The ARANEUS Project: Extending Database Techniques to the World Wide Web

G. Mecca[1], P. Merialdo[1,2], A. Masci[2], and G. Sindoni[2]

[1] D.I.F.A. – Università della Basilicata
[2] D.I.A. – Università di Roma Tre
{mecca,merialdo,masci,sindoni}@dia.uniroma3.it

Abstract. The ARANEUS project aims at developing tools for data-management on the World Wide Web. We have implemented a system, called a Web-base Management system, for managing Web data. The system is designed to support several classes of applications: (i) high-level access to data in the Web; (ii) design, implementation and maintenance of Web sites; (iii) cooperative applications on the Web. We discuss the lessons learned from our experiences with the system, ranging from database-style query interfaces to popular Web sites, to the design and implementation of several sites, among which an integrated Web museum, which correlates data coming from several virtual museums on the Web.

1 Introduction

The enormous growth of the World Wide Web suggests that it will soon become not only a uniform interface for sharing data, but also a standard, world-wide distributed computing platform. Next-generation information systems are likely to be based on HTTP-like protocols, hyper-textual front-ends and platform-independent programming languages. This will clearly have a strong impact on the role played by *data* in such systems. Traditional concepts, such as *data-independence* from applications, design methodologies, and even the concept of DBMS need to be reconsidered in this new framework. In our perspective, database management systems will evolve into new and more sophisticated forms of repositories capable of dealing with these new requirements in data manipulation.

We call a *Web-Base* a collection of data of heterogeneous nature, and more specifically: (*i*) highly structured data, such as the ones typically stored in relational or object-oriented database systems; (*ii*) semistructured data, in the Web style. We can simplify by saying that it incorporates both databases and Web sites. We call a *Web-Base Management System (WBMS)* a system for managing such Web-bases, i.e, a system providing functionalities for both database and Web site management. It is natural to think of it as an evolution of ordinary DBMSs, in the sense that it will play in future generation Web-based Information Systems the same role as the one played by database systems today. Coherently

with the nature of the Web, the system should be fully distributed: databases and Web sites may be either local or remote resources.

We have developed a system, the ARANEUS *Web-Base Management System* (a demo of which will be given at SIGMOD'98 [34]), which meets the above requirements. Original features of the system are: (*i*) a data model called ADM for Web documents and hypertext; (*ii*) several languages for wrapping, querying, creating and updating Web sites; (*iii*) methods and techniques Web site design and implementation. The goal of this paper is the discussion of the various experiences we made in developing the ARANEUS WBMS and using it as a support for Web-based applications of various kinds. The focus of the paper is not on the technical novelties of the system, which have been described elsewhere [11, 10, 12], but on an *a posteriori* evaluation of choices, on the lessons learned, and on the evolution of the system due to our experiments with it.

In Section 2, we introduce the architecture of the system and briefly summarize its main modules. Then, in the following Sections, we concentrate on the discussion of our experiences with the system. We refer to three main classes of applications that a WBMS should support: (*i*) *queries*, (*ii*) *Web site design, implementation and management*, and (*iii*) *integration of Web sites*.

In Section 3, we illustrate the main problems concerned with wrapping and querying Web sites, discussing our experiments with the data model, various approaches to wrapping a data-source, and several alternative paradigms we have tested for expressing queries on a Web site. In Section 4, we concentrate on the process of Web-site design and implementation. Here, the focus is on the benefits of adopting a specific design methodology, in the spirit of information systems [14], for Web site design. Finally, experiences at points above are somehow summarized in Section 5 by a site-integration application, the *Integrated Web Museum*, developed using data coming from the Uffizi [5], Louvre [4] and Capodimonte [2] Web sites. Related work is discussed in Section 6.

2 The ARANEUS Web-Base Management System

The ARANEUS WBMS introduces a number of new tools and techniques for managing Web-bases. The overall architecture is shown in Figure 1. The system is implemented in Java and runs on any Java-enabled platform.

The **User Interface** is completely written in HTML, so that end-users and administrators can access the system from any client on the network. The system allows to manipulate data coming from different sites. Sites are divided in *external sites*, i.e., sites administered by third parties, over which the WBMS has no direct control, and *local* or *internal sites*, i.e., sites created and administered by the system, over which it has complete control. Operations allowed on external sites are essentially queries and local warehousing of site data. On the contrary, beside queries, also updates and restructurings can be done on local sites.

