The Ecobase Project*: Database and Web Technologies for Environmental Information Systems

Ecobase project members **

[1]UFRJ, [2]INRIA, [3]PUC-Rio, [4]IME-Rio, [5]U. Paris 5, [6]Embrapa, [7]UNIRIO, [8]U. Paris 6

Abstract

A very large number of data sources on environment, energy, and natural resources are available worldwide. Unfortunately, users usually face several problems when they want to search and use relevant information. In the Ecobase project, we address these problems in the context of several environmental applications in Brazil and Europe. We propose a distributed architecture for environmental information systems (EIS) based on the Le Select middleware developed at INRIA. In this paper, we present this architecture and its capabilities, and discuss the lessons learned and open issues.

1 Introduction

Over the last years, governments have recognized that environmental information could have a profound impact on our ability to protect our environment, manage natural resources, prevent and respond to disasters, and ensure sustainable development. All these issues emphasize the need to circulate and exchange information and also to combine information across different disciplines using processing programs. Internet as a global network makes it possible to better share environmental information among various users (scientists, administrations, general public, etc.).

Unfortunately, when users want to search and use environmental information on the Web, the following problems occur [TS97]. First, data is not referenced by data suppliers and therefore is hard to locate, or data is referenced under specific classification criteria that are domain-specific. Second, data is hard to access: either private, or at a too high cost, or requiring costly pre-processing (e.g., data must be re-entered manually from paper documentation) or format translation, or still too long acquisition administrative procedures, etc. Third, accessed data sets are hard to use because they are inconsistent or non-compatible (e.g. access to long time series where standard data collection techniques have not been applied, thereby making adjacent time series not compatible). This may entail detailed data identification such that corrections can be made (either in-house or by the data supplier); however, such data identification is often not present. Fourth, accessed data need to be processed by remote, complex programs. These programs, typically written in a 3GL language, implement image manipulation algorithms, weather index analyses and many other useful functions. However, sharing data sources and programs across many users through the Web may be very difficult because of the high cost of locating, extracting and using relevant resources. Fifth, the quality of retrieved data is hard to assess (accuracy, "first-hand" versus derived, timeliness, etc). It is often hard to compare data produced using different scientific models because of a lack of documentation about the underlying computational process.

What is needed is an environmental information system (EIS) that eases access to and manipulation of a large variety of heterogeneous, distributed data and program sources. We can distinguish between three main categories of users based on the data each user needs from an EIS: end-users, brokers and data providers. End-users (e.g., general public, policy-makers) need to locate and extract data that match their interest, or appropriate servers to retrieve data of the desired level of quality or run complex programs. Brokers (e.g., environmental scientists, public administrations) construct the servers for end-users. Data providers (e.g., biologists, geologists) collect data and want to distribute them as widely and as easily as possible. They may also want to provide access to their complex programs.

In the Ecobase project, we address these problems in the context of several environmental applications in Brazil and Europe. The project started in 1998 to share research and experience in EIS between four Brazilian universities (PUC-Rio, UFRJ, IME-RJ, UNIRIO) located in Rio de Janeiro and INRIA. The Caravel database group at INRIA

* Project sponsored by CNPq, Brazil and INRIA, France. http://www.uniriotec.br/ecobase.

** The Ecobase project members are: Luc Bouganim[2], Maria Claudia Cavalcanti[1], Françoise Fabret[2], Maria Luiza Campos[1], François Llirbat[2], Marta Mattoso[1], Rubens Melo[3], Ana Maria Moura[4], Esther Pacitti[5], Fabio Porto[3], Margareth Simoes[6], Eric Simon[2], Asterio Tanaka[7], Patrick Valduriez[8]. The project was managed by Asterio Tanaka and Patrick Valduriez.

has gained experience with EIS through the Thetis European project. Thetis has led to the definition of a general component-based architecture for EIS [HNL+99], and the development of a new middleware system, called Le Select [AMS+00]. In parallel, the Brazilian universities have gained experience in the management of metadata for environmental data, and integration of spatial data.

This paper reports on the main results of the project and discusses open issues. In Section 2, we describe the three environmental applications and their requirements. Section 3 presents our distributed architecture for EIS, which is based on Le Select. Section 4 presents the main capabilities of our EIS design. Section 5 addresses the lessons learned and open issues.

