To Weave the Web *

Paolo Atzeni [1], Giansalvatore Mecca [2], Paolo Merialdo [1]

[1] Dipartimento di Informatica e Automazione — Università di Roma Tre
Via della Vasca Navale, 84
00146 Roma – Italy
email: {atzeni,merialdo}@inf.uniroma3.it

[2] D.I.F.A. — Università della Basilicata
Via della Tecnica, 3
85100 Potenza – Italy
email: mecca@dis.uniroma1.it

WWW: http://poincare.inf.uniroma3.it

Abstract

The paper discusses the issue of views in the Web context. We introduce a set of tools and languages for managing and restructuring data coming from the World Wide Web. We present a specific data model, called the Araneus Data Model, inspired to the structures typically present in Web sites. The model allows us to describe the scheme of a Web hypertext, in the spirit of databases. Based on the data model, we develop two languages to support a sophisticate view definition process: the first, called Ulixes, is used to build database views of the Web, which can then be analyzed and integrated using database techniques; the second, called Penelope, allows the definition of derived Web hypertexts from relational views. This can be used to generate hypertextual views over the Web.

*This work was supported by Università di Roma Tre, MURST, and Consiglio Nazionale delle Ricerche.

1 Introduction

As a consequence of the explosion of the World Wide Web [16], an increasing amount of information is stored in repositories organized according to loose structures, usually as hypertextual documents, and data access is based on browsing and information retrieval techniques.

Due to their intuitive nature, browsing and searching present severe limitations [13]. Also, they offer little or no support to a global view of information in the Web, nor to the actual extraction and manipulation of data: specific effort is required to use Web data as input to subsequent computations or to correlate values in Web pages. A step to overcome this problem has been made with recent proposals of query languages for the Web in the database style [24, 25, 28]. They tend to see the Web as a huge collection of essentially unstructured documents, organized as a graph, and allow to pose queries based on the structure of the graph. Little assumption is made on the inner structure of HTML documents. Other proposals [19, 26] aim at integrating data from the Web (and also other sources).

In this paper, we present the approach to the management of Web data as attacked in the ARANEUS project carried out by the database group at Università di Roma Tre. Our approach is based on a generalization of the notion of *view* to the Web framework. In fact, in traditional databases, views represent an essential tool for restructuring and integrating data to be presented to the user. Since the Web is becoming a major computing platform, we believe that also in this field a sophisticate view mechanism is needed, with novel features due to the semi-structured nature of the Web. First, *hypertextual views* should be offered; in fact, the Web can be considered as a uniform interface for sharing data, both in Internet and Intranet environments; thus, in this context, restructuring and presenting data under different perspectives requires the generation of *derived Web hypertexts*, in order to re-organize and re-use portions of the Web. To do this, data from existing Web sites must be extracted, and then queried and integrated: these manipulations can be better attained in a more structured framework, in which traditional database technology can be leveraged to analyze and correlate information.

Therefore, there seem to be different view levels in this framework: *(i)* at the first level, data are extracted from the sites of interest and given a database structure, which represents a first *structured view* over the original semi-structured data; *(ii)* then, further database views can be built by means of re-organizations and integrations based on traditional database techniques; *(iii)* finally, a *derived hypertext* can be generated offering an alternative or integrated hypertextual view over the original sites. In the process, data go from a loosely structured organization — the Web pages — to a very structured one — the database — and then again to Web structures.

In this paper, we present a set of tools that support such a view definition process. Since step *(ii)* is based on rather standard database techniques, we focus on the other two steps. A key point in our approach consists in the use of a specific formalism, the ARANEUS *Data Model* (ADM), for describing the structure of a Web hypertext. Based on the data model, we propose two languages to define views over the Web: the ULIXES language, to define database views over a site, and, symmetrically, the PENELOPE language, to generate a derived hypertext from a database.

The ARANEUS Data Model

To extract data from a Web site, as a very first step we derive a *scheme*[1] of the site. The use of a specific data model is central in our approach: in fact, in order to reason about hypertextual data, we need to describe its logical organization, abstracting from the physical organization (the pages).

We say that ADM is a *page oriented* model, in the sense that the main construct of the model is that of *page scheme*, used to describe the structure of a set of homogeneous pages in a site; the main intuition behind the model is that an HTML page can be seen as an object with an identifier, the URL, and several attributes, one for each relevant piece of information in the page. The attributes in a page can be either *simple*, like text, images, binary data or links to other pages, or *complex*, that is, lists of items, possibly nested. ADM also provides *heterogeneous union* and a *form* type, specifically needed to model the organization of Web pages.

In order to see pages as instances of page schemes we apply suitable text extraction procedures based on the EDITOR language [15], a new language for searching and restructuring text. EDITOR programs are a procedural language for manipulating text, based on "cut & paste" operations. Each page is seen as an object with several extraction methods, one for each attribute of the page; these methods access the HTML source of the page and extract the value of the corresponding attribute. In essence, EDITOR programs act as "wrappers" for pages.

Based on these ideas, the scheme of a Web site can be seen as a collection of page schemes, connected using links. This structured view of the site abstracts some of the properties of the pages, reflecting the user's perspective, and provides a high-level description to be used as a basis for successive manipulations.

It is worth noting that this approach generalizes the models adopted in other query languages for the Web [24, 25, 28], in which pages are considered as essentially unstructured objects; this means that ADM can be used also when there is little structure. However, in addition, it allows to model existing structures and regularities, when considered interesting. In fact, in those sites in which pages referring to similar concepts can be considered as having the same scheme, the model provides a concise and effective description of the site's content.[2]

Database Views and Hypertextual Views over the Web

Based on the scheme, ULIXES and PENELOPE support a two-way data-restructuring process. First, we use ULIXES to build relational views over the Web. Here, the knowledge about the structure based on the ADM description of the site highly facilitates the process of extracting data. To provide a flexible paradigm to access data, we reconsider the issue of *path expressions* [22] in this framework. We introduce the notion of *navigational expression* as a means to express a set of navigations — i.e. paths through pages — in a site; in fact, to access data in the Web, it is natural to start from some entry-point, like, for example, a home page, and navigate until data of interest are found.

