instance, in environmental and biodiversity applications, a broader scenario is a "landscape", that contains multiple habitats, vegetation types, land uses, which are inter-related by many spatio-temporal relationships. A local study may focus on vegetation patches (and thus certain data sources and granularity), or in insect-plant interactions (totally distinct data sources and granularity). However, vegetation patches will influence insect-plant interactions and vice-versa. Hence, an open problem here is the work on interactions across multiple scales, including constraint maintenance throughout all levels.

Along the same lines, LIS is dedicating considerable effort in the management of biodiversity data (our work on agricultural applications, also ongoing, was exemplified in section 2.2). Biodiversity information systems are concerned with the environment and natural resources, to help scientists manage information on species and their relationships. The main data source on species are the so-called "collection records" – species occurrence data describing where a given animal (or plant) was observed, when, where, by whom, and using which methodology. This kind of record must be correlated with other kinds of data, most of which geo-referenced (e.g., habitat or climate variables), but also textual data sources such as taxonomic classifications, historical records and others.

Typical queries in such systems combine textual information (e.g., on species – where and when they were observed, by whom and how) with geographic, historical and taxonomic information, characterizing the ecossystems where species were observed, and their spatial distribution. Most queries on such data are of textual nature, and, in some cases, allow direct interaction with maps. One of the lines of research in LIS has been that of designing and developing data storage, mining, querying and visualization tools to provide more flexible mechanisms to handle these data. One of the prototypes available from our site (http://proj.lis.ic.unicamp.br/biocore) concerns the creation of UNICAMP's virtual zoology museum. The final goal is to make all data available to experts, to allow for multiple correlations among records of distinct kinds. Additionally, non-experts will be able to navigate the museum's collection Figure 3 shows a screen copy of a taxonomic tree built dynamically from the database, with which scientists interact to construct their query.

Taxonomic trees are not unique, and evolve with scientific discoveries. Since data are updated in the database, this must be reflected by changing the tree in the interface. Dynamic visualization is the least of the problems. Data evolution (and updates) underly two challenges associated with handling biodiversity records. The first is to keep track of the evolution of species names, and their use in research (biodiversity) papers. The second concerns establishing correlations among species, to better design habitat and interaction models. These challenges are being approached from two directions – storage and semantic issues. On the storage and management perspective, we are combining work on temporal databases and version mechanisms. From the semantics point of view, LIS' work is investigating the use of ontologies. Here, the work is extending the Aondê platform[Daltio and Medeiros 2008], a service that supports several operations on arbitrary ontologies – both developed within LIS' projects and imported from the Web. Aondê allows posing distinct kinds of queries and updates on ontologies – including ontology difference and merge, construction of ontology views, ontology pruning, ranking and integration.

The museum's physical collections have over 30 thousand vertebrates (fish, amphibia, reptiles, birds, mammals) and 59 thousand invertebraes (mostly insects, especially butterflies). This is now being enhanced with over 200 thousand marine invertebrates. Each such animal gives origin to a multitude of records and digital content – basic collection records, annotations, images and sound. The sound data repository alone contains over 2Tbytes of recordings that are being ported to a database [Cugler et al. 2011], while other kinds of data are being subject to distinct data-related research (e.g., on query mechanisms, annotations, content-based retrieval and ontology management – see previous section). A recent result concerns the ongoing development of a multimodal query processing framework [Fedel 2011], which supports distinct user profiles (expert and non-expert alike), and allows combining content-based and text-based parameters. Figure 4 shows a screen prototype designed for this kind of multimodal query, in which users can combine textual and content-based (images) input parameters to query the collections databases. Query processing starts with traditional techniques on text (area 1 of the figure), proceeds to ontology queries (area 3) and combines the result to content-based retrieval (area 2). Similar to the example in agriculture given by figure 2, this figure embodies over 4 years' work in query processing, ontology management and image processing.