2 Environmental applications

In this section, we introduce three environmental applications dealt with in Ecobase: coastal zone management, agro-ecological and economical zoning, and biopheneomenon corrosion control.

2.1 Coastal Zone Management (Thetis)

The Thetis european project (1997-2000) (http://caravel.inria.fr/Fcontract_THETIS.html) addresses applications of coastal zone management (CZM) over the Mediterranean region. CZM is a methodology for the management of coastal resources with the ultimate goal of improving the development of coastal zones, e.g. by reducing pollution. Environmental scientists and public institutions working on CZM need to access, integrate and visualize data matching their interests from several multinational distributed data sources, across many scientific disciplines such as marine biology, oceanography, chemistry and engineering. There is a wealth of accumulated information about the Mediterranean zone including data and images in heterogeneous databases, files, spreadsheets, video and audio data. The data sources are also fairly autonomous and complex, making it hard to integrate relevant information. Furthermore, scientists need to access program sources such as mathematical models for simulating physical processes of coastal circulation, wave generation, sediment transport, etc. These programs are typically written in conventional programming languages such as C and Fortran, are very complex, run on special-purpose platforms, and can take several hours to execute. They also have their own syntax and semantics, and have different resolution or accuracy.

The objective of the Thetis project was to build an EIS for Mediterranean CZM with transparent access to data and program sources via the Web. Thetis addresses the traditional problems of mediator systems (dealing with large numbers of autonomous, heterogeneous and distributed data sources). But it also addresses major new challenges: accessing autonomous, complex programs; servicing various populations of users (scientists, public institutions) with various levels of expertise; and providing collaborative and interactive capabilities to consolidate and aggregate data.

2.2 Agro-Ecological Economical Zoning (Embrapa)

The Agro-Ecological Economical Zoning project at Embrapa (http://www.cnps.embrapa.br) deals with agriculture and environmental planing in Brazil. There are mainly two types of zoning: the Pedo-Climatic zoning (PCZ) and the Ecological-Economic zoning (EEZ). The former deals with the spatial integration and analysis of climatic and soil aspects in order to evaluate areas suitable for a specific crop, with a spatio-temporal scale. The climatic risk of sowing at the wrong time is assessed together with the land availability for growing a certain crop. EEZ deals with environmental aspects (soil, climate, geology, geomorphology) together with anthropic aspects (land use, land cover) in order to determine the land vulnerability. The spatial analysis of the bio-physical and social-economicalaspects and the related integration (land vulnerability and economical potential) subdivides a region according to its best use, thereby classifying each state of the country in: conservation, preservation, consolidation and expansion zones.

PCZ and EEZ require the integration of many distributed data sources, stored in different formats (files, spreadsheets, maps, conventional and spatial databases). Knowledge-based systems, geographical information systems (GIS), statistical and geostatistical systems, simulation models and decision-support systems are used for simulating crop growth, climate risks, land availability and multiple integration analysis.

The data sources are quite heterogeneous: raw information - soil database, crop requirements database, knowledge rules; derived information - land suitability (for each crop), climate suitability (for each period for a certain crop); spatial information - generated through GIS, organized in maps: soil, land and climate suitability, and zoning maps [TBC+98].

2.3 Corrosion Control (SIMBio)

The SIMBio (System for Interpretation and Modeling of Biophenomena) project [CSL+00] deals with bio-corrosion monitoring on oil platforms over the Brazilian coastal zone. The goal of the system is to help scientists to study corrosion caused by bacteria. Biologists work on bio-corrosion of oil pipes and oceanographers work on ocean stream behavior. But both may be involved in the same environmental problem; for instance, oil spills from underwater pipes.

In order to identify the main cause of these bio-corrosion events, scientists have to collect heterogeneous distributed data and apply an adequate model to analyze the event. First, scientists collect water, soil or pipe samples from the region under investigation. Then, laboratory analyses provide numerical data sets from these samples, such as chemical components' indexes. These data sets are then interpreted or analyzed by means of scientific models in order to derive new data, or some useful conclusion. Scientists from different disciplines have their own models. However, the analysis of oil spills usually requires combining multiple models originating from different disciplines. The choice of a model is usually guided by an archive of previous case studies.