For each site, the system manipulates data of heterogeneous nature, and essentially HTML pages and database tables. With respect to external sites, database tables contain data extracted from the site and materialized locally to

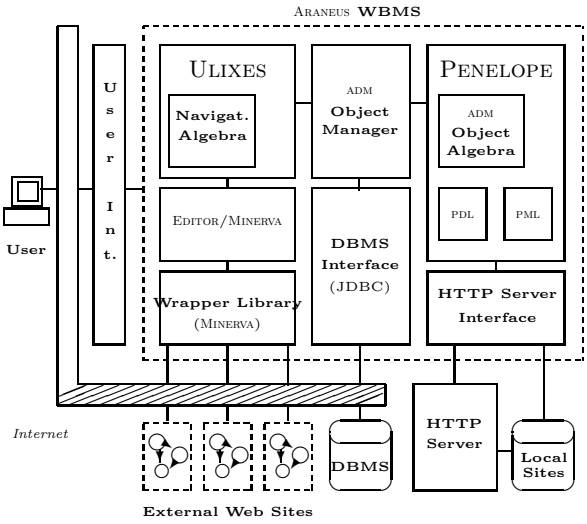

Fig. 1. Architecture of the Araneus WBMS

serve as a basis for further computations. In the case of local sites, tables are used to manage data to be published on the site. Due to this heterogeneous nature of Web-bases, several data models and languages are used. In the following, we briefly discuss the key features and the corresponding components of the system.

2.1 Data Models

In our perspective, structured data are essentially database tables; hence, the data model used to describe structured data is the relational model, and the corresponding language is SQL. Note that other models for structured data, for example object-based, were possible. However, the adoption of a relational model allows us to leverage standard, wide-spread and robust technology and to accomplish a real platform independence. In fact, we assume that the system has access to a (possibly remote) DBMS—as shown in Figure 1—and uses it to store and manipulate tables. Strictly speaking, the DBMS is not part of the WBMS itself; it is more appropriate to say that the WBMS relies on a DBMS to handle tables. Any table-based DBMS—relational or object-oriented—fits in the system; the standard SQL-based protocol JDBC is used to communicate with the database.

A new model, called adm [11] has been developed for handling Web documents. adm represents a key component of the system. In adm each page is seen as a URL-identified nested object. Similar pages are grouped into *page-schemes*, which recall the notion of relation scheme or class in databases. Inheritance is not provided, in favor of *heterogeneous union*, which in our experience better models semistructured hypertexts. Also, other types such as *forms* are introduced to

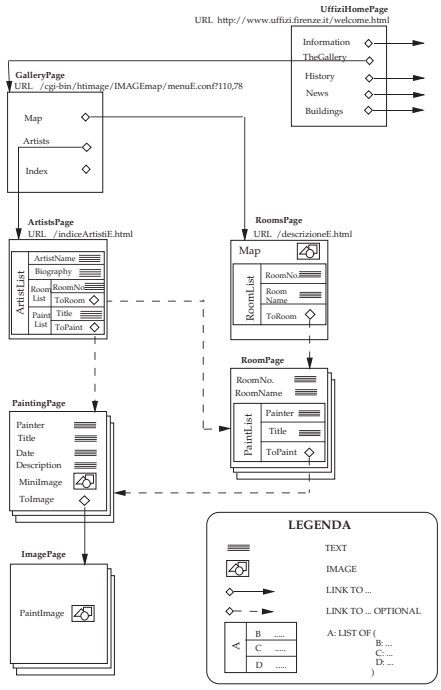

Fig. 2. ADM Scheme for the Uffizi Web site

model Web-specific features. A *Web site* is a collection of page-schemes connected by links. Figure 2 shows a graphical representation of an ADM scheme, with an explanation of the graphical primitives. Further examples of ADM schemes can be found on our Web site [1].

In the system, ADM objects are handled by the ADM **Object Manager** module; it manipulates nested objects by decomposing them and using the DBMS to store them in flat tables.

2.2 Queries over Hypertexts

The system allows to query Web-sites b means of a suitable language called ULIXES [11], which essentially implements a **Navigational Algebra** [35]. External sites need to be wrapped in order to extract data from pages and see them as instances of page-schemes. Wrappers are written using a text-management language, called EDITOR [10]. The system uses an extensible **Wrapper Library** to store wrappers; whenever a page of a certain page-scheme is to be accessed, its source is downloaded from the network and then processed by the corresponding wrapper, which extracts attribute values and stores the resulting ADM object using the **Object Manager**.

2.3 Hypertext Creation and Management

New sites can be created and administered using the system. Here, the idea is that data to be published on the site are stored in the database. Pages in the site can be either materialized or kept virtual. The module that supports the process of defining and maintaining new sites is called PENELOPE [11]. It incorporates the ADM **Object Algebra**, a nested relational algebra with URL invention that can be used to generate ADM views, and therefore HTML pages, over database tables. These views are declaratively defined using the PENELOPE *Definition Language* (PDL), and maintained using the PENELOPE *Manipulation Language* (PML). When a new site is to be generated, its ADM scheme is first designed. This process is supported by a specific design methodology [12]. The new pages may include links to existing pages, and so the new site can be linked to existing ones.