3.2 Managing digital content - Components, Composition and Authoring.

Our current research concerning this topic is targeted to the Web. DCCs (see section 2.2) are the foundations of a component-based authoring environment we developed [Santanchè et al. 2009]. The environment is built on top of a Web-based digital library, in which reuse is integrated into the authoring process. Technologies like mashup, Web syndication, on demand software routines and libraries subsidize a client-centric approach to combine distributed resources. This research is directed towards content authoring and management on the Web in two directions:

(i) Expanding the classical approach of multimedia authoring – usually targeted to presentations – to other scenarios. Authoring should be able to support, for example, scientists in their modeling

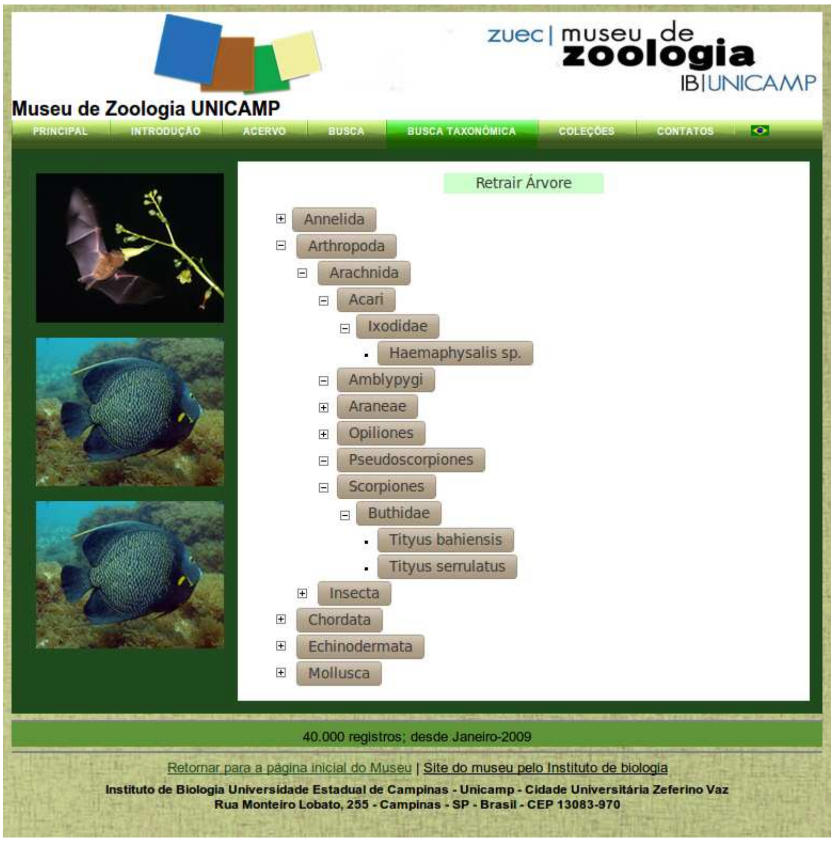

Fig. 3. Screen shot of a prototype of the virtual zoology museum - Taxonomic search tree built dynamically to support queries on species collected

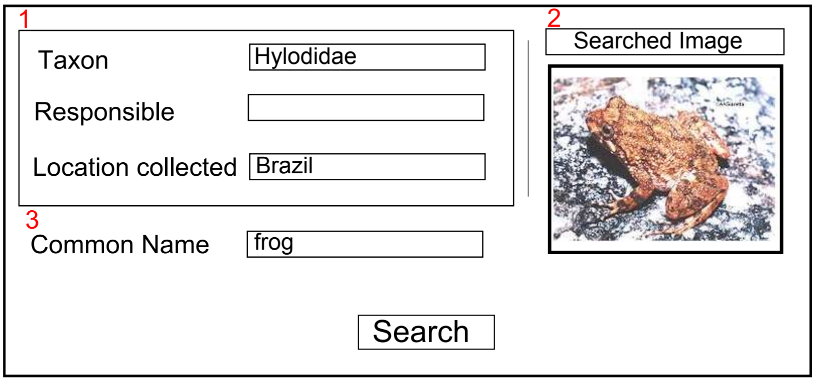

Fig. 4. Multimodal search – combining content-based image retrieval, common names, and taxonomic information.

tasks, data fusion, visual analysis [Santanchè and Baumann 2010].

(ii) Producing a new Web-based authoring environment, which empowers collaboration and amplifies the spectrum of units that can be reused and combined across the Web.