[1] Following [8] we could rather use the term *data guide*.

[2] Pages in a site can be often reconducted to a relatively small number of different types; in fact, in order to reduce design and maintenance costs, large sites tend to reduce the amount of heterogeneity among their pages, trying to gain structure [8].

The main idea, here, is that each navigation can be seen as a tuple of values and each navigational expression as a relation. These relations, which can be locally materialized and queried using any database query language, represent a *database view* over a site.

Once data has been extracted, it can be easily manipulated locally. For example, it can be correlated with local data sets, or perhaps integrated with data coming from different sites, using known techniques (see, for example [30, 23]).

Finally, PENELOPE can be used to present the resulting database as a derived hypertext, that is, a *hypertextual view* over the original data sources that can be explored using a Web browser. PENELOPE uses ADM to describe the structure of the resulting hypertext, and allows to define new page schemes to present data under the form of derived pages, which are correlated using a suitable URL invention mechanism borrowed from object-oriented databases in order to generate a complex hypertext.[3] In essence the derived hypertext offers an alternative and integrated view over the original sites; it can involve both new pages and existing portions of the Web, in order to re-organize and re-use data.

Based on these ideas, the view definition process can be summarized as shown in Figure 1: *(i)* Web sites of interest are identified and relevant data are described using ADM; pages are wrapped using the EDITOR language in order to abstract their representation with respect to the physical structure; *(ii)* the ULIXES view language is used to define a set of database views over the sites of interest; these views can be analyzed, manipulated and integrated; *(iii)* finally, the resulting table-based data are restructured using PENELOPE to generate a derived hypertext to be browsed by the user, possibly involving existing portions of the Web.

We have chosen to model database views as relations; however, all the ideas of the paper can be easily adapted to work with object-oriented or object-relational structures as well.

Organization of the Paper

Section 2 provides a discussion of related work. Section 3 introduces the ARANEUS data model (ADM). The ULIXES view language is discussed in Section 4, whereas the language PENELOPE is introduced in Section 5. Due to space limitations, the presentation is mainly informal. All examples in the paper refer to the Database and Logic Programming Bibliography server at Trier [27]. In order to simplify the presentation, only a fragment of the corresponding site is discussed, and some of the features have been simplified. Moreover, all names in the paper are completely fictional. An on-line demo of a prototype of the view languages can be found at the ARANEUS project home page [2], where more examples on different Web sites are presented.

[3] In order to keep track of the structure, PENELOPE automatically adds *meta-information* to the derived pages. Meta-tags are used to mark relevant attributes inside pages, so that their values can be easily extracted from the HTML sources. For space limitations, we do not comment on these issues here.

