

# Digital Library Services Supporting Information Integration over the Web<sup>\*</sup>

Tarcisio Lima<sup>1,2</sup>, Amit Sheth<sup>3</sup>, Naveen Ashish, Mukesh Guntamadugu,  
Sriram Lakshminarayan<sup>4</sup>, Narayanan Palsena, and Dilpreet Singh

Large-Scale Distributed Information Systems Lab./Computer Science Dept.  
415 Boyd GSRC, University of Georgia, Athens GA 30602 USA  
{[tarcisio](mailto:tarcisio@cs.uga.edu), [amit](mailto:amit@cs.uga.edu), [naveen](mailto:naveen@cs.uga.edu), [mukesh](mailto:mukesh@cs.uga.edu), [srir](mailto:srir@cs.uga.edu), [palsena](mailto:palsena@cs.uga.edu), [singh](mailto:singh@cs.uga.edu)}@cs.uga.edu  
<http://lsdis.cs.uga.edu/~adept>

And also at:

<sup>1</sup>Computer Science Dept., Federal Univ. of Juiz de Fora, MG, 36036-300 Brazil

[tlima@dcc.ufjf.br](mailto:tlima@dcc.ufjf.br) <http://www.dcc.ufjf.br>

<sup>2</sup>CS and Statistics Dept., Univ. of Sao Paulo, Sao Carlos, SP, 13560-970 Brazil

[tarcisio@icmc.sc.usp.br](mailto:tarcisio@icmc.sc.usp.br) <http://www.icmc.sc.usp.br>

<sup>3</sup>Taalee, Inc., 263 W Clayton St., Suite 5, Athens, GA, 30601 USA

[amit@taalee.com](mailto:amit@taalee.com) <http://www.taalee.com>

<sup>4</sup>Yahoo! Inc., 3420 Central Expressway, Santa Clara, CA, 95051 USA

[srir@yahoo-inc.com](mailto:srir@yahoo-inc.com) <http://www.yahoo.com>

**Abstract.** Our research and development activities in digital libraries raised relevant features in supporting Web information integration. Underlain by an in house multi-agent based architecture, the main achievements so far have been prototyped as services: (a) various semantic interoperability niches, by the use of inter-ontological relationships built onto *iscapes* (a means of specifying information requests using embedded context sensitive information); (b) integrated access to information, by automating *metabase* (a database of metadata) creation; (c) a framework for creating *iscapes* and metadata modeling; and (d) information processing, by query planning and cost modeling of Web sources. A real-world application scenario illustrates how geographical and environmental Web-based information systems can benefit from appropriating these facilities.

## 1 Introduction

Most mediation and information brokering (e.g., [1], [2], [3]) systems integrate information from multiple sources and allow end users to pose information requests on their repositories. Some mediation systems also allow use of ontologies and 'is-a' relationship (and corresponding subsumption based reasoning) (e.g., [4], [5]). However, very few systems (e.g., [6]) let users explore more complex and meaningful semantic relationships or support information requests built upon information correlations involving semantic relationships.

The overall goal of the Alexandria Digital Earth Prototype Project at the University of Georgia (ADEPT<sub>UGA</sub>) [7] is to develop digital library sources, which are capable of representing geospatial information and meta-information collections in a Digital Earth metaphor. The

---

<sup>\*</sup> This research is primarily based on and part of UGA (University of Georgia) work on the Alexandria Digital Earth Prototype (ADEPT Project, 1999-2004), which is supported by the NSF (National Science Foundation), USA, under grant no. IS IRI-9411330 (Prime contractor: University of California, Santa Barbara. PI: Terrence Smith). Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect the view of the NSF. First author: UGA's ADEPT Project Local Coordinator for its first terms (Fall 1999, Spring and Summer 2000). Second author: UGA's ADEPT Project Principal Investigator.

Digital Earth is a framework for integrating a wide variety of geo-referenced data, including natural, cultural, and historical components. It grows organically with the Web, by contributions from governments, universities, corporations, and individuals [8]. Our context is the development of services that provide a supportive environment for modeling and learning about different complex geographical/environmental phenomena. The ADEPT<sub>UGA</sub> is part of the Digital Libraries Initiative Phase Two (1999-2004), a multi-agency initiative seeking the development of the next generation of digital libraries, advancing the use and usability of globally distributed, networked information resources[9].