Interoperable semantics and ontologies pervade many topics of LIS research. The central question in this line of research is: how to capture the implicit semantics of content produced by users, transforming it in a machine interpretable representation, in order to expand the participation of machines in content indexation, discovery and handling. It is subdivided in three sub-lines: folksonomized ontologies; In Loco semantics; and generation and management of annotations, with implications on data quality.

The knowledge systematically organized and formalized in ontologies can be enriched and contextualized by the implicit knowledge from folksonomies – i.e., organic taxonomies emerging from people tagging in social environments. Our research in this line proposes a "folksonomized" ontology, a confluence of a formal ontology enriched with social knowledge extracted from folksonomies, e.g., to be used in content search and classification. On the other hand, ontologies are supplied with contextual data, which can improve relationship weighting and inference operations. Folksonomized ontologies can be also used as tools to analyze ontology quality and to help the process of ontology evolution, e.g., showing the discrepancies between the emergent knowledge of a community and the formal representation of this knowledge in the ontology.

Whenever authors produce documents, they follow some patterns, carrying implicit semantics, essential to interpret documents. Some of these patterns are shared by a wide range of document types, while others are more specialized. For example, a learning module, a financial report and a scientific paper share some patterns of construction, e.g., the definition of the main title in the beginning, the division in labeled sections etc. On the other hand, scientific papers have specific patterns to define references – sometimes also present in learning modules – while financial reports have patterns to present budget tables. This observation leads to the essence of our In Loco Semantics [Santanchè and Silva 2010], a methodology for content creation, extraction and transformation, designed to enhance the interpretation and consequently to improve the semantics extracted from this content, by recognizing predefined patterns followed by an author during content creation – the annotation patterns. Examples of these patterns are the application of predefined formatting styles (e.g., title, author etc.), the spatial disposition of fields in a table (e.g., labels left and values right) and so on.

Our current research is going towards interpreting such patterns according to their context, leading to richer semantic interpretations of the content. The first target are spreadsheets produced by scientists (and thus eScience), given that spreadsheets are a primary data source (and tool) used by scientists to store and analyze their data. By recognizing patterns that scientists use to build spreadsheets, we aim to explicit their schemas using semantic Web open standards and to store them in repositories.

We have also been exploiting the semi-automatic construction of annotations of geospatial data [Macário and Medeiros 2009; Macário et al. 2009; Sousa 2010], in which the annotation of data sources is driven by scientific workflows. Annotations are associated with ontologies and stored in repositories, to support retrieval of data sources based not only on metadata standards, but also on annotation semantics. Ongoing work considers annotations to derive information on data quality.

3.3 Load balancing and data migration on the Web

Presently, we are also conducting research in an infrastructure to support data migration on the Web. In particular, this is being investigated in the context of (agricultural) supply chains, which requires work on workflows and composition of inter-organizational business processes encapsulated by Web services. The composition is handled via service choreographies. In this infrastructure, business processes are described in WS-CDL (Web Services Choreography Description Language), and they are performed by a set of coordination managers that execute WS-BPEL (Business Process Execution Language for Web Services) compositions. This also involves developing a mechanism for sharing choreography descriptions through the use of semantic annotations based on ontologies.

Additional research along the same lines concerns the availability and dependability of data providers and services [Nakai et al. 2010; 2011]. The replication of web servers on geographically distributed datacenters allows service providers to tolerate disastrous failures and to improve query response times.

A key issue for good performance of worldwide distributed web services is the efficiency of the load balancing mechanism used to distribute client requests among the replicated servers. LIS research in this topic has involved the analysis of several solutions and the implementation of a simulator that is used to compare those solutions and support the proposal of new ones, built on top of realistic internet models. Our mechanisms have outperformed well known solutions in an extensive set of simulations.