Scientists apply model after model, according to some heuristics, generating a model application sequence, which can be represented as a scientific workflow. The observation of a possible sign of bio-corrosion, a prevention study or even a simple investigation may start a new case study. Bio-corrosion scientists work with a limited quantity of models, through which they can reach some conclusion. However choosing the most adequate model is not simple. In a distributed and multidisciplinary scientific environment, scientists need to browse metadata descriptions, in order to understand models out of their scope of expertise. Thus it is important to describe and represent scientific workflows, models as well as their associated program implementations, to help scientists in choosing the right model.

2.4 Summary of Requirements

We can summarize the main requirements of these environmental applications as follows:

- To locate and efficiently extract relevant and accurate information from a possibly very large number of autonomous, heterogeneous data sources over Internet. This suggests the use of mediator technology.
- To analyze and interpret data using simulation models and other complex analytical programs, thereby generating new value-added data. This suggests the ability to manage scientific models and heterogeneous programs with specific metadata and workflow techniques.
- To store data that is either supplied by data providers, or produced as a result of the two previous tasks. This suggests the use of data warehouse technology.

3 EIS Architecture

To address the requirements of the applications presented before, we adopt a common component-based architecture. In this section, we present this architecture, which is based on the Le Select middleware.

3.1 General Architecture

The architecture is multi-tiered with a client layer, an application layer, a middleware layer and a resource layer (see Figure 1). The lowest layer includes all resources (data and programs) shared by the environmental applications. These resources are published via the Le Select middleware. There are two kinds of application services: extraction service and scientific model management. The client layer provides a Web-based interface to application services.

The extraction service sends queries to the middleware layer to extract data from distributed sources. The result of the queries are appropriately structured by the extraction service and loaded into a data staging area (e.g., a database). This service uses the middleware layer to extract the metadata associated with data sources in order to build a metadata repository or a data warehouse.

The scientific model service distinguishes between regular users and publishers. A regular user basically searches for some scientific solution to a given problem. On the other hand, a publisher is a scientist who proposes a scientific model and wants to share it with other users. The model is typically implemented as a data transformation program. However, it is not trivial to describe the transformation in a way users can easily access, understand and exploit.

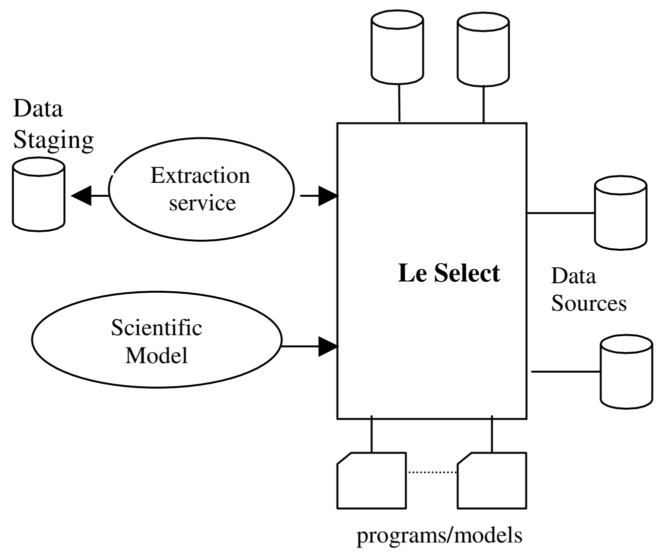

Figure 1: *EIS architecture*

In this architecture, the default is to provide a virtual integration of all resources. Quite often, data and programs cannot be replicated (e.g., for privacy reasons). In addition, the replication of infrequently used data is not cost-effective. However, data replication is necessary when one wants to provide an integrated metadata repository (as in PCZ and EEZ) or when data need be extracted and fed into a data transformation chain (as in SIMBio). In the later case, extracted data must be archived to enable later analysis of the results produced by the data processing chain.

3.2 Le Select Middleware System

Le Select is a successor to Disco [TRV98], also developed at INRIA. From systems like Disco, Le Select retains the general principles of mediator/wrapper architectures, while offering unique features to share both data and programs. The general objective is to allow resource owners to easily publish their resources to the community, give a uniform and integrated view of published resources to potential users, and let them manipulate the available resources through a high-level language. Data remain in their original form and need not be copied or transformed to be published. Similarly, programs remain installed in their original configuration and computer platform.

The publication of a resource requires the installation of a Le Select server at some Internet site (called a publishing site), the writing or configuration of a wrapper at that site, and its registration within the Le Select server. Table wrappers give a uniform representation of data as relational tables, whose columns can take values of user-defined data types, and transform SQL queries into the particular language of the data source. There are generic data wrappers (e.g., XML and JDBC wrappers), which can be easily configured by the publisher. Published data can be either materialized in some store, or computed on-demand by the data source.