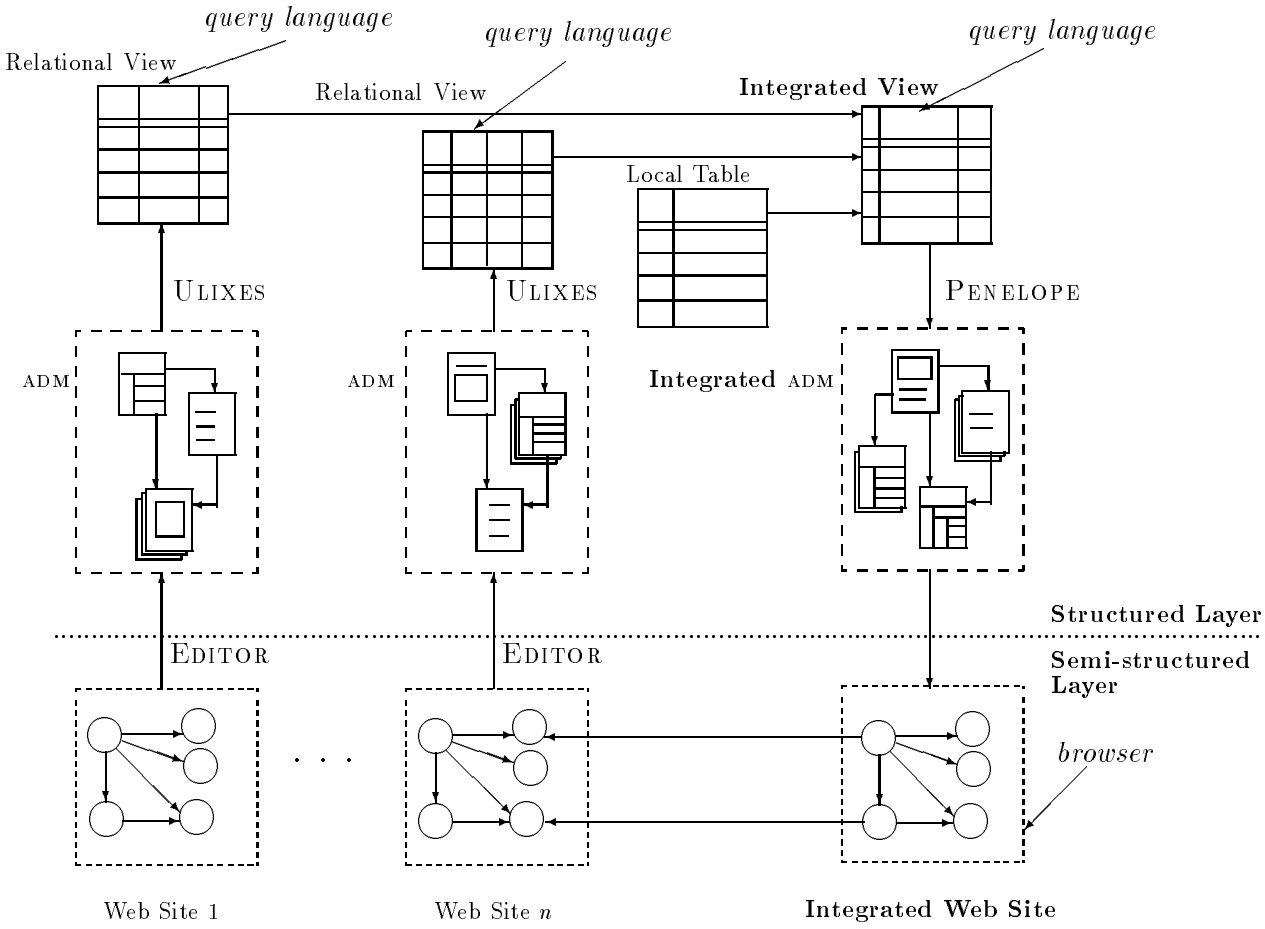

Figure 1: The view definition process in ARANEUS

2 Related Work

The ARANEUS data model can be considered as a subset of ODMG [18], in the sense that the notion of page-scheme can be assimilated to the notion of class. However, there are some important differences, motivated by the nature of HTML documents: first, there is only one collection type in ADM, namely, lists; moreover, inheritance is not present in ADM, whereas heterogeneous union is supported; also, identifiers in ADM – that is, URLs – are visible to the user and can be queried like any other value; finally, ADM provides a *form* construct, which is specific of the Web framework.

Several query systems for unstructured data have been recently proposed. W3QL [24] allows for expressing both structure specifying queries, based on the organization of the hypertext, and content queries, mainly based on information retrieval techniques. Moreover, the language has been designed to be highly extensible, and tools for managing Web forms are presented. WebSQL [28] is another query language in the spirit of W3QL, in which the effort is towards formalization and some interesting ideas are introduced, such as the one of query locality. Other interesting ideas, like the one of restructuring, are introduced in the WebLog query language [25], whose syntax and

semantics are based on logic. In all of these languages there is a very simple notion of scheme, and pages are considered within a single type, i.e. as nodes in a graph, with at most a fixed set of attributes. Moreover, the idea of derived Web structures is not investigated. ARANEUS builds on these proposals, trying to generalize them. In fact, if needed, ADM allows to see the Web as a collection of essentially unstructured pages connected using links (see Section 3); however, in addition, it can also be used to describe the inner structure of pages by means of a scheme, thus providing a finer description of data organization.

Other proposals, namely Lorel [12] and UnQL [17], aim at querying heterogeneous and semi-structured information. With respect to ARANEUS, these languages adopt a *lightweight* data model to represent data, based on labeled graphs, and concentrate on the development of powerful query languages for these structures; on the contrary, ADM provides more complex structures, such as lists of tuples. Moreover, in both these proposals, there is no notion of scheme similar to the one of ARANEUS. In fact, we consider the scheme very important in order to reason about data organization and to use high-level tools for manipulating data.

In ARANEUS, unstructured HTML documents are analyzed to extract their structure; this has several points in common with the management of textual data in the OQL-doc project [10, 9]. Although OQL-doc is not a query language designed for the Web, it shares with ARANEUS the idea of giving structure to unstructured data, and then use high-level database languages to pose queries. However, in OQL-doc, this activity is based on the use of context-free grammars and *structuring schemes*. When dealing with HTML documents, grammars show important limitations. First, the structure is not always completely defined; moreover, the structure can be irregular; finally, HTML documents often contain *errors*, in the sense that they do not fully comply to HTML grammar rules; missing tags are a common example of these errors. To overcome these problems, we adopt a different approach, based on the EDITOR language [15], which, being procedural, allows for a more flexible manipulation of text than traditional parsers.

The use of path expressions [22] in the ULIXES view language inherits some ideas from all of the previous languages. In fact, in Section 4 we show how it is possible to express some WebSQL queries using ULIXES. However, while all of the previous languages allow for expressing recursive paths, here, for the sake of simplicity, we discuss only simple path expressions without recursion. This is sufficient for expressing most "real-life" navigations. However, the language can be extended with recursion in order to enhance the expressive power.

Other proposals [19, 26] aim at integrating data from the Web. The TSIMMIS system [19] extracts data from heterogeneous (possibly semi-structured) sources, and correlates them in order to generate an integrated database representation. Specific translators are written for the various sources. Information Manifold [26] focuses on databases accessible through an interface based on fill-in forms, and provides a specific support for querying, on the basis of declarative descriptions of the contents. These techniques can be used in ARANEUS in order to correlate tabular data and generate integrated views.

Several commercial database systems (see, for example, [6, 3]) now provide functionalities for

the automatic generation of pages. However, they mainly allow for generating a single page at a time, containing a set of database tuples; usually, the skeletons of pages are kept inside the database; pages contain specific HTML tags, which specify that, in order to fill one page, an SQL query is to be run against the database: once the query has been executed, the resulting table is inserted in the body of the page returned to the user. The language PENELOPE described in this paper can be used to this end; however, it can also generate a whole hypertext based on the database content. Finally, the issue of hypertext structure is not addressed in commercial products.

3 ADM: A Logical Data Model for Web Hypertexts

The ARANEUS Data Model (ADM) is a variant of ODMG [18], specifically tailored to the Web context. We say that it is *page oriented* in the sense that it recognizes the central role that pages play in this framework.

Each Web page is considered as an object with an identifier (the URL) and a set of attributes. We introduce the notion of a *page scheme*, which resembles the notion of relation scheme in relational databases or class in object-oriented databases, to model sets of homogeneous pages. Attributes of a page may have simple or complex type. Simple type attributes are *monovalued* and correspond essentially to text, images or links to other pages. Beside monovalued attributes, Web pages often contain collections of objects, that is, multivalued attributes. We model them using *lists* of tuples. Component types in lists can be in turn multivalued, and therefore nested lists may arise. It should be noted that we have chosen lists as the only multivalued type since repeated patterns in Web pages are physically ordered. Consider for example the page in Figure 2 from the DB & LP Bibliography Server at Trier [27]. It refers to fictional publications of author Leonardo Da Vinci. The page has a monovalued attribute, the name of the author; it also has a multivalued attribute, consisting of the list of works; each item in the list can in turn be described as a tuple having attributes such as the title, the authors and so on. Note that, for each author in the site there is a similar page, and all of these pages share the same structure.