4. FINAL REMARKS - MULTI-FOCUS COOPERATIONS

Research on data and information management has many ramifications and requires cooperation of many kinds of expertise in Computer Science, and LIS has always been a good example of this. An example of a large effort in this sense was LIS' work on the BIOTA database system, for the FAPESP-BIOTA biodiversity information system, the first of its kind in Brazil, aggregating data collected by over 300 scientists from a wide variety of biodiversity-related domains [Fagundes and Medeiros 2000]. The screen copies of prototypes shown in this article also hide long-term cooperations with experts from other domains. Behind Figure 2 lie 16 years of joint work with experts in agricultural sciences, and Figures 3 and 4 are examples of 13 years collaboration with experts in biodiversity. However, this article is restricted to the work developed within the laboratory, by projects coordinated at LIS, and where the students involved developed (or are developing) their work in the laboratory. This corresponds to 14 PhD theses, 70 MSc dissertations, 40+ journal papers and 200+ conference papers, having been assisted by over 80 undergraduate student scholarships. From a strict point of view, several results cannot be classified as "database research", but they are all certainly associated with some aspect of information management. Table I gives an overview of rough numbers associated with the cooperations among the authors of this article, and their work within LIS. It also gives examples of some journals and conferences where papers were published.

Contribution	Databases	Image Processing	Soft. Engineering	Networks	Interfaces
Journal Papers	27 TKDE, VLDB Journal, Geoinformatica, IJDL, Sigmod Record	13 PR, IVC, CAM	3 Multimedia Systems	2 IJCIS	–
Conference Papers	140 CIKM, SBBD, ACM Sigspatial, ECDL, JCDL, Geoinfo, ICDE	50 SIBGRAPI, ICME, VISAPP, ISVC, ICIP, ISMM, ISM	10 SBES, ICEIS	5 IDCS, SBRC	10 CLIHC, HCI, ECCE, ACM Sigspatial
MSc Defences	58	9	1	1	2
PhD Theses	9	3	—	1	1

Table I. Multidisciplinary work in LIS: papers and defences according to areas in Figure 1

Very many people outside UNICAMP also contributed to the results, from groups in Brazil and abroad. In Brazil, strong academic cooperation links have involved, among others, DCC-UFMG, UNIFACS, PUC-Rio, INPE, CNPTIA-EMBRAPA and CIN-UFPE. Abroad, one must single out groups in France (LAMSADE and CNAM), USA (Georgia Tech, U. of Maine and Virginia Tech), Germany (universities of Munster, TU Berlin and Jakobs) and the Netherlands (Eindhoven). Last but not least, there was also intensive research that concerned industrial and/or economic needs. Examples were the work in network infrastructures (associated with CPqD), in security (and fingerprint databases), and in coffee monitoring (co-sponsored by COOXUPE, the largest coffee cooperative in the world).

REFERENCES

Andaló, F. A., Miranda, P. A. V., da Silva Torres, R., and Falcão, A. X. Shape feature extraction and description based on tensor scale. *Pattern Recognition* 43 (1): 26–36, 2010.

Baccarin, E. *Automated Negotiation of Multi-party Contracts in Agricultural Supply Chains*. Ph.D. thesis, IC - Unicamp, 2009. Supervisor E. Madeira, co-supervisor C. B. Medeiros.

Baccarin, E., Madeira, E., Medeiros, C. B., and van der Aalst, W. SPICA's Multiparty Negotiation Protocol: Implementation using YAWL. *International Journal of Cooperative Information Systems*, 2011. accepted.

Barros, R. M. L., Menezes, R. P., Russomanno, T. G., Misuta, M. S., Brandão, B. C., Rivero, P. J. F., Leite, N. J., and Goldenstein, S. K. Measuring handball players trajectories using an automatically trained boosting algorithm. *Computer Methods in Biomechanics and Biomedical Engineering* 14 (1): 53–63, 2010.

Carneiro, A. P. *Análise de Desempenho de Métodos de Acesso Espaciais Baseada em um Banco de Dados Real*. M.S. thesis, IC - Unicamp, 1998. Supervisor G. Magalhães.

Carvalho, C. A., Baranauskas, M. C. C., and Miranda, E. E. Sistemas de Geoinformacao: Tendencias em Tecnologias para Acesso e Interação de Tomadores de Decisão. In *Procs. of Brazilian Symposium on GeoInformatics*. Campos do Jordão, Brazil, 2010.

Costenaro, W. *Characterization of spatial DBMS for performance analysis (In Portuguese)*. M.S. thesis, IC - Unicamp, 1997. Supervisor G. Magalhães.