For instance, pollution measurements data and satellite images on land use are published at Rio via a table wrapper that exports two tables Poll (region_id, date, value) and Veg (region_id, image). Scientists in Paris publish a Fortran program, which computes the vegetal cover percentage within a satellite image. This program is published by means of a table wrapper that exports a table VegCover (image, coverage). Similarly, scientists in São Paulo publish a program that computes a pollution index from pollution measurements and internal data (not published via Le Select), using a mathematical model. The table wrapper for this program exports a table PollIndex (measurements, index).

Data processing programs are represented as specific «Le Select programs» that take a set of relational tables as input, a set of parameters as arguments, and return a set of relational tables as output.

Published resources can be manipulated through a high-level query language. All resources exported by wrappers (i.e., tables and Le Select programs) have universal names based on their wrapper's URL. Le Select supports a standard SQL select statement to query tables exported by multiple distributed wrappers. For instance, scientists in Brasília wanting to correlate water pollution indexes, computed by a program in São Paulo, with the vegetal cover percentage computed by a program in Paris on data located in Rio de Janeiro, could issue the query:

```
SELECT P.region_id, I.index, C.coverage
FROM Poll P, Veg V, PollIndex I, VegCover C
WHERE P.region_id = V.region_id
   AND I.measurements = P.value
   AND C.image = V.image and I.index > 1.5
   AND C.coverage < 0.3
```

Le Select's language also includes a JOB EXECUTE statement to trigger the asynchronous execution of a Le Select program. Each input table is specified by means of SELECT statements, and arguments are passed by value. Programs execute at the site where they are published, and their wrappers are responsible for getting their operand data from possibly remote Le Select servers, invoking the underlying program, and making their result available as a relational table through a table wrapper.

Unlike Disco, Le Select has a fully distributed peer-to-peer architecture composed of multiple publishing sites (see Figure 2). Each publication site has a complete Le Select server capable of publishing local resources, accessing local or remote resources (published by other servers), as well as processing (optimizing and executing) SQL queries. Thus, all resources published in the network can be queried from any Le Select server. Furthermore, the schema of data and the signatures of Le Select programs are only known to the wrappers that publish them. There is no notion of global catalog and integrated schema.

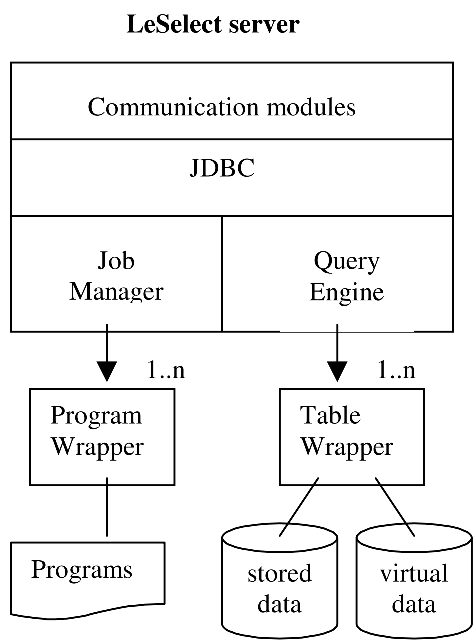

Figure 2: *Architecture of a Publishing Site*

4 EIS Capabilities

In this section, we present in more details the main capabilities of our EIS: extraction service, scientific model management, metadata management and query processing.

4.1 Extraction Service

The extraction service assists end-users to build a customized target database that fulfills the needs of decision-making applications. First, it enables end-users to browse the metadata published by a publication site in order to discover database schema definitions that can be re-used to build their target database schema. This is achieved by translating structural metadata expressed in XML into a DDL script that constructs the database schema. Second, once a target database is available, the extraction service enables extracting data from a publication site and using them to populate parts of the target database relations. This is supported by a user-friendly graphical interface that lets end-users define the mapping between the source and target database schema.

4.2 Scientific Model Management

A scientific model is built based on some assumptions that constrain its use. For instance, a segmentation model may consider a single geographic region. A program implementation of a given scientific model does not need to consider such information in its processing. However, in order to correctly apply the model, it is important to verify whether the assumptions are valid. Therefore, describing a model also means defining the assumptions under which the model can be used. Model management introduces two new problems: (i) how to describe models; and (ii) how to monitor the distributed usage of models, programs and data across scientists.