The novel feature in maximizing the value of these distributed heterogeneous information resources – addressed initially in the InfoQuilt project [10] and now developed in its follow-on effort, the ADEPT<sub>UGA</sub> – is the introduction of a different kind of information request and integration mechanism called *iscape* (*information landscape*). Iscapes are useful to understand and model geographical/environmental phenomena, typically involving complex relationships between them. Iscapes contain meta-information constructed specifically to represent and explore semantic relationships, in order to facilitate the bridging of the semantic gaps among multiple individual Web information sources [11].

Let us consider an interdisciplinary group of consultants (engineers, geographers, and environmentalists) helping the Federal Government in planning the demographic expansion in a critical area subject to volcanic activities. This real-world application scenario comes in support of one of the suggestions for Digital Earth scenarios that were sampled by the "First Inter-Agency Digital Earth Working Group" [12], an effort on behalf of NASA's inter-agency Digital Earth Program. Among all intricate facets of the proposed situation, one of the problems that raises and needs to be analyzed is: "*How does <volcano> affect the <environment> in the <area>?*". This is typical for an iscape specification: **affect** is the semantic inter-ontological relationship, on top of which we represent the contextualized natural phenomenon to be studied (section 2). The following sections focus on the salient areas identified in order to address the issues regarding the design, construction, and execution of iscapes. At the end we show the ADEPT<sub>UGA</sub> prototype process and some conclusive remarks.

## 2 Semantic Interoperability Using Inter-ontological Relationships

We have developed a framework for the definition of the ontologies and a schema for the generic definition of relationships [7]. Here we discuss how to achieve semantic interoperability among heterogeneous multimedia data using inter-ontological relationships.

### 2.1 "Affect" Relationship into the Iscape Design

The "affect" relationship is of our special interest as it plays a powerful role in describing interactions among elements of different granularities, and is probably the most interesting semantic relationship to study natural phenomena. The domain in the previous real-world application scenario is *natural disasters*. Several ontologies were developed, among them 'Volcano' and 'Environment' are the ones being used in our particular iscape example. Figure 1 shows the "affect" relationship into the iscape design<sup>5</sup>.

Each of these ontologies has sub-components and these sub-components may have an effect on each other. We can immediately see that the "affect" relationship is the composition of sub-relations that may have some conditions if they have to be true. These conditions are the only computable entities, using which we can determine whether the given relation holds for the set of terms from different ontologies at hand. They are used as user modifiable parameters for exercising different possible results. However, these conditions are not strictly neces-

---

<sup>5</sup> Although reflecting a real-world phenomenon, it should be noted that this is a relatively simplified yet coherent exercise with the Digital Earth, that should be refined through actual experience.

sary to generate a relationship statement, in which case it is just a description of how the relationship works.

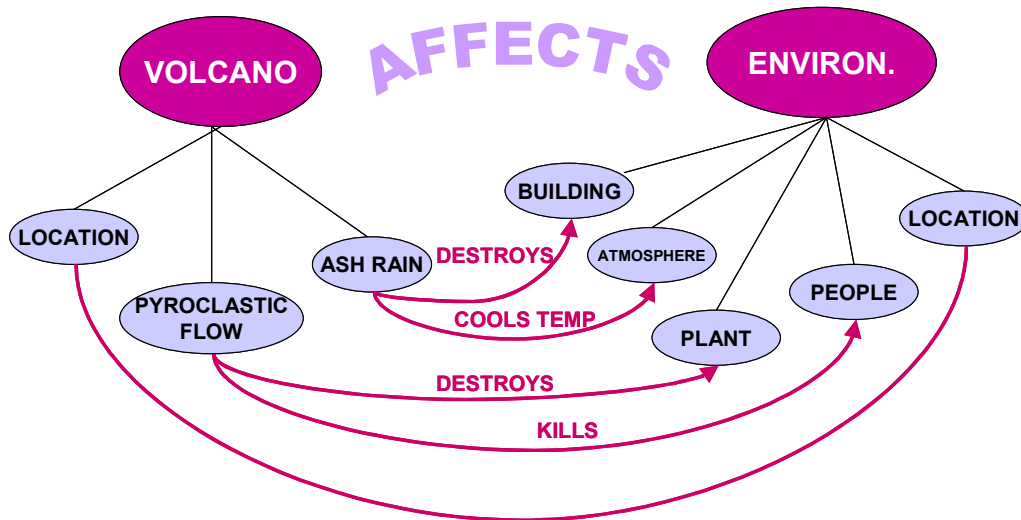


Figure 1 Iscape design: inter-ontological relationships and conditions involved [13]