Cugler, D. C., Medeiros, C. B., and de Toledo, L. F. Managing Animal Sounds - Some Challenges and Research Directions. In *Procs. of e-Science Workshop - within Congresso da Sociedade Brasileira de Computacao*. Natal, Brazil, pp. 18–27, 2011.

Cura, L. M. V., Leite, N. J., and Medeiros, C. M. B. An architecture for content-based retrieval of remote sensing images. In *Procs. of IEEE Int'l Conf. on Multimedia and Expo*. New York, USA, pp. 303–306, 2000.

da S. Torres, R. and Falcão, A. X. Content-based image retrieval: Theory and applications. *Revista de Informática Teórica e Aplicada* 13 (2): 161–185, 2006.

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

da S. Torres, R., Medeiros, C. B., Goncalves, M. A., and Fox, E. A. A Digital Library Framework for Biodiversity Information Systems. *International Journal on Digital Libraries* 6 (1): 3–17, 2006.

da S. Torres, R., Silva, C., Medeiros, C. B., and Rocha, H. V. Visual Structures for Image Browsing. In *Procs. of ACM Int'l Conf. on Information and Knowledge Management*. New Orleans, USA, pp. 167–174, 2003.

Daltio, J. and Medeiros, C. B. Aondê: An Ontology Web Service for Interoperability across Biodiversity Applications. *Information Systems* 33 (7-8): 724–753, 2008.

Daltio, J., Medeiros, C. B., Gomes Jr, L. C., and Lewinsohn, T. A Framework to Process Complex Biodiversity Queries. In *Procs. of ACM Symposium on Applied Computing*. Fortaleza, Brazil, pp. 30–39, 2008.

de A. Oliveira, M. and Leite, N. A multiscale directional operator and morphological tools for reconnecting broken ridges in fingerprint images. *Pattern Recognition* 41 (1): 367 – 377, 2008.

Digiampietri, L. A. *Management of Bioinformatics Scientific Workflows*. Ph.D. thesis, IC - Unicamp, 2007. Supervisor J. C. Setubal, co-supervisor C. B. Medeiros.

dos Santos, J. A., Faria, F. A., Calumby, R. T., da S. Torres, R., and Lamparelli, R. A. C. A Genetic Programming approach for coffee crop recognition. In *Procs. of IEEE International Geoscience and Remote Sensing Symposium*. Honolulu, USA, pp. 3418–3421, 2010.

dos Santos, J. A., Ferreira, C. D., and da S. Torres, R. A Genetic Programming Approach for Relevance Feedback in Region-based Image Retrieval Systems. In *Procs. of Brazilian Symposium on Computer Graphics and Image Processing*. Campo Grande, MS, pp. 155–162, 2008.

dos Santos, J. A., Ferreira, C. D., da S. Torres, R., Gonçalves, M. A., and Lamparelli, R. A. C. A relevance feedback method based on genetic programming for classification of remote sensing images. *Information Sciences* 181 (13): 2671–2684, 2011.

Escalona, M., Torres-Zenteno, A., Gutierrez, J., Martins, E., da S. Torres, R., and Baranauskas, M. C. Development Process for Web Geographic Information Systems. In *Procs. of Int'l Conf. on Enterprise Information Systems*. Galway, Ireland, pp. 112–117, 2008.

Fagundes, A. S. and Medeiros, C. B. Implementing a Metadata Database for an Environmental Information System. In *Procs. of Brazilian Symposium on Databases*. João Pessoa, Brazil, pp. 101–114, 2000.

Faria, F. A. *Uso de Técnicas de Aprendizagem para Classificação e Recuperação de Imagens*. M.S. thesis, Instituto de Computação, Unicamp, 2010. Supervisor R. da S. Torres.

Faria, G., Medeiros, C. B., and Nascimento, M. A. An Extensible Framework for Spatio-Temporal Scientific Database Applications. In *Procs. of Int'l Conf. on Scientific and Statistical Database Management*. Capri, Italy, pp. 202–206, 1998.

Fedel, G. S. *Multimodal search to support biodiversity research (in Portuguese)*. M.S. thesis, IC - Unicamp, 2011. Supervisor C. B. Medeiros.