There are some specific aspects in this framework with no counterpart in traditional data models. First, an important construct in Web pages is represented by *forms*. Forms are used to execute programs on the server and dynamically generate pages. ADM provides a *form type*, which is essentially considered as a *virtual list of tuples*; each tuple has as many attributes as the fill-in fields of the form, plus a link to the resulting page;[4] such lists are virtual since tuples are not stored in the page but correspond to submissions of the form. Second, usually a site includes pages that have a special role and are unique, in the sense that there are no other pages with the same structure. Typically, at least the home page of each site falls in this category. For the sake of homogeneity, we also model these pages by means of page-schemes. Both of these aspects can be seen in the page in Figure 3 from [27]. The page is unique since no other page in the site has the same structure. Moreover, an essential feature of the page is the form for searching the author

[4] Forms introduce several specific data types, such as *checkboxes*, *radios* or *selections*. We ignore these aspects here for simplicity, and consider only attributes of type text. All ideas can be easily generalized to the most general case.

Figure 2: A page containing the publications by Leonardo Da Vinci

database: by specifying a string corresponding to the name of an author, the list of publications of that author is returned. We see the form as a virtual list of tuples with two attributes: the first is the text entered as a string to search; the second is a link to the page generated by the corresponding search; each tuple associates a result page with a search keyword; tuples in the list cannot be directly accessed, but they must be "built up", in the sense that a keyword has to be specified in order to generate a result page.

Other constructs of the model are due to the semi-structured and heterogeneous nature of Web data; besides *optional* attributes, which allow to describe pages with missing features, ADM also provides a *heterogeneous union* type [21], which seems more appropriate to this context than traditional class inheritance [8]. Consider again the page in Figure 3 from [27]. The search form has a rather involved behavior: when a string is specified, the database of author names is searched; if a single name matching the query string is found, the author's page with his/her publications is returned; on the contrary, if the query string matches several names, a different page containing the list of all matching names, along with links to the corresponding pages, is returned. Union is essential to model this behavior: in this case, we say that the form returns a link to a type which is the union of two different page-schemes, one for author pages and the other for name-index pages.

We can define ADM types as follows. Given a set of *base types* containing the type `TEXT`, a set of *attributes* and a set of *page-scheme names*, the set of ADM *types* is recursively defined as follows (each type is either monovalued or multivalued):

- each base type is a *monovalued* ADM *type*;

Figure 3: The page for searching the author database

- LINK TO D is a *monovalued* ADM *type*; D, the destination of the link, is *(i)* either a page-scheme name, P, *(ii)* or a *union type*, P_1 UNION P_2 UNION $\ldots P_n$, where each P_i is a page-scheme name;

- LIST OF$(A_1 : T_1, A_2 : T_2, \ldots, A_n : T_n)$ is a *multivalued* ADM *type*, if $A_1, A_2, \ldots, A_n$ are attributes and $T_1, T_2, \ldots, T_n$ are ADM types;

- FORM$(A_1 : T_1, A_2 : T_2, \ldots, A_n : T_n)$ is a *multivalued* ADM *type*, if $A_1, A_2, \ldots, A_n$ are attributes and $T_1, T_2, \ldots, T_n$ are ADM types; exactly one attribute has type LINK TO D: this is used to denote the URL of the page generated in response to the submission of the form.

An ADM *page-scheme* has the form $P(A_1 : T_1, A_2 : T_2, \ldots, A_n : T_n)$, where P is a page name, each A_i is an attribute and each T_i is an ADM type. Attributes may be labeled *optional*. The page-scheme may be labeled as *unique*.

Here are some examples of page-schemes from [27] declared in the ARANEUS *Data Definition Language (DDL)*. Consider first page-scheme `AuthorPage`, which describes pages such as the one in Figure 2.

```
PAGE SCHEME AuthorPage
              Name:        TEXT;
              WorkList:    LIST OF (Authors:        TEXT;
                                    Title:          TEXT;
                                    Reference:      TEXT;
                                    Year:           TEXT;
                                    ToRefPage:      LINK TO ConferencePage UNION JournalPage;
                                    AuthorList:     LIST OF (Name:        TEXT;
                                                             ToAuthorPage:LINK TO AuthorPage
                                                                          OPTIONAL;););
END PAGE SCHEME
```

The page has one monovalued attribute (the `Name`) plus a multivalued attribute (the `WorkList`), containing the list of publications; this, in turn, is a set of nested tuples. Note how we choose to

model information about publications: for each paper, we extract the title, the reference (conference or journal the paper was published in) and the year; moreover, we also have a list of authors; for each author, we report the name and a link to the corresponding page in the site; this link is optional since it is not present for all authors. However, we choose to have a slightly redundant representation, in the sense that we also extract a string containing the names of all authors. This has several advantages in terms of querying the site, allowing to pose conditions on all author names, and can be done since the page-scheme is only one of the possible descriptions of information in the page, reflecting the user's personal view of the site, and can be adapted on the basis of efficiency needs. Each citation also contains a link, `ToRefPage`, to the corresponding conference or journal page; since pages for conferences and journals have different structure, we use a union type to model the link.

Note that, to see actual pages in the site as instances of the page-scheme, we need to access the HTML source and apply suitable text restructuring procedures. Attribute values are extracted from the HTML source using the EDITOR language [15], a formalism for text manipulation. In the current implementation we wrap pages using Java classes; every page-scheme in the site corresponds to a specific class, with one method for each attribute; each method implements an EDITOR program that accesses the HTML source and returns a complex value for the attribute.

Another example is the page-scheme `AuthorSearchPage`, describing the page in Figure 3.

```
PAGE SCHEME AuthorSearchPage UNIQUE
               URL /indices/a-tree/index.html
               NameForm:   FORM (Name:       TEXT;
                                 Submit:     LINK TO AuthorPage UNION IndexPage);
END PAGE SCHEME
```

The page-scheme is `UNIQUE`, in the sense that it has a single instance in the site, whose URL is esplicitly mentioned. It also contains an attribute of type FORM. The form has two attributes: the keyword to search for, and the link to the search-result page. Note how the link has a union type; in fact, as discussed above, based on the search results, pages with different structures are returned. Note also that the actual page in Figure 3 also contains an index of all author names based on initials: however, we consider it irrelevant in order to navigate the site and choose not to model it.