3 Querying the Web

The first experiences with our system were aimed at investigating the chances of re-applying traditional database concepts, such as the ones of *data-model*, *scheme* and *query* to the Web. In the attempt to find a reasonable database-like abstraction from Web data, we had soon to face the need of choosing the right perspective under which we wanted to consider the Web. It is in fact apparent that the richness of the Web leaves great freedom in choosing the level of abstraction for looking at data. It has also been recently argued [29] that a full-fledged data-model for the Web should encompass descriptions of data at different levels, as follows;

- early attempts [28, 36] at defining query techniques for the Web essentially considered the *whole Web as a huge database*; at this level, the graph abstraction seems to be the only reasonable one, since heterogeneities are too big to abstract any common features beside the link topology;
- an alternative is to consider *the Web as a collection of sites, each site being a separate database*; this view of the Web is justified by the fact that Web sites are often quite specific in their offer of data and services, and therefore their description can benefit from more accuracy in the semantics of data and their relationships;
- another extreme may concentrate on *single documents, i.e. HTML pages, each seen as a separate data source*; in some cases, this is also reasonable, since there are pages containing complete databases (for example, stock prices or flight schedules) that might be worth querying by themselves. This approach is substantiated by a large body of research in the field of document query languages [33, 25, 6].

It can be seen that each of these levels represents somehow a compromise between richness of information and accuracy of the description: seeing the Web as a database allows in fact to capture its treasure of information, but at the same

time it forces to shift towards a very high-level description, which captures little of the semantics of data. On the contrary, the description of a single page may be very accurate, but shows little of its connections with the rest of the Web.

When we first started our study, we concentrated on the "Web site as a database" perspective, which in our opinion represents a good compromise, and investigated how much of the database inheritance could be re-used in this context. As a first consequence of this, our study completely ignored the problem of *finding relevant data* in the Web, on which other approaches had concentrated so far. We considered the case in which a user has already identified a bunch of sites containing information of his/her interest, and wants to be able to flexibly access and manipulate data in the site, perhaps to be used as a basis for further computations.

3.1 The Data-Model

One of the key features of our approach was the attempt to re-consider in the Web world the cardinal notion of database scheme as an intensive, compact description of a larger extension of data. Now, the Web contains a plethora of sites, of very different nature. Clearly, only sites whose informative content is of a certain size (in the order to hundreds of pages or above) are worth applying database-style techniques (small sites can be well explored using a browser). Hence, *large Web sites* became the target of our analysis, as a preferred counterpart of databases on the Web. Interestingly, we soon realized that, although the Web as a whole may be well considered a very unstructured data-source, from the logical viewpoint large Web sites are often very structured, and therefore well suited to be described using database techniques.

We have so far developed descriptions for several Web sites, coming from different domains, ranging from museums, to bibliography servers, to sport sites. From the logical viewpoint, all of these sites have a rather tight structure, and can be described in a very compact way using database structures. Note that none of these sites come from an underlying database. Clearly, in other database-generated sites we have examined the structure is even tighter. Based on the analysis of these sites, we refined our definition of the ADM data model, adopting a subset of ODMG, and enriching it with union types in place of hierarchies and forms. In fact, there is no clear counterpart to inheritance hierarchies in this field, but on the contrary *union types* are essential to model heterogeneity among pages; this is especially true for links, whose target page may in some cases belong to the union of several page-schemes (example: for a paper, a link to a conference or a journal page). Also, the form-based access mechanism in the Web, often used in alternative to links, needed to be incorporated in the data model.

The ADM description of a site is derived *a posteriori*, on the basis of a "reverse engineering" process. So far, this is done mainly by hand, with the support of several tools, but we are now studying techniques to automate this process. What we do is essentially to run robots on the site to map the site content and graph topology; then, based on the graph, we examine a sample of pages, to

identify their attributes. Each set of logically homogeneous pages is described by a page-scheme. Even for very large sites, this activity can be completed in a few hours. This of course does not conclude the reverse engineering phase, since appropriate wrappers need to be written in order to build, from HTML pages in the sites, an internal representation of data as instances of the data model. The process of writing wrappers for a Web site can be much more delicate and demanding than expected.

3.2 Wrapping

Wrapping a site consists essentially in mapping logical access to attribute values in a page at the ADM level, to physical access to text in the HTML source. The natural candidate to write wrappers for HTML documents seemed at first to be context-free grammar parsers. These have in fact been used with success in other, more controlled frameworks [6, 16]. Unfortunately, we soon realized that HTML pages are far too complex to be captured by ordinary parsers. In fact, although sites have often a rather tight *logical* structure—i.e., pages can be easily split in homogeneous sets, all pages in a set having essentially the same attributes—they are often very heterogeneous from the *physical* viewpoint, that is, the actual HTML of two apparently similar pages may be radically different. There are a number of reasons for this, the main one being that HTML documents often present heterogeneities and exceptions, or even *errors*; in fact, browsers do not parse the HTML sources they access, and, even in the presence of errors, they manage to display the corresponding page anyway; as a consequence, it is rather frequent the case of pages that do not fully comply to HTML grammar rules.

We soon realized that more flexibility with respect to the one granted by grammars was needed. So, we resorted to a *procedural language*: EDITOR. EDITOR [10] is a language to search and restructure textual documents. It is implemented as a Java-based abstract data type for managing documents, upon which searches can be performed using simple patterns, and restructuring by cutting and pasting regions among documents. Being procedural, EDITOR behaves well in presence of exceptions, which can be explicitly caught in the control-flow of the language, and managed separately on the basis of a case-by-case analysis.