Ferreira, C. D. *Recuperação de Imagens com Realimentação d Relevância baseada em Programação Genética*. M.S. thesis, Instituto de Computação, Unicamp, 2007. Supervisor R. da S. Torres.

FERREIRA, R. C., MATOS, P. S., ADAM, R. L., LEITE, N. J., AND METZE, K. Application of the minkowski-bouligand fractal dimension for the differential diagnosis of thyroid follicular neoplasias. *Cellular Oncology* vol. 28, pp. 331–333, 2006.

FILETO, R., LIU, L., PU, C., ASSAD, E., AND MEDEIROS, C. B. POESIA: An Ontological Workflow Approach for Composing Web Services in Agriculture. *VLDB Journal* 12 (4): 352–367, 2003.

GATTO, S. D. *Factors that affect the performance of spatial joins in based in real case studies (In Portuguese).* M.S. thesis, IC - Unicamp, 2000. Supervisor G. Magalhães.

GONÇALVES, M. A. AND MEDEIROS, C. B. Constructing Geographic Digital Libraries Using a Hypermedia Framework. *Multimedia Tools and Applications - an International Journal* 8 (3): 341–356, 1999.

GONCALVES, M. A. AND MEDEIROS, C. B. Initiatives that Center on Scientific Dissemination. *Communications of the ACM* 41 (4): 80–82, 1998. Special Issue, Digital Libraries.

KASTER, D., MEDEIROS, C. B., AND ROCHA, H. Supporting Modeling and Problem Solving from Precedent Experiences: The Role of Workflows and Case-Based Reasoning. *Environmental Modeling and Software* 20 (6): 689–704, 2005.

KOGA, I., MEDEIROS, C. B., AND BRANQUINHO, O. Handling and Publishing Wireless Sensor Network Data: a Hands-on Experiment. *Journal of Computational Interdisciplinary Science*, 2011. Accepted.

KONDO, A. A., MEDEIROS, C. B., MADEIRA, E. R. M., AND BACARIN, E. Web Service-based Traceability in Food Supply Chains. In *Procs. of Int'l Conf. on Web Information Systems and Technologies.* Barcelona, Spain, pp. 171–177, 2007.

LEITE, N. J., RIVERO, P. J. F., BREZINKOFER, R., AND BARROS, R. M. L. Morphological algorithm to track markers in 3d kinematic data analysis. In *Procs. of Congress of the International Society of Biomechanics.* pp. 192 – 192, 1999.

LEITE, N. J. AND TEIXEIRA, M. D. An idempotent scale-space approach for morphological segmentation. In *Procs. of Mathematical Morphology and its Applications to Image and Signal Processing.* pp. 291–300, 2000.

MACÁRIO, C. G. N. AND MEDEIROS, C. B. 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.

MACÁRIO, C. G. N., SOUSA, S. R., AND MEDEIROS, C. B. Annotating Geospatial Data based on its Semantics. In *Procs. of ACM SIGSPATIAL Conference.* Seattle, USA, 2009.

MARIOTTE, L., MEDEIROS, C. B., DA S. TORRES, R., AND BUENO, L. TIDES - a new descriptor for time series oscillation behavior. *Geoinformatica* 15 (1): 75–109, 2011.

MEDEIROS, C. B. AND CILIA, M. Maintenance of Binary Topological Constraints through Active Databases. In *Procs. of Int'l Workshop on Advances in Geographic Information Systems.* Baltimore, USA, pp. 127–134, 1995.

MEDEIROS, C. B., PEREZ-ALCAZAR, J., DIGIAMPIETRI, L., PASTORELLO, G. Z., SANTANCHE, A., DA S. TORRES, R., AND MADEIRA, E. WOODSS and the Web: Annotating and Reusing Scientific Workflows. *Sigmod Record* 34 (3): 18–23, 2005.

MONTOYA-ZEGARRA, J. A., PAPA, J. P., LEITE, N. J., DA S. TORRES, R., AND AO, A. X. F. Rotation-Invariant Texture Recognition. In *Procs. of Int'l Symposium on Visual Computing.* Lake Tahoe, USA, pp. 193–204, 2007.