## 2.2 Mapping Functions

We can define the "affect" relationship in a general way as follows:

$$\begin{aligned}
 &(\exists x \mid x \in \mathbf{ASC}) \text{ and } (\exists y \mid y \in \mathbf{BSC}) \\
 &(\mathbf{FN}(x) \textit{ operator } \mathbf{FN}(y))^* \Rightarrow [\mathbf{ASC} \textit{ relation } \mathbf{BSC}] \\
 &[\mathbf{ASC} \textit{ relation } \mathbf{BSC}]^* \Rightarrow \mathbf{A} \textit{ affects } \mathbf{B}
 \end{aligned}$$

where: **A** and **B** are ontologies; **ASC** and **BSC** are sets containing the sub-components of **A** and **B**, respectively; **FN** is an enclosing function belonging to *FunctionSet*; *operator* and *relation* are sets belonging to *OperatorSet* and *RelationSet*, respectively.

*FunctionSet*, *OperatorSet* and *RelationSet* are defined for the specific domain(s) that we are dealing with. *FunctionSet* typically defines the different enclosing functions that would be used in the relationships. *OperatorSet* defines the various relational operators that are used. It should be noted that these operators could be overloaded, providing different functionality for different enclosing functions. For example, the '=' operator would have to equate weight in a way that is different that equating volume. Finally, the *RelationSet* defines the different sub-relations that form part of the main relationship in the particular domain. For the particular *natural disasters* domain above we have:

$$\begin{aligned}
 \text{FunctionSet} &= \{\text{area, location, time, size, magnitude, height, depth}\} \\
 \text{OperationSet} &= \{<, >, =, <=, >=, \text{INTERSECT}\} \\
 \text{RelationSet} &= \{\text{increases, decreases, kills, destroys, cools temperature of...}\}
 \end{aligned}$$

A partial expression for our iscape would be modeled as in figure 2:

$$\begin{aligned}
 &[\text{Area (Pyroclastic Flow)} \textit{ INTERSECT } \text{Area (People Habitat)}] \text{ and} \\
 &[\text{Time (Volcano)} = \text{Time (Environment)}] \Rightarrow [\text{Pyroclastic Flow} \textit{ kills } \text{People}] \\
 &[\text{Area (Ash Rain)} \textit{ INTERSECT } \text{Area (Building)}] \text{ and} \\
 &[\text{Volume (Ash Rain)} > 100 \text{ cubic meters}] \Rightarrow [\text{Ash Rain} \textit{ destroys } \text{Building}] \\
 &[\text{Location (Volcano)} = \text{Location (Environment)}] \text{ and } [\text{Size (Ash Particle)} < 2 \text{ micron}] \text{ and} \\
 &[\text{Height (Ash Eruption)} > 500 \text{ meters}] \Rightarrow [\text{Ash Rain} \textit{ cools temperature } \text{Atmosphere}] \\
 &\Rightarrow [\text{Volcano}] \textit{ affects } [\text{Environment}]
 \end{aligned}$$

Figure 2 Iscape design: mapping functions

## 2.3 Use of Fuzzy Mapping Functions and Operators

Location of the volcano is geo-referenced and location of a given place is similarly calculated. The "location" function takes the ontology and returns a point. If we were to compare two locations using a singular point, we would hardly get any matches. We know that the effects of a volcanic activity would be felt around an area rather than focused at a point. We may assume a large volcano's effect is felt in a 10-mile radius and the analyzed point should fall within this circular area. The '=' operator is overloaded with respect to the 'Location' function. The concept of overloading is not new, but fits in nicely in the current context.