However, it has the usual disadvantage of procedural programming, namely, the lack of declarativity. In fact, the programmer has to explicitly take care of all details of the wrapping, and this in some cases makes the code rather long and boring to write; even more serious, this has a negative influence on the process of maintaining a wrapper. In fact, Web sites are essentially dynamic, and tend to evolve. Although, in our experience, it is quite difficult that the overall logical scheme of a site changes—it never happened to the sites we have examined in the last couple of years—it may happen that the *presentation* of data inside pages, i.e., the physical HTML organization, changes—this happened to one of the sites we had wrapped and forced us to maintain our wrappers.

Having a more compact and declarative formalism for defining wrappers became soon a major necessity, in order to be able to easily change and maintain them. This originated MINERVA [20], an attempt to find a compromise between a

declarative, grammar-based approach, and the flexibility of procedural programming. MINERVA essentially incorporates an explicit exception handling mechanism inside context-free parsers: the user defines a set of productions, describing the grammar of the document; then, the system generates the actual wrapper by translating these declarative specifications into EDITOR code. During the parsing, if one of the productions fails, an exception is raised. This can be explicitly captured by the grammar using the `EXCEPTION` clause associated with every production. The clause reports a piece of (procedural) EDITOR code, which is run against the document in the attempt to restructure it and take care of the exception. If this succeeds, the parsing goes on where it was suspended.

3.3 Query Paradigms and Query Interfaces

Once wrappers for a site have been written, the user can express queries using ULIXES [11]. ULIXES is an SQL-like language for ADM objects, which essentially implements a navigational algebra [35]. Each query returns a set of tuples; these are built by navigating a path in a scheme, with selection and projection conditions.

Following is an example of ULIXES query on the Uffizi Web site [5], whose scheme is shown in Figure 2; the query retrieves the title and position (room name) of all paintings by Michelangelo in the museum. Whenever the query is executed, it starts downloading HTML pages from the site, it wraps them, computes the query and stores the result in a local database table called `MichelangeloPaintings`.

```
DEFINE TABLE MichelangeloPaintings (Title, RoomName)
AS           RoomsPage.RoomList.ToRoom
             → RoomPage.PaintList.ToPaint
             → PaintingPage
IN           UffiziScheme
USING        PaintingPage.Title, RoomPage.RoomName
WHERE        PaintingPage.Painter LIKE '%Michelangelo%'
```

Implementing the prototype of ULIXES posed several interesting challenges. In fact, the system works on ADM objects that are purely virtual, and have to be constructed on the fly by downloading and wrapping actual pages on the site. When the system tries to cross a link and accesses a new page, due to the presence of union types it cannot statically determine the type—i.e., page-scheme—of the page. Therefore, it is forced to adopt a form of dynamical type casting in order to select the right wrapper and methods to apply to the page. This has a number of subtleties. The solution we adopted was to determine, for each page-scheme, a "certificate", that is, an invariant property of HTML sources that uniquely allows to identify instances of that page-scheme. When a page is downloaded, we check its certificate to determine its type and then correspondingly wrap and store it. Another interesting problem was related to managing forms. In our approach, a form is seen as a virtual list that associates a link to a result page with each set of parameters. The language hides all details

relative to crossing forms (parameters are specified as conditions in the `WHERE` clause) but this requires a different treatment for different form methods (`GET` or `POST` with or without redirection).

ULIXES was primarily conceived as a language for writing applications on the Web. However, it may also be used as a tool for casual queries. This is especially useful for very large Web sites where retrieving data based on complex conditions is hardly feasible by simple browsing. To experiment the effectiveness of the language, we made a prototype available to a number of users. From this experience, we have learned that—although quite intuitive—the syntax of the language discourages non-expert users from writing queries from scratch. Casual users found annoying the need to explicitly specify the path to be navigated in the scheme. We therefore started thinking to alternative query interfaces for the system.

First, we have considered the avenue of a visual paradigm; to do this, we developed POLYPHEMUS, a diagrammatic query interface for expressing ULIXES queries. The user can specify a query by interacting with a graphical representation of the ADM scheme; a path is selected by simple mouse clicks on page-schemes, and selection and projections can be easily specified along the path using dialog windows. Then, the user can invoke ULIXES to run the query on the site, and browse the results. We are testing the effectiveness of the formalism. It seems that the main advantages of such approach stand in a more natural perception of the way the site is navigated to reach data of interest.

However, a major drawback in terms of performance is that in this way the user is forced to choose a path in the site, and, in the frequent case in which alternative paths to the same data are available, it is not guaranteed that the chosen one is also the optimal one. In fact, the presence of alternative access paths is a peculiarity of Web hypertexts.

An alternative and promising approach consists in building relational views over a site, and allow users to express queries over these abstractions. We essentially give users the illusion of interacting with a bunch of database tables, which can be queried using SQL.

In general, a declarative query will admit different translations, corresponding to different navigation paths to get to the data. We leave it to the system to translate these declarative queries into navigation of the underlying hypertext. A specific optimizer, CIRCE [35], generates a number of query plans and, based on a cost model, selects an optimal one.