We define the notion of ADM *scheme* as a set of page-schemes. We can represent the scheme as a directed multigraph; nodes in the scheme graph are page-schemes; each unique page-scheme in the diagrammatic representation is denoted as a single page, whereas non-unique page-schemes are represented as "stacks" of pages; edges are used to denote links. A fragment of the DB & LP Bibliography server scheme is shown in Figure 4, which also contains an explanation of the other symbols.

Based on this perspective, at the instance level, a site can be seen as a graph in which links connect trees corresponding to pages. In fact, each instance of a page-scheme is a tree (because of its nested structure), and may contain links to other instances. Nodes of trees (and therefore of the overall graph) are essentially tuples; each tuple attribute may either have a simple value or be

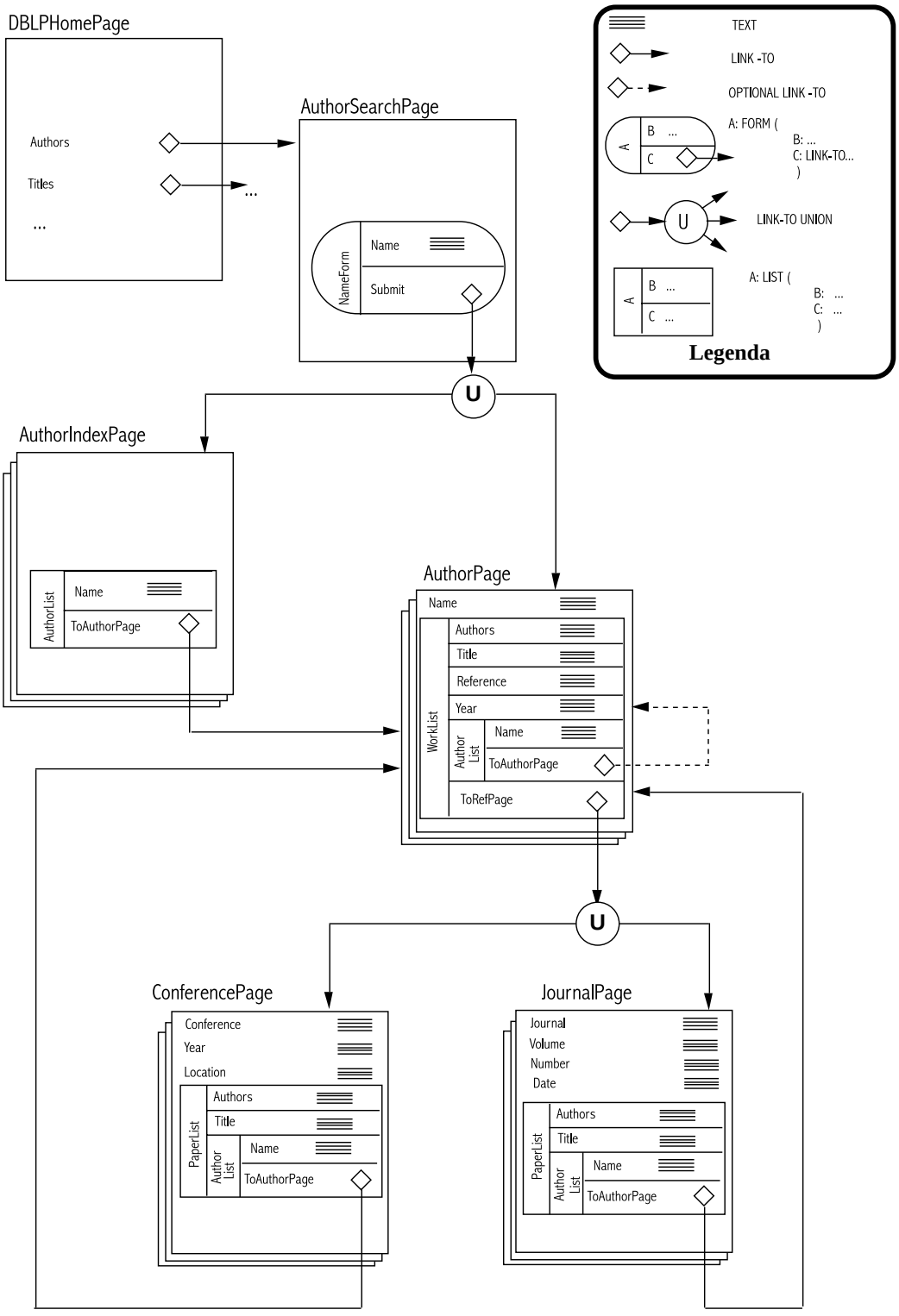

Figure 4: A portion of `DBLPScheme`, the ADM scheme for the the DB & LP Bibliography Server

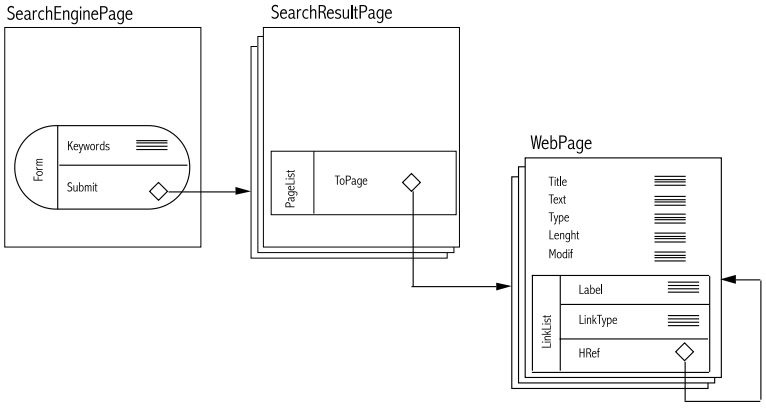

Figure 5: `WebScheme`, the ADM scheme for the World Wide Web

the root of a subtree; optional attributes may have a null value.

Note that ADM generalizes the data models used in other Web query languages [24, 25, 28]. For example, in WebSQL [28], the Web is seen as a graph of documents connected by links; each document has a fixed set of attributes: the URL, the title, the document text, the type, ecc. The whole Web is described using a simple relational scheme with two (virtual) relations, as follows:

```
Document(url, title, text, type, length, modif)
Anchor(base, href, label)
```

Each tuple in relation `Document` describes a single page, whereas a tuple in relation `Anchor` corresponds to a link from page with URL `base` to page with URL `href`, with the associated label. Documents can be accessed in two ways: *(i)* either by specifying their URL or *(ii)* by specifying a set of keywords; in this case, one or more search engines like, for example [1], are contacted in order to find the URL of all documents mentioning the specified keywords; then, these URLs are used to start the navigation. A distinctive feature of the language is that it allows to distinguish between *internal* and *external* links; a link is said to be internal when the two pages are inside the same physical server; it is external otherwise.

We can easily give a similar scheme of the whole Web using ADM. The scheme, called `WebScheme`, is reported in Figure 5: it contains three page-schemes, as follows. `WebPage` is used to model a generic Web page; for each page, attributes `URL, title, text, type, length` and `modif`, all of type `TEXT` are reported; attribute `LinkList` models the list of all links in the page; for each link, the label is reported, plus the link type, i.e. internal or external. It is worth noting that all of these pieces of information can be easily extracted from any page, with little effort. Two more page-schemes are used to model the presence of the search engine(s): `SearchEnginePage` models the search-engine form as a (virtual) list of tuples associating the search-result page with the specified keywords; `SearchResultPage` describes pages produced as a result of the search: each of these pages contains a list of URLs, namely those corresponding to documents containing the specified keywords.

Based on this scheme, in the following section we discuss how it is possible to ask very general queries, such as "*retrieve all documents in the Web mentioning Java*", using ULIXES. This shows

that our approach is highly scalable: ADM allows to model the organization of Web pages at different levels of granularity, going from very general and unstructured representations, like in `WebScheme`, to rather tight and structured representations for specific portions, like for example `DBLPScheme` in Figure 4.

4 ULIXES: Defining Relational Views over the Web

In this section, we present ULIXES, a language for the definition of relational views over the Web. ULIXES has been designed to be a simple and flexible language for extracting data from the Web based on an ADM scheme. The data extraction process is based on the notion of *navigation* in the site. Navigations in ULIXES are expressed using *navigational expressions*, i.e. path expressions [22] denoting paths in the site graph. In our perspective, a site offers in essence a set of navigations through pages in the site; these navigations allow to follow links between different pages, but also to explore the hierarchical structure of a page: they represent a natural means to query the page. Consider for example the scheme in Figure 4. Suppose we are interested in reaching all author pages in the site. To do this, we can start from the author search page (`AuthorSearchPage`), whose URL is known, submit an empty form, and reach the search result page; this is an index of all authors in the site, that is an instance of page-scheme `AuthorIndexPage`; for each author, the corresponding page can be reached by following the associate link. These navigations in the site can be specified using the following *navigational expression*, in which the dot operator (.) denotes navigations inside pages, and the *link operator* ($\rightarrow$) is used to follow links:

`AuthorSearchPage.NameForm.Submit`$\rightarrow$`AuthorIndexPage.AuthorList.ToAuthorPage`$\rightarrow$`AuthorPage`

The semantics of the expression can be interpreted as all possible paths in the site obtained by starting from the unique instance of page-scheme `AuthorSearchPage`, submitting an empty form, traversing the `AuthorIndexPage` and then reaching an `AuthorPage`. Each of these navigations can be represented as a tuple of values, one value for each monovalued attribute associated to nodes in the navigation; thus, each navigational expression can be represented as a relation, in the relational model sense. Based on these ideas, we can associate a relation, i.e. a set of tuples to each navigational expression. Given a navigational expression, N, we call SEM(N) the corresponding relation. We assume that attributes are suitably renamed whenever needed.

Given the relational nature of navigations, the definition of relational views over ADM schemes can be directly based on navigational expressions. We have a `DEFINE TABLE` statement to be used for this purpose, with the form:

`DEFINE TABLE`	$R(B_1, B_2, \ldots, B_n)$
`AS`	N
`IN`	S
`USING`	$A_1, A_2, \ldots, A_n$
`WHERE`	$c_1, c_2, \ldots c_k$

where: *(i)* R is a relation name and $B_1, B_2, \ldots, B_n$ are attributes; *(ii)* S is an ADM scheme; *(iii)* N is a navigational expression over S; *(iv)* $A_1, A_2, \ldots, A_n$ are attributes of SEM(N); and *(v)* $c_1, c_2, \ldots, c_k$ are a set of conditions over the attribute values. The semantics of a `DEFINE TABLE` can be defined on the basis of previous notions: relation R is the projection onto $A_1, A_2, \ldots, A_n$ of the selection of SEM(N) with respect to the conjuction $c_1, c_2, \ldots, c_k$, with each A_i renamed to B_i, that is:

$$R = \rho_{B_1 \leftarrow A_1, \ldots, B_n \leftarrow A_n}(\pi_{A_1, A_2, \ldots, A_n}(\sigma_{c_1 \wedge c_2 \wedge \ldots \wedge c_k} \mathrm{SEM}(N)))$$

In the current implementation, we see each page as a *nested relation* [11, 14], in which list attributes are modeled using tables. Due to the absence of duplicates, our relations can be decomposed in flat relations,[5] and the semantics of navigational expressions can be easily defined using joins. To do this, for each page in a navigation, we generate the associated table, and then join them using a local SQL engine.