While comparing time frames, we may use the knowledge that the eruption effects are felt for several weeks around the volcano surroundings rather than at a specific time. If the topic were earthquakes, then this time frame is likely to be in seconds or, at the maximum, in minutes. Thus, given a time and date, we can do temporal matching based on the entity at hand. The main theme in using enclosing functions and overloaded operators is that this scheme provides for interoperating with entities of different granularities and types. We can achieve geo-spatial and temporal interoperabilities, both of them of capital importance while dealing with geographical/environmental systems on the Web.

## 2.4 Simulation Operations

Results from the above mapping functions and computations can be used for further machine processing or human use. For example, simulation operations could be exercised over the Web as an iscape could provide built-ins for running a simulation model. In our iscape example, we could have chosen to simulate the natural population growth surrounding the volcano using appropriate growth models, comparing the results and so forth.

## 2.5 Relationship Information Store

Putting all the ideas together, we arrived at the concept of a relationship store. We developed RELATE [13] to keep track of the different ontologies and relationships created. It gives us a powerful manipulating scheme and allows to infer several facts about the data. Some of the interesting things that can be done are:

- simple questions and answers, explaining how one entity relates to another;
- generalized questions like "How do all natural disasters affect the environment?" – even though we may not have direct information about natural disasters affecting the environment, the repository contains information about how entities which are natural disasters affect the environment;
- open-ended questions like "What are the factors affecting B?" or "What all does A affect?" can be answered with ease, since these types of questions essentially require multiple matching at either end;
- automatically inferring transitive relationships: every time a relationship "A affects B" is created, the repository is cross-checked to see if any other entities affect A or if B affects any entity. Once a match is found, the subset of attributes of the common entity is examined to make sure the intersection is not null;
- using the synonym store we can construct queries using the "like" relation. Thus, any query posed with an ontology not described in our system, but described as a synonym for one of the existing ones, will yield an answer.

## 3 Integrated Access to Information

We may need to create a database of information about entities in a particular domain where the data is extracted, integrated and fused from multiple heterogeneous and autonomously created Web information sources. Such a database is referred to as a *metabase*, a single and quality source of information eliminating the need of browsing individual sources thereafter,

eliminating duplicate results, and providing higher recall besides being fast and efficient. We developed MÉTIS<sup>6</sup> [14] to automate the process of creating metabases. Figure 3 reflects the Métis architecture, briefly described hereafter.

### 3.1 Automating Metabase Creation

The data from each Web source can be extracted using the most appropriate extraction tool for the source, whereas a database wrapper is written for each database to convert the data represented in the internal schema of the database into attribute value pairs. Once the data is structured into generic objects, *mapping rules* (guidelines to convert the source data into a uniform representation) declaratively specified for each source and a set of mapping functions can be used to map the extracted or wrapped data into a canonical representation for objects of each domain. The source objects is then either merged with the existing object for that entity or inserted as a new entity into the metabase. This unification and matching is achieved through a set of *integration rules* (guidelines for object matching) for each domain and matching functions for object matching.