4 Site Creation and Management

We have experienced and tuned our tools in the generation and maintenance of many Web sites in different contexts. Beside some toy applications, we want to mention two official sites of large size: the Faculty of Engineering at University of Basilicata [3], containing more than 500 pages of information about people, education and initiatives of the faculty; and the Web site of the IV Surgical Clinic, at the Umberto I Hospital in Rome (in preparation).

In both cases, the design was conducted according to the ARANEUS Web Site Design Methodology [12], which is based on a clear separation of data management tasks from page design and implementation. Data management is supported by the DBMS, in which all data to be published on site are stored. The logical structure of pages is designed with the help of ADM. Then, PENELOPE [11] is used to map the hypertext structure onto the database and generate pages.

4.1 Different Design Levels and the Need of a Methodology

A key feature of our approach to Web site design that we want to emphasize here is the clear distinction among three different levels: (*i*) database design; (*ii*) hypertext design; (*iii*) presentation design. The separation is justified by the observation that the three levels are largely independent. For example, differently from databases, hypertexts are very redundant data-sources; the same piece of information can occur several times in a site (consider, for example, the name of a course, repeated in every page in which there is a link to the course page); also, to make browsing more effective, usually several different access paths to the same information are given (access to publications may be either by research topic or by year etc.); on the contrary, we would like to store data with as little redundancy as possible, in order to avoid inconsistencies or update problems. Also, the organization of a site may be restructured to make it more effective even without changing the underlying data. Therefore, the right approach seems to have all data stored in a database, and design the site as a hypertext view over the database.

Similarly, the presentation of data in a page may often vary even if the underlying logical structure remains the same. In our approach, we associate an HTML template file with each page-scheme. The HTML template completely specifies the layout of pages corresponding to that page-scheme, and can be changed independently from the page-scheme structure. When PENELOPE generates an instance of a page-scheme, it extracts attribute values from the database and merges them with the corresponding template. It is worth noting that, in our experience, designing an appealing layout for pages is a complicated activity, some times even more delicate than designing the site itself.

4.2 Maintaining the Site and the "Push or Pull" Problem

One aspect, often underestimated, of a Web site life-cycle is the need to maintain the site. Leaving aside major restructurings of the hypertext organization or changes in the presentation, it is still necessary to periodically update the database. Also, updates to the database should be correspondingly reflected on the site.

Note that two different approaches to Web publishing are possible in this context. Database products on the market adopt *pull* techniques, in which pages contain calls to the DBMS, and, when the user requests a new page, such calls are evaluated, the page is generated on the fly and returned to the browser. The

main advantage of this approach is that pages always reflect the most recent database state; however, there are at least two limitations associated with it.

First, if the underlying database has to be used also for other ends—for example, like a repository for a company information system—frequent accesses to the Web site may considerably increase the load on the database and can slow down the overall performance; on the other side, creating a new database especially intended for Web publishing purposes may not be economically feasible and poses further problems to guarantee consistency between the two repositories.

Second, the resulting Web site is strongly platform-dependent: the HTTP server needs a specific DBMS as a back end to serve pages, which often contain non-standard tags to invoke the execution of scripts; this means, for example, that such a site cannot be mirrored or distributed over the network, nor moved to another platform without also migrating the DBMS.

An alternative is represented by a *push* approach, in which data are materialized in HTML files and 'pushed' to the site. This clearly solves the problems above, since the resulting site is standard and the HTTP server works independently from the DBMS; however, in this case, the management of pages in presence of updates is more complex; in fact, when the database is updated, also materialized HTML files need to be correspondingly maintained to reflect the change.

Since push techniques are becoming increasingly popular on the Web due to the appearance of *channels*, i.e., sites that periodically deliver pages or portions of sites directly to the client machine, we decided to support both approaches: in a site, pages can be kept virtual, and generated on-the-fly, or materialized in HTML files. In order to enforce consistency between the database and HTML pages, we had to develop a page-update language and a suitable algorithm for incremental page maintenance [41]. The page update language, called PENELOPE *Manipulation Language* (PML), provides two instructions: `Generate` and `Remove`, which can refer to: (*i*) the whole site; (*ii*) all instances of a page-scheme; or (*iii*) pages that satisfy a condition in a `Where` clause. The page-maintenance algorithm takes as input a database update, and returns a minimal set of PML instructions needed to correspondingly update the pages. In essence, when an update to the database is requested to the system, it automatically generates a *mixed transaction*, in which SQL updates to database tables and PML updates to pages are combined in order to guarantee consistency between the two. The transaction is then atomically executed against the database and the Web site.

4.3 Queries and Meta-Information

When implementing PENELOPE, one of our objectives was to be able to seamlessly extend the query process described in the previous section also to internal sites. This somehow required to keep track of the ADM scheme of the site when generating HTML files, and make it accessible to ULIXES without the need of developing ad-hoc wrappers. The way we did it is by adding *meta-tags* to HTML pages, in order to mark the structure. These meta-tags are embedded in HTML

comments, and thus are completely transparent to ordinary Web browsers; however, they can be used by ULIXES in order to dynamically wrap the page and extract relevant pieces of information. Observe that our tagging mechanism is fully embedded in HTML and does not require extensions. Also, it can be a basis for the definition of extensions of other query languages, such as *W3QS* [28] or *WebSQL* [36], in order to exploit the schema information in querying the site.

5 Site Integration

The two activities described in the previous sections can be nicely coordinated in a larger framework, aiming at integrating data coming from different Web sites. This form of "data-oriented cooperation" requires to identify a number of Web sites containing homogeneous data, extract pieces of information from the sites, correlate these data and then publish everything in a new site, which offers an integrated and possibly reorganized perspective over the original ones.

This is what we have done in developing the *Integrated Web Museum* [1], a Web site containing data and images about paintings in the Uffizi, Louvre and Capodimonte Web sites. This initiative has had an unexpected success, showing that there is great interest for this kind of applications: ever since the site has been indexed by some popular index servers, thousands of accesses per day have been registered.

The museum was developed according to a several-step process that we can summarize as follows: (*i*) *relational view definition and data extraction:* once the original sites have been described in ADM and wrapped, we identify portions of interest and extract them using ULIXES; each ULIXES query defines a relational view over the original sites; (*ii*) *relational view integration:* these views are then processed using the local relational DBMS, to generate an integrated view; also local tables and/or wrapped data sources (files) can be incorporated in the integration process; (*iii*) *hypertextual view definition:* finally, a new Web site is generated as an hypertextual view—defined using PENELOPE—over the integrated relational view. Our approach is therefore to define the global system as a view over the original data sources (see [43] for a discussion of alternative approaches). In the following, we discuss the main issues we have dealt with in developing this application, trying to highlight the problems we faced, and the solutions we adopted.

5.1 Dealing with Schematic and Semantic Heterogeneities

The overall process is rather involved because of the different view levels and models. These levels take care of eliminating heterogeneities between the original data sources. These are essentially of two kinds [39]: *schematic heterogeneities*—i.e., related to the way data are organized at the original data sources—and *semantic heterogeneities*, i.e. related to the way data are represented at the original data sources.

We use our multi-layer view approach to progressively reduce schematic heterogeneities. We want to emphasize the fact that, on the Web, logically similar data may be organized in a radically different way due to presentation choices. Coherently with our approach, we split the two aspects and use the relational level to reason about the logical organization of data, and the hypertext one to reason about the presentation. As an example, it can be seen from Figure 2 that paintings in the Uffizi Gallery are organized by rooms (one page for each room with a list of paintings); on the contrary, they are organized in collections (or departments) in the Louvre and Capodimonte sites (we have omitted the schemes for space reasons; they can be found at [1]). However, beside the specific access paths, the actual data have a very similar logical structure, namely a collection of artists and a collection of works. This form of heterogeneity is dealt with by ULIXES, which takes care of giving a table-based perspective over data in hypertext form.

Once this has been done, integrating the two data-sets amounts to finding the union of two collections. Our approach to this step is rather conservative, in the fact that the logic of the integration process—what is sometimes called the *mediator layer*—is based on SQL, i.e., the integrated relational view is defined as an SQL view over the component tables defined by ULIXES. This, in our experience, is satisfactory in the majority of cases.

Semantic heterogeneities are much harder to manage. A key problem we faced was that of inconsistencies between identifiers. For example, Michelangelo is reported as "Buonarroti Michelangelo" on the Uffizi and Capodimonte Web sites, and simply as "Michelangelo" on the Louvre Site. There are less known authors which are reported with three different names in the three sites. Even in different pages of the same site, the same author may in some cases be referenced to with different names. We were forced to manually write conversion tables in order have unique identifiers for each painter. Statistical approaches seem to be necessary, but they are far from being a final solution to the problem.

6 Related Work

Several systems recently presented in the literature address the problem of managing data coming from the Web. However, they either concentrate on designing query languages for the Web, or on developing tools for Web site generation.

W3QL [28], WebSQL [36] and WebOQL [8], and, with slightly different focus, Lorel [7] and UnQL [17] are prominent examples of query systems designed for semistructured and Web data. The main difference with respect to ARANEUS stands in the choice of the data model: all of these proposals adopt variants of a simple graph-based data model, and concentrate on the development of query languages for these structures. Also, there is no notion of scheme similar to the one of ARANEUS. Other proposals—WebLog [30], ADOOD [24] and FLORID [27]—advocate the use of logic as a formalism for querying the Web.

A notion of scheme similar to the one introduced in ARANEUS has been recently used in WGLog [21], whose aim is at studying graph-based query lan-

guages for the Web, and in WAG [18], which also studies mining and integration problems in the Web framework.

Languages and tools for wrapper generation were first studied in the context of textual databases (see, for example, [6]), and then with specific reference to the Web [26, 9, 42]. All of these approaches use variants of grammars to describe patterns in documents. As discussed in the previous sections, grammars are not completely satisfactory in this context. We therefore try to combine the advantages of declarativity with the flexibility of a procedural language for dealing with exceptions.

Other proposals, namely TSIMMIS [19] and the Information Manifold [31] aim at integrating data from heterogeneous sources, including the Web. These techniques can be used in ARANEUS in order to correlate tabular data and generate integrated views. Integration and queries over data coming from the Web is also the goal of the WebSuite project [15], a collection of modules for wrapping, querying, translating [37] and integrating data from the Web.

Web-site generation is another fertile area (see, for example, [22, 40, 38]). These proposals deal with the problem of implementing a Web site as a view over a set of data sources. They mainly adopt a graph-based model, in the spirit of OEM [19, 7], and have no notion of schema of a site. A different approach is the one undertaken in AutoWeb [23], in which a data model inspired by hypermedia authoring is used to describe a site. Deciding the organization of data in the site is an activity supported by a specific design methodology; based on this design phase, data stored in a relational database is translated into HTML pages.

Several commercial database systems now provide functionalities for the automatic generation of pages. However, also in that case, no data model is used to describe pages and hypertexts. Moreover, these proposals tend to adopt a pull approach to Web publishing, whereas we also support materialized approaches.

7 Conclusions

The ARANEUS Web-base Management System represents a proposal towards the definition of a new kind of data-repository, to serve as a basis for Web-based information systems. We have presented several experiences maturated in the development of the system, and in its use as a support for implementing a number of applications. A number of promising research directions still need to be investigated, as follows.

Inferring Structure from the Web: the reverse-engineering of a Web site would greatly benefit from techniques for automatic schema-finding and wrapping; also, this would reduce the need for schema and wrapper maintenance in the case one site undergoes a major restructuring.

Optimization Techniques for Web Queries: it has been shown in [35] than cost models and optimization strategies for the Web may be significantly different from the ones developed for traditional and object-oriented databases; if the Web becomes the preferred medium for data exchange and access, this issue requires careful investigation.

Efficient Algorithms for Change Detection in Web Sites: if database views are warehoused locally to reduce computation costs, then it is necessary to develop algorithms for efficiently checking Web sites in order to detect updates, i.e., insertions or deletions of pages.

Algorithms for Incremental Hypertext View Maintenance: related to the previous point, if new sites have been built as an hypertext view over existing ones, and pages have been materialized locally, whenever a change is detected in the original pages, the hypertext view should be correspondingly maintained.

References

1. The ARANEUS Project. `http://poincare.dia.uniroma3.it:8080/Araneus`.
2. The Capodimonte Museum Web site. `http://capodimonte.selfin.net`.
3. Faculty of Engineering at University of Basilicata. `http://www.ing.unibas.it`.
4. The Louvre Web site. `http://www.louvre.fr`.
5. The Uffizi Web site. `http://www.uffizi.firenze.it`.
6. S. Abiteboul, S. Cluet, V. Christophides, T. Milo, G. Moerkotte, and J. Siméon. Querying documents in object databases. *J. of Digital Libraries*, 1(1):5–19, 1997.
7. S. Abiteboul, D. Quass, J. McHugh, J. Widom, and J. Wiener. The Lorel query language for semistructured data. *J. of Digital Libraries*, 1(1):68–88, 1997.
8. G. O. Arocena and A. O. Mendelzon. WebOQL: Restructuring documents, databases and Webs. In *Fourteenth IEEE International Conference on Data Engineering (ICDE'98), Orlando, Florida*, 1998.
9. N. Ashish and C. Knoblock. Wrapper generation for semistructured Internet sources. In *Proceedings of the Workshop on the Management of Semistructured Data*, 1997.
10. P. Atzeni and G. Mecca. Cut and Paste. In *Sixteenth ACM SIGMOD Intern. Symposium on Principles of Database Systems (PODS'97)*, 1997.
11. P. Atzeni, G. Mecca, and P. Merialdo. To Weave the Web. In *Intl. Conf. on Very Large Data Bases (VLDB'97)*, 1997.
12. P. Atzeni, G. Mecca, and P. Merialdo. Design and maintenance of data-intensive Web sites. In *VI Intl. Conf. on Extending Database Technology (EDBT'98)*, 1998.
13. P. Atzeni, G. Mecca, and A. Mendelzon Eds. Proceedings of the workshop on the Web and Databases (WebDB'98) (in conjunction with EDBT'98) `http://-poincare.dia.uniroma3.it:8080/webdb98`, 1998.
14. C. Batini, S. Ceri, and S. B. Navathe. *Conceptual Database Design: an Entity-Relationship Approach*. Benjamin and Cummings Publ. Co., 1993.
15. C. Beeri, G. Elber, T. Milo, Y. Sagiv, O. Shmueli, N. Tishby, Y. Kogan, D. Konopnicki, P. Mogilevski, and N. Slonim. WebSuite – a tools suite for harnessing Web data. In *Proceedings of the Workshop on the Web and Databases (WebDB'98)*, 1998.
16. G. E. Blake, M. P. Consens, P. Kilpeläinen, P. Larson, T. Snider, and F. W. Tompa. Text/relational database management systems: Harmonizing SQL and SGML. In *First Intl. Conf. on Applications of Databases, (ADB'94), Vadstena, Sweden. LNCS 819*, pages 267–280. Springer-Verlag, June 1994.
17. P. Buneman, S. Davidson, G. Hillebrand, and D. Suciu. A query language and optimization techniques for unstructured data. In *ACM SIGMOD Intl. Conf. on Management of Data (SIGMOD'96), Montreal, Canada*, pages 505–516, 1996.
18. T. Catarci, L. Iocchi, D. Nardi, and G. Santucci. Conceptual views over the Web. In *Proceedings of the Fourth Workshop on Knowledge Representation meets Databases (KRDB'97) (in conjunction with VLDB'97)* `http://sunsite.-informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-8/`, 1997.
19. S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. D. Ullman, and J. Widom. The TSIMMIS project: Integration of heterogenous information sources. In *IPSJ Conference, Tokyo*, 1994.

20. V. Crescenzi and G. Mecca. Grammars have exceptions, 1998. Submitted for Publication.
21. E. Damiani and L. Tanca. Semantic approaches to structuring and querying Web sites. In *IFIP Working Conference on Database Semantics*, 1997.
22. M. Fernandez, D. Florescu, J. Kang, A. Levy, and D. Suciu. STRUDEL – a Web site management system. In *ACM SIGMOD Intl. Conf. on Management of Data (SIGMOD'97), Tucson, Arizona*, 1997. Exhibits Program.
23. P. Fraternali and P. Paolini. A conceptual model and a tool environment for developing more scalable, dynamic, and customizable Web applications. In *VI Intl. Conf. on Extending Database Technology (EDBT'98)*, 1998.
24. F. Giannotti, G. Manco, and D. Pedreschi. A deductive data model for representing and querying semistructured data. In *Proceedings of the Workshop on Logic Programming Tools for Internet Applications (in conjunction with ICLP'97)*, 1997.
25. G. H. Gonnet and F. W. Tompa. Mind your grammar: a new approach to modelling text. In *Thirteenth Intl. Conf. on Very Large Data Bases (VLDB'87)*, 1987.
26. J. Hammer, H. Garcia-Molina, J. Cho, R. Aranha, and A. Crespo. Extracting semistructured information from the Web. In *Proceedings of the Workshop on the Management of Semistructured Data*, 1997.
27. R. Himmeroeder, G. Lausen, B. Ludaescher, and C. Schlepphorst. On a declarative semantics for Web queries. In *Fifth Intl. Conf. on Deductive and Object-Oriented Databases (DOOD'97)*, 1997.
28. D. Konopnicki and O. Shmueli. W3QS: A query system for the world-wide web. In *Intl. Conf. on Very Large Data Bases (VLDB'95)*, 1995.
29. D. Konopnicki and O. Shmueli. Bringing database functionality to the WWW. In *Proceedings of the Workshop on the Web and Databases (WebDB'98)*, 1998.
30. L. Lakshmanan, F. Sadri, and I. N. Subramanian. A declarative language for querying and restructuring the Web. In *6th Intl. Workshop on Research Issues in Data Engineering (RIDE-NDS'96)*, 1996.
31. A. Y. Levy, A. Rajaraman, and J. J. Ordille. Querying heterogeneous information sources using source descriptions. In *Intl. Conf. on Very Large Data Bases (VLDB'96), Mumbai(Bombay)*, 1996.
32. M. Ley. DataBase systems and Logic Programming bibliography site. `http://-www.informatik.uni-trier.de/~ley/db/index.html`.
33. A. Loeffer. Text databases: A survey of text models and systems. *Sigmod Record*, 23(1):97–106, March 1994.
34. G. Mecca, P. Atzeni, A. Masci, P. Merialdo, and G. Sindoni. The Araneus Web-base management system. In *ACM SIGMOD Intl. Conf. on Management of Data (SIGMOD'98), Seattle, Washington*, 1998. Exhibition Section.
35. G. Mecca, A. Mendelzon, and P. Merialdo. Efficient queries over Web views. In *VI Intl. Conf. on Extending Database Technology (EDBT'98)*, 1998.
36. A. Mendelzon, G. Mihaila, and T. Milo. Querying the World Wide Web. *Journal of Digital Libraries*, 1(1):54–67, April 1997.
37. T. Milo and S. Zohar. Schema-based data translation. In *Proceedings of the Workshop on the Web and Databases (WebDB'98)*, 1998.
38. F. Paradis and A. M. Vercoustre. A language for publishing documents on the Web. In *Proceedings of the Workshop on the Web and Databases (WebDB'98)*, 1998.
39. A. P. Sheth and J. A. Larson. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys*, 22(3):183–236, September 1990.
40. G. Simeon and S. Cluet. Using YAT to build a Web server. In *Proceedings of the Workshop on the Web and Databases (WebDB'98)*, 1998.
41. G. Sindoni. Incremental maintenance of hypertext views. In *Proceedings of the Workshop on the Web and Databases (WebDB'98)*, 1998.
42. D. Suciu Ed. Proceedings of the workshop on the management of semistructured data (in conjunction with ACM SIGMOD 1997) `http://www.research.att.com/-~suciu/workshop-papers.html`, 1997.
43. J. D. Ullman. Information integration using logical views. In *Sixth Intl. Conf. on Data Base Theory, (ICDT'97)*, 1997.