NAKAI, A., MADEIRA, E., AND BUZATO, L. E. DNS-based Load Balancing for Web Services. In *Procs. of Int'l Conf. on Web Information Systems and Technologies.* Porto, Portugal, 2010.

NAKAI, A., MADEIRA, E., AND BUZATO, L. E. Improving the QoS of Web Services via Client-Based Load Distribution. In *Procs. of Simposio Brasileiro de Redes de Computadores e Sistemas Distribuidos.* Campo Grande, Brazil, 2011.

NAKAI, A. M., MACÁRIO, C. G., MADEIRA, E., AND MEDEIROS, C. B. An Infrastructure for Sharing and Executing Choreographies. In *Procs. of Int'l Conf. on Web Information Systems and Technologies.* Madeira, pp. 80–89, 2008.

OLIVEIRA, J., GONÇALVES, M., AND MEDEIROS, C. B. A Framework for Designing and Implementing the User Interface of a Geographic Digital Library. *International Journal of Digital Libraries* 2 (2-3): 190–206, 1999.

OLIVEIRA, J., MEDEIROS, C. B., AND CILIA, M. A. Active Customization of GIS User Interfaces. In *Procs. of IEEE International Data Engineering Conference.* Birmingham, UK, pp. 487–496, 1997.

PASTORELLO, G. Z., GOMES JR, L. C., MEDEIROS, C. B., AND SANTANCHÈ, A. Sensor Data Publication on the Web for Scientific Applications. In *Procs. of Int'l Conf. on Web Information Systems and Technologies.* Madeira, pp. 137–142, 2008.

PASTORELLO JR, G. Z., DALTIO, J., AND MEDEIROS, C. B. A Mechanism for Propagation of Semantic Annotations of Multimedia Content. *Journal of Multimedia* vol. 5, pp. 332–342, 2010.

PASTORELLO JR, G. Z., MEDEIROS, C. B., RESENDE, S., AND ROCHA, H. Interoperability for GIS Document Management in Environmental Planning. In E. Zimanyi (Ed.), *Journal on Data Semantics III.* Lecture Notes in Computer Science, vol. 3534. Springer-Verlag, pp. 100–124, 2005.

PEDRONETTE, D. C. G. AND DA S. TORRES, R. Distances Correlation for Re-Ranking in Content-Based Image Retrieval. In *Procs. of SIBGRAPI Conference on Graphics, Patterns and Images.* Gramado, Brazil, pp. 1–8, 2010.

Penatti, O. A. B. and Torres, R. S. Color descriptors for web image retrieval: A comparative study. In *Procs. of SIBGRAPI Conference on Graphics, Patterns and Images*. Campo Grande, Brazil, pp. 163–170, 2008.

Pires, F. *An Automated Environment for Modelling Geographic Applications*. Ph.D. thesis, IC-Unicamp, 1997. Supervisor C. B. Medeiros, In Portuguese.

Prado, A. B., Baranauskas, M. C., and Medeiros, C. B. Cartography and Geographic Information System as Semiotic Systems: A Comparative Analysis. In *Procs. of Symposium on Advances in Geographic Information Systems*. Washington, USA, pp. 161–166, 2000.

Rivero, P. J. F., Leite, N. J., and Barros, R. M. L. A flexible software for tracking of markes used in motion analysis. *Computer Methods and Programs in Biomedicine* 72 (2): 155 – 165, 2003.

Rivero, P. J. F., Leite, N. J., and Barros, R. M. L. Tracking soccer players aiming their kinematical motion analysis. *Computer Vision and Image Understanding* 101 (2): 122 – 135, 2006.

Rocha, S. R. *Analysis of the performance of spatial joins: a comparison between analytical and experimental methods (In Portuguese)*. M.S. thesis, IC - Unicamp, 2003. Supervisor G. Magalhães.

Santanchè, A. and Baumann, P. Component-based Web Clients For Scientific Data Exploration Using The DCC Framework. In *Procs. of Int'l Conf. on Geographic Information Science*. Zurich, Switzerland, pp. 1–5, 2010.

Santanchè, A. and Medeiros, C. B. User-centered Multimedia Building Blocks. *Multimedia Systems* 12 (4-5): 403–421, 2007a. Springer.