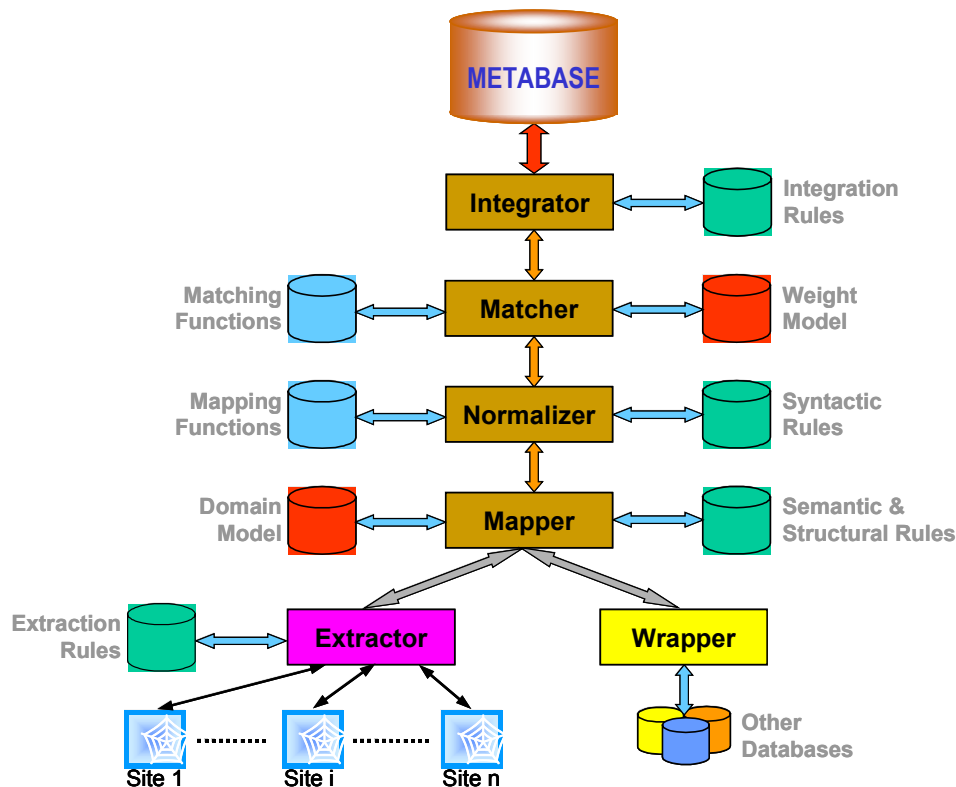


Figure 3 Métis architecture overview [14]

### 4 Framework for Creating Iscapes and Modeling Metadata

The creation of iscapes can be a difficult task as the creator needs to have an understanding of the underlying domain knowledge. Hence, an Iscape Builder Toolkit for easy creation and deployment of iscapes has been developed, using the Java Swing library. As part of it, we have developed an Iscape Definition Language (IDL) [11]. IDL serves as a formal framework to specify what ontologies, attributes and inter-ontological relationship parameters are involved in the construction of an iscape.

<sup>6</sup> Métis is synonym for "mestizo", a person in a mixture of races that makes him/her unique, with the main characteristics of his/her ancestors. In our context, it represents the metabase unifying the best information from multiple heterogeneous sources.



## 5.1 Query Planning and Optimization

Our aim was the development of a query planner for Web based mediators producing high quality query plans. One of the important contributions was the development of a cost model and query cost estimation technique for wrapped Web sources [15]. Traditional database use factors such as average number of tuples, number of blocks access, blocking factor etc. to estimate the cost of executing a query. However, such factors are not relevant for wrapped Web sources, where the cost of executing a query depends on other factors, such as number of pages retrieved, time to retrieve a single page from the server, organization of information to be extracted on the pages on the Web source, etc.

We also explored approaches for choosing query plans with low execution cost from the space of alternative query plans for a given query. We may also consider materializing or caching some data locally to improve the response time for expensive queries. Besides, we have incorporated the concept of simulation (sub-section 2.4) as a specialized operation in our system. As simulations allow configurable parameters, which can be changed for evaluating different results, the simulation operation is highly applied while studying natural phenomena.

## 6 ADEPT<sub>UGA</sub> Prototype Process

To efficiently handle the execution of iscapecs, the process is as follows:

- **Phase 1.** The domains are selected. Ontologies are built using the Iscape Builder Toolkit. Various attributes and relationships existing across ontologies are modeled. Existing distributed data sources are closely studied to model available metadata. The metabase is created according to the ADEPT<sub>UGA</sub> metamodel and domains chosen, using Métis.
- **Phase 2.** The administrator logs into the system and launches the Iscape Builder. The Iscape Builder leads him to construct an iscape using domain specific ontologies and relationships. The iscape is stored in the metabase. Modifications of the iscape is done using the options provided in the user interface.
- **Phase 3.** Users log on to the system and see the iscapecs created. The iscapecs have configurable parameters like selection of ontologies, relationships and attribute values. They change these parameters for "what-if" based new information requests on the digital earth concepts and phenomena. Upon execution of an iscape with user selected parameters, the multi-agent system presents metadata and data from multiple relevant sources. The user analyzes the results to understand causal effects for the domain in study and its supporting information.

Several iscapecs have been already defined and developed as a demonstration of their processing. Due to a lack of space here, we invite the reader for a live demonstration of their execution that can be accessed through the ADEPT project's Web pages [7].