As an example, suppose we are interested in the authors, titles and references for all papers by Leonardo da Vinci in VLDB conferences. We can generate such a relation using the following statement:

```
DEFINE TABLE VLDBPapers (Authors, Title, Reference)
AS           AuthorSearchPage.NameForm.Submit → AuthorPage.WorkList
IN           DBLPScheme
USING        AuthorPage.WorkList.Authors, AuthorPage.WorkList.Title,
             AuthorPage.WorkList.Reference
WHERE        AuthorSearchPage.NameForm.Name = 'Leonardo Da Vinci',
             AuthorPage.WorkList.Reference LIKE '%VLDB%'
```

In this expression, we are essentially giving a name, `VLDBPapers`, to a relation corresponding to the navigational expression `AuthorSearchPage.NameForm.Submit → AuthorPage.WorkList`; the resulting relation, reported in Figure 6, includes only a subset of attributes, namely those listed in the `USING` clause: the authors, title, and reference for each of the selected papers. There are several things to note with respect to this example; first, during the navigation, data are filtered using conditions in the `WHERE` clause, so that only paths ending with a paper in a VLDB conference are considered. Moreover, the `WHERE` clause is also used to fill-in the form in a completely transparent way by specifying that the name attribute of the form must be equal to 'Leonardo da Vinci'. Note also that in the navigational expression we require that the `Submit` link returns a page over scheme `AuthorPage`, that is, we are selecting one of the possibilities in the union type. The language has specific mechanisms for dealing with union types, so that, if the page returned is of the correct type, it navigates it; otherwise, it simply returns an empty result.

Consider now a more complex example: suppose we are now interested in finding all conferences in which both Leonardo and Michelangelo published at least one paper. In this case, starting from the `AuthorPage` of Leonardo Da Vinci, we need to inspect all papers published at conferences, reach the corresponding `ConferencePage` and navigate it to find papers whose author is Michelangelo. This can be done as follows:

[5] This is due to the fact that we suppose nested structures to be in *Partitioned Normal Form (PNF)* [29]

Authors	Title	Reference
M. Buonarroti, Leonardo Da Vinci	The Man in the Circle	VLDB 1486: 73-90
Leonardo Da Vinci, M. Buonarroti	Algorithms for Checking Numbers in Databases	VLDB 1487: 150-167
N. Copernico, Leonardo Da Vinci, G. Galilei	Gas Dynamics and Possible Flying Vehicles	VLDB 1489: 159-166
N. Copernico, Leonardo Da Vinci, G. Galilei	The Virgin of the Rocks: Study of the Use of Light in Paintings	VLDB 1488: 159-166

Figure 6: The result of a `DEFINE TABLE` statement

```
DEFINE TABLE Conference (Name, Year)
AS           AuthorSearchPage.NameForm.Submit → AuthorPage.WorkList.ToRefPage →
                 ConferencePage.PaperList.AuthorList
IN           DBLPScheme
USING        ConferencePage.Conference,
             ConferencePage.Year
WHERE        AuthorSearchPage.NameForm.Name = 'Leonardo Da Vinci',
             ConferencePage.PaperList.AuthorList.Name = 'M. Buonarroti'
```

In this case, we start from the form, access Leonardo Da Vinci's page, and navigate to reach all conference pages; in the `ConferencePage`, we inspect all papers in the `PaperList`, and, for each paper, the name of all authors in the `AuthorList`. Using the `WHERE` clause we select only those paths reaching conferences in which there are papers by Michelangelo.

Note that ULIXES can be used to ask more general queries, involving the whole Web. Consider for example the following query from [28]: "*find all links to applets from documents mentioning Java*"; based on the Web scheme in Figure 5, the query can be expressed in ULIXES as follows:

```
DEFINE TABLE JavaApplets(PageURL, AppletURL)
AS           SearchEnginePage.Form.Submit → SearchResultPage.PageList.ToPage →
                 WebPage.LinkList
IN           WebScheme
USING        WebPage.URL,
             WebPage.LinkList.HRef
WHERE        SearchEnginePage.Form.Keywords = 'Java',
             WebPage.LinkList.Label LIKE '%applet%'
```

In this case, we use the search engine to find all Web pages mentioning Java; then, we inspect the links in these pages, and choose the ones pointing to applets.

We would like to emphasize the effectiveness and flexibility of the chosen approach. It is effective, in the sense that it provides a high-level tool for extracting data; note that computing the shown queries only by means of browsing would require a significant effort for the user. At the same time, the approach is flexible since, once a relational view has been defined and a table has been generated, any high-level query language – relational or object-oriented – can be used to access data, provided that it can manipulate tables. These relational views can then be integrated with other data sources, local data sets or also views over different sites. As an example, suppose we generate a large table containing the references of all authors in the DB & LP Bibliography

Authors	Title	Reference	Year	ToRefPage
Leonardo Da Vinci, N. Copernico, G. Galilei	The Hammer Code: A Study on the Nature of Water	PODS 1485: 200-215	1486	http:\\... /pods85.html
...	...	...	...	...
N. Copernico, Leonardo Da Vinci, G. Galilei	Gas Dynamics and Possible Flying Vehicles	VLDB 1489: 159-166	1489	http:\\... /vldb89.html
Leonardo Da Vinci, N. Copernico	Querying and Updating Relational Databases	PODS 1489: 101-109	1489	http:\\... /pods89.html
Leonardo Da Vinci, G. Galilei	On the Nature of Earth, Wind and Fire	TODS 34(2): 212-230 (1489)	1489	http:\\... /tods34.html
...	...	...	...	...
M. Buonarroti, N. Copernico, Leonardo Da Vinci, G. Galilei	Monalisa and the Management of Multimedia Data in Databases	TCS 75(3): 291-308 (1490) (1490)	1490	http:\\... /tcs75.html

Figure 7: All papers by Leonardo with years of publication.

server [27]. We may think of applying the same process to other bibliography servers, like, for example [5], and then integrate the two views to obtain a larger set of references. Based on the integrated database, a new, derived site can be then generated using the language PENELOPE.

5 PENELOPE: Generating Derived Hypertexts

The approach discussed in the previous section is interesting but could be considered as extraneous to the Web framework, where users access information by navigating hypertexts. We thus would like to extend the view paradigm in such a way that, once data have been retrieved, they are presented to the user as a hypertext. Here, we show how relational views can be transformed back into Web hypertexts, whose pages have a structure that does not appear in the existing site(s). This restructuring technique can be used to define a derived site, that is, a hypertextual view over the input sites or over a database, but also as a support to casual queries, where the user wants to browse the results (this could be particularly useful with respect to complex queries with large results).

In order to reach this goal, we introduce PENELOPE, which allows the definition of new page-schemes according to which data will be organized. Let us illustrate the process by means of an example, again on [27]. Suppose we have used ULIXES to navigate the site and build the table `DaVinciPapers` in Figure 7, containing a tuple of the form *(Authors, Title, Reference, Year, ToRefPage)* for each paper by Leonardo Da Vinci. We want to re-organize papers, dividing them on the basis of the year of publication. To do this, for each year in the table, we generate a page containing the list of papers published by the author in that year; moreover, a unique page containing the list of all years is created to provide access to year pages. We need a *restructured scheme* as described in Figure 8; the two page-schemes are called `DaVinciYearsPage` and `YearPage`. The structure of the pages can be defined using the following `DEFINE PAGE` statements. Note that attributes of the source table `DaVinciPapers` are enclosed in angle brackets `<>`.

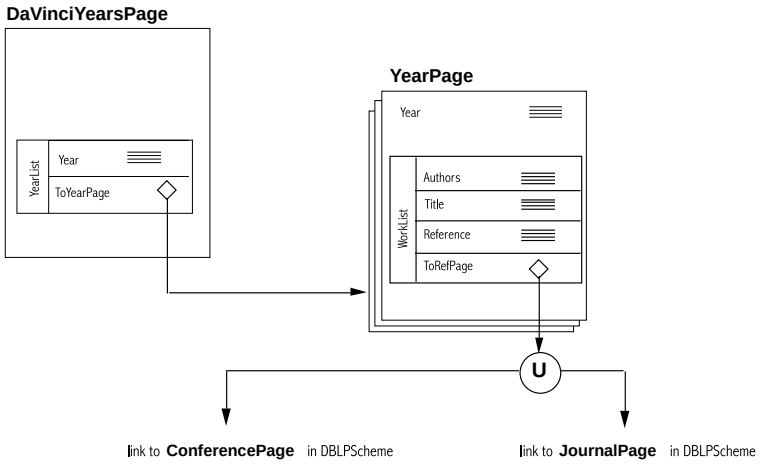

Figure 8: New page-schemes to organize papers by year

```
DEFINE PAGE YearPage
AS   URL      URL(<Year>);
     Year:         TEXT <Year>;
     WorkList:     LIST OF  (Authors:  TEXT    <Authors>;
                             Title:     TEXT    <Title>;
                             Reference: TEXT    <Reference>;
                             ToRefPage: LINK TO ConferencePage UNION JournalPage <ToRefPage>);
FROM DaVinciPapers

DEFINE PAGE DaVinciYearsPage UNIQUE
AS   URL 'result.html';
     YearList:     LIST OF  (Year:       TEXT    <Year>;
                             ToYearPage:  LINK TO YearPage    (URL(<Year>)));
FROM DaVinciPapers
```

These statements generate the HTML code for the new pages. The first statement defines `YearPage` as a page-scheme with a monovalued attribute, the year, and a multivalued attribute corresponding to the list of papers by Leonardo published in that year. A page will be created for each different year; clearly, URLs for these pages have to be generated by the system, and each time a page is created, a new, different URL is needed. We use function terms to generate URLs; in fact, term `URL(<Year>)` specifies that the system has to generate an URL for each page over scheme `YearPage`, and that the URL must be uniquely associated with the year value.[6] The `DEFINE PAGE` statement also describes how pages must be filled-in starting from attributes in the relation. For example, the definition of attribute `Year` of type `TEXT` in page-scheme `YearPage`, specifies that its values come from attribute *Year* of relation `DaVinciPapers`.

The second statement defines `DaVinciYearsPage` as a unique page-scheme with a multivalued attribute `YearList`, corresponding to the list of years; note how, in this case, a local, constant URL, `result.html`, is assigned to the corresponding instance. Since we declare the page-scheme as unique and indicate a single URL, we are assuming that a unique page will be generated by the statement. In the page, each item in the year list must be linked to the corresponding year page;

[6] This technique is somehow similar to the use of *Skolem functors* to invent new OID's in object-oriented databases [20].

to do this, we use as a value for the link the same function term used to generate URLs of years pages, i.e. `URL(<Year>)`.

Note that, for each paper in `YearPage`, we also want to access the corresponding conference or journal page. However, we do not re-create these pages in the derived hypertext, but instead we access those in the original site. To link these pages, we use URLs of existing pages, i.e. values of attribute `ToRefPage` in the table, which have been extracted using ULIXES. In essence, by using the two languages together, we generate a hypertext made of some new pages, offering a different perspective over data, plus an existing portion of the original site.

`DEFINE PAGE` statements are based on these ideas. *Local URLs* are used to identify new pages; they can be either constant strings, or strings built using the function symbol `URL` from attributes in relations. For example `result.html` is a constant local URL, whereas `URL(<Year>)` denotes a local URL built from values of attribute `Year`. Thus, a `DEFINE PAGE` statement has the form:

```
DEFINE PAGE  P [UNIQUE]
AS           S
FROM         R
```

where: *(i)* P is a new page-scheme name; *(ii)* R is a relation; and *(iii)* S describes the page structure, by specifying the page attributes, their type, and their correspondence with attributes of R. The `UNIQUE` keyword is optional; it is used to specify that the defined page-scheme is unique.[7]

The semantics of these statements can be informally defined as follows:

- relation R is extended by adding local URLs to each tuple as new attributes; first, the page URL is generated and added to the relation as a new attribute, `URL`; then, for any structure of the form LINK TO P (*LocalURL*), an attribute `ToP` is added to the relation; for each of these attributes, if the function term `URL()` has been used, a file name is generated based on the value of the specified attribute; we also suppose that attributes are renamed according to the page structure, if needed; consider the statement in the previous example, defining page-scheme `DaVinciYearsPage` starting from relation `DaVinciPapers` reported in Figure 7. Adding local URLs yields table *(a)* in Figure 9; in the table, URLs to link year pages have been generated by the systems from values of the `Year` attribute;

- this relation is then projected onto the attributes occurring in the page structure, S, plus the URL; since only the year and the link to `YearPage` are reported in the `DEFINE TABLE` statement, table *(b)* in Figure 9 is produced;

- next, the projected relation is *nested* [29] according to structures