To address (i), Le Select's publication mechanisms can be used to describe models, similarly to the way it is used to describe programs. Indeed, models also have inputs, outputs and constraints. However, the differences between these definitions must be clarified. When defining model input data types, the meaning of those input data types is more important than their internal representation. Another difference concerns the constraints. Some constraints may be specific to the program implementing the model while some others may be valid for any implementation of the model. Finally, if there are program and model definitions, and a program is an implementation of a model, then the program definition should have a reference to the model it implements. (ii) is an open issue that we discuss in the next section.

4.3 Metadata Management

The multitude of metadata standards [MCB99], designed for specific domains, yields metadata incompatibility in the process of heterogeneous resource integration. Metadata standards for environmental data do not properly address structural components of data repositories. In the Ecobase context, we have worked on a three-layer architecture to support access and extraction processes of environmental data captured from heterogeneous and distributed repositories [TS97]. The first layer represents the information consumers (public and private entities, scientists, etc.). The second layer represents the brokers (those responsible for information integration) and the extractor agents, who communicate directly with each data repository through the use of mediators and wrappers. The third layer corresponds to the data generated by data producers. To support data extraction from heterogeneous sources, we developed a metamodel to support metadata [MPT00] describing the structure of each type of data source. Structural descriptors can be published by Le Select and are of extreme relevance, as extractors need to know how data is organized to correctly access data. The idea is to enrich LeSelect servers with additional metadata descriptions (associated to data sources, programs, models, etc.) that could be collected and organized in a metadata repository to be used by search and retrieval engines.

4.4 Query Processing

Queries like the one in Section 3.2 may involve expensive functions and large objects (e.g. images), and thus may be very inefficient. One possible query execution plan for this query is to join relations Poll and Veg at the Rio site, apply PollIndex on the resulting tuples at the São Paulo site, apply VegCover on tuples which satisfy the predicate over PollIndex at the Paris site and finally transmit tuples which satisfy the predicate over VegCover to the original site in Brasília. A naive execution of this plan, without optimization, can yield very high response time, which stems from multiple image transportation through the network (from Rio to São Paulo, and to Paris), repeated expensive function invocation (VegCover) and sequential execution.

The example shows why standard query execution strategies fail in our context. Indeed, it is reasonable to consider that the time to execute relational operators, including joins, and the time to transfer relational data are negligible compared to the time to process scientific programs and transport large objects. Thus, the problem is not to focus on join ordering in order to minimize the cost incurred by processing joins, but minimizing data transportation and the number of expensive function calls, and maximize parallel execution [BFP+01].

5 Lessons Learned and Open Issues

The experience gained with the development of environmental applications using our EIS architecture taught us several important lessons. In this section, we summarize these lessons and discuss open issues.

Distributed EIS architecture. Our fully distributed peer-to-peer architecture based on Le Select is well suited for environmental applications. It eases data and program publication without requiring an integrated schema. Since any publishing site is powered by a Le Select server, published resources can be easily accessed and combined.

Program publishing. Publishing programs in addition to data sources, and the ability to embed calls within SQL queries proved very useful when dealing with autonomous and heterogeneous data sources. Publishing programs that perform complex data transformations and updates (as in the extraction application service) through program wrappers brings two advantages. First, the extraction/loading facility offered by the program can be shared between different applications. Second, the monitoring of the program, e.g., for refreshing the data repositories, can be delegated to another application that acts as a client to Le Select. An open issue here is the automatic generation of wrappers for programs.

Query processing. Processing queries that deal with expensive functions and large objects requires new optimization techniques [BFP+01]. Furthermore, for applications such as EEZ and PCZ, it is important to find relationships between spatial information, which requires computing spatial joins [LEM00]. However, introducing and optimizing spatial joins in a distributed system like Le Select remains an open issue.

Scientific workflows. Scientific workflow management should handle arbitrary data processing chains. Similarly to data, models and programs, data processing chains should be published through Le Select. This raises the open issue discussed in [CSL+00] of specification formalism, based on metadata metamodel standards.