Santanchè, A. and Medeiros, C. M. B. A Component Model and Infrastructure for a Fluid Web. *IEEE Transactions on Knowledge and Data Engineering* 19 (2): 324–341, 2007b.

Santanchè, A., Mota, M., Costa, D. P., Oliveira, N., and Dalforno, C. O. Componere – Autoria na Web baseada em Componentes. In *Procs. of Simp. Brasileiro de Sistemas Multimídia e Web*. Fortaleza, Brazil, pp. 45–54, 2009.

Santanchè, A. and Silva, L. A. M. Document-centered Learning Object Authoring. *Learning Technology Newsletter* 12 (1): 58–61, Jan., 2010.

Sasaoka, L. K. and Medeiros, C. B. Access Control in Geographic Databases. In J. Roddick et al. (Ed.), *Int'l Workshop on Conceptual Modeling for Geographic Information Systems*. Lecture Notes in Computer Science, vol. 4231. Springer, pp. 292–308, 2006.

Schimiguel, J., Baranauskas, M., and Medeiros, C. B. Usabilidade de Aplicações SIG Web na Perspectiva do Usuário: um Estudo de Caso. *Revista IP - Informatica Publica* vol. 8, pp. 7–22, 2006.

Schimiguel, J., Baranauskas, M. C. C., Medeiros, C. B., and Simoni, C. A. C. Guiding the Process of Requirements Elicitation with a Semiotic Approach - A Case Study. In *Procs. of Int'l Conf. on Human Computer Interaction*. Lawrence Erlbaum Associates, Inc, 2005.

Schimiguel, J., Melo, A. M., Baranauskas, M. C. C., and Medeiros, C. B. Accessibility as a Quality Requirement: Geographic Information Systems on the Web. In *Procs. of Latinamerican Conference on Human-computer interaction*. Cuernavaca, Mexico, pp. 8–19, 2005b.

Seffino, L., Medeiros, C. B., Rocha, J., and Yi, B. WOODSS - A Spatial Decision Support System based on Workflows. *Decision Support Systems* 27 (1-2): 105–123, 1999.

Silva, F. R. *An approach to detect outliers in categorical data (In portuguese)*. M.S. thesis, IC - Unicamp, 2004. Supervisor G. Magalhães.

Silva, W. V. *An architecture based in generation of predicates for obtention of spatial association rules (In Portuguese)*. M.S. thesis, IC - Unicamp, 2003. Supervisor G. Magalhães.

Sousa, S. R. *Management of Semantic Annotations of Data on the Web for Agricultural Applications*. M.S. thesis, IC - Unicamp, 2010. Supervisor C. B. Medeiros.

Souza, S. M. F. M. *Um Benchmark voltado a Análise de Desempenho de Sistemas de Informaï£¡ï£¡es Geográficas*. M.S. thesis, IC - Unicamp, 1995. Supervisor G. Magalhães.

Torres, R. S. *An Environment for Managing Images and Spatial Data for Biodiversity Applications Development*. Ph.D. thesis, IC-Unicamp, 2004. Supervisors C. B. Medeiros and A. Falcão.

Vasconcelos, R. C. *A Comparative Analysis of the Use of Relational and Object-oriented models in Geographic Information Systems (In Portuguese)*. M.S. thesis, IC - Unicamp, 1996. Supervisor G. Magalhães.

Vilar, B., Malaverri, J., and Medeiros, C. B. A Tool based on Web Services to Query Biodiversity Information. In *Procs. of Int'l Conf. on Web Information Systems and Technologies*. Lisbon, Portugal, pp. 305–310, 2009.

Weske, M., Vossen, G., Medeiros, C. B., and Pires, F. Workflow Management in Geoprocessing Applications. In *Procs. of ACM Symp. Geographic Information Systems*. Washington, USA, pp. 88–93, 1998.

Zegarra, J. A. M., Papa, J. P., Leite, N. J., Torres, R. S., and Falcï£¡o, A. X. Learning how to extract rotation-invariant and scale-invariant features from texture images. *EURASIP Journal on Advances in Signal Processing* (691924): 1 – 15, 2008.