## 7 Conclusion

Significant progress has been achieved in all of the above. Although the context was to provide a learning environment for studying geographical/environmental phenomena, the novel feature of an iscape could be understood as a general high level information request over the Web involving any domain of knowledge. We realized that our new ontological reasoning service, provided by the RELATE store, represents a sophisticated conceptual model of terms and richer relationships, capturing semantics in a promising way. It is challenging as it still requires new means of improvement in order to better support this semantic layer. After constructing and testing the main suite of software agents of our digital library, we also realized that the packaging of our agent technology framework greatly contributes with general services that are strategically useful while dealing with information integration over the Web.

**Acknowledgements.** We are thankful to key past contributors including Clemens Bertram, Krishnan Parasuraman, Vineet Mahajan and Subhajit Ray as they apportioned the initial substrata to this project.

## References

1. R. Bayardo, W. Bohrer, R. Brice, A. Cichocki, G. Fowler, A. Helal, V. Kashyap, T. Ksiezyk, G. Martin, M. Nodine, M. Rashid, M. Rusinkiewicz, R. Shea, C. Unnikrishnan, A. Unruh, and D. Woelk, InfoSleuth: Agent-Based Semantic Integration of Information in Open and Dynamic Environments. *Proceedings of the 1997 ACM International Conference on the Management of Data*. 1997.
2. S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom, The TSIMMIS Project: Integration of Heterogeneous Information Sources, *Stanford University*, 1994.
3. V. Kashyap, A. Sheth, Information Brokering Across Heterogeneous Digital Media – A Metadata-based Approach, *The Kluwer International Series on Advances in Database Systems*, Volume 20, Boston: Kluwer, 248pp., 2000 (August).
4. Y. Arens, C. Chee, C. Hsu, and C. Knoblock, Retrieving and Integrating Data from Multiple Information Sources, *International Journal of Intelligent and Cooperative Information Systems*, 1993.
5. E. Mena, V. Kashyap, A. Sheth, and A. Illarramendi, OBSERVER: An approach for query processing in global information systems based on interoperation across pre-existing ontologies. *Proceedings of the First IF-CIS International Conference on Cooperative Information Systems*, 1996.14. M. Guntamadugu, Métis: Automating Metabase Creation from Multiple Heterogeneous Sources. *Master Thesis. University of Georgia*, 2000.
6. A. Sheth, V. Kashyap, and T. Lima, Semantic Information Brokering: How Can a Multi-agent Approach Help?, in *Cooperative Information Agents III*, Lecture Notes in Artificial Intelligence, M. Klusch, O. Shehory, G. Weiss (Eds.), Vol. 1652, Berlin et al.: Springer-Verlag, 292-311, 1999 (July).
7. T. Lima (Coord.), A. Sheth, N. Ashish, M. Guntamadugu, S. Lakshminarayan, N. Palsena, and D. Singh, ADEPT Project at UGA. <http://lsdis.cs.uga.edu/~adept>, last accessed on March 02, 2001.
8. T. Crockett, Digital Earth: A New Framework for Geo-referenced Data. Quarterly Newsletter of the Institute for Computer Applications in Science and Engineering, Vol. 7, No. 4, December 1998. <http://www.icase.edu/RQ/archive/v7n4/DigitalEarth.html>, last accessed on March 02, 2001.
9. Digital Library II Phase 2 Initiative (1999-2004). <http://www.dli2.nsf.gov/>, last accessed on March 02, 2001.
10. InfoQuilt Project. <http://lsdis.cs.uga.edu/proj/iq/iq.html>, last accessed on March 02, 2001.
11. N. Palsena, A Framework for Creating Information Landscapes and Modeling Metadata in the Context of Digital Earth. *Master Thesis. University of Georgia*, 2000.
12. R. Kahn et al., Digital Earth User Scenario Suggestions, First Inter-Agency Digital Earth Workshop, Greenbelt, MD, June 1998. <http://digitalearth.gsfc.nasa.gov/Scenarios199806.htm>, last accessed on March 02, 2001.
13. S. Lakshminarayanan, Semantic Interoperability in Digital Libraries Using Inter-ontological Relationships. *Master Thesis. University of Georgia*, 2000.
14. M. Guntamadugu, Métis: Automating Metabase Creation from Multiple Heterogeneous Sources. *Master Thesis. University of Georgia*, 2000.
15. D. Singh, Query Planning and Optimization for Mediator Based Web Sources. *Master Thesis. University of Georgia*, 2000.