Scientific models. Publishing scientific models is important and can be done with Le Select's program wrappers. However, monitoring the distributed use of models, programs and data across scientists is an open issue. One approach we are investigating is to let model users define their requirements through model views and have the model publishers provide mappings from their programs to these views. Then, an event monitoring system could register successful mappings and program executions.

Metadata management. Whatever formalism is used to describe resources (data, programs, models, etc.), it must support high-level expressions where variables can range over data and metadata indistinctly [GST98]. It should be based on an expressive formal model, general enough to accommodate all kinds of resources and comprehensive enough to describe semantic and structural characteristics of resources. However, it is not clear yet which metadata framework has the required richness and precision. An interesting approach we are investigating is based on ontologies, which provide powerful constructs to capture richer relationships between concepts [BMW00]. Another issue is the management of a metadata catalog service for exchanging information over resources within the EIS architecture.

Replication. For environmental applications like SIMbio that consolidate information from different sites, a crucial problem arises when base data change at a high frequency rate while there is a strong need to keep a fresh view of the base data. Consider for instance an application that tracks the evolution of an oil spill. Lazy master replication can be used with efficient algorithms that improve freshness [PSM98]. However, an interesting finding is that all base data in Ecobase are time stamped, which eliminates the problem of maintaining replica consistency as addressed in [PMS99]. More work is needed to design refreshment algorithms that further exploit such property.

References

[AMS+00] M. Amzal, I. Manolescu, E. Simon, F. Xhumari, A. Lavric. Sharing Autonomous and Heterogeneous Data Sets and Programs with Le Select, http://caravel.inria.fr/Fprototype_LeSelect.html.

[BFP+01] L. Bouganim, F. Fabret, F. Porto, P. Valduriez. Processing Queries with Expensive Functions and Large Objects in Distributed Mediator Systems. *Int. Conf. on Data Engineering*, Heidelberg, Germany, April 2001.

[BMW00] R. Braga, M., Mattoso, C. Werner. Using Ontologies for Domain Information Retrieval. *DEXA Int. Conf.,* Greenwich, UK, Sept 2000.

[CSL+00] M.C. Cavalcanti, E. Simon, F. Llirbat, M. Mattoso, M.L. Campos. Scientific Experiments Management in Heterogeneous Distributed Database Systems, Technical Report, COPPE-UFRJ, July 2000.

[GST98]H. Galhardas, E. Simon, A. Tomasic. A Framework for Classifying Scientific Metadata. *AAAI Workshop on AI and Information Integration*. Madison, Wisconsin, August 1998.

[HNL+99] C. Houtsis, C. Nikolaou, S. Lalis, S. Kapidakis, V. Christophides, E. Simon, A. Tomasic. Towards a Next Generation of Open Scientific Data Repositories and Services. *CWI Quaterly*, 12(2), 1999.

[LEM00] A. Lima, C. Esperança, M. Mattoso. A Parallel Spatial Join Framework using PMR-Quadtrees. *DEXA Int. Conf.,* Greenwich, UK, Sept 2000.

[MCB99] A. Moura, M. L. Campos, C. M. Barreto. A Survey on Metadata for Describing and Retrieving Internet Resources, *World Wide Web Journal*, 1, 1999.

[MPT00] A.M. Moura, H. Perez, A. Tanaka. A Metadata Model for Supporting Data Extraction from Environmental Information Systems, *Int. Conf. on Geographic Information Science*, Savannah, Georgia, 2000.

[PMS99] E. Pacitti, P. Minet, E. Simon: Fast Algorithms for Maintaining Replica Consistency in Lazy Master Replicated Databases, *Int. Conf. on Very Large Databases*, Edinburgh, 1999.

[PSM98] E. Pacitti, E. Simon, R. Melo: Improving Data Freshness in Lazy Master Schemes, *IEEE Int. Conf. on Distributed Computing Systems*, Amsterdam, 1998.

[TBC+98] A. Tanaka, S. Behring, C. Chagas, S. Fuks. The Brazilian Geo-referenced Soil Information System, *World Congress of Soil Science*, Montpellier, France, 1998.

[TRV98] A. Tomasic, L. Raschid, P. Valduriez. Scaling Access to Heterogeneous Data Sources with DISCO. *IEEE Trans. on Knowledge and Data Engineering*, 10(5), 1998.

[TS97] A. Tomasic, E. Simon. Improving Access to Environmental Data Using Context Information. *ACM SIGMOD Record*, 26(1), March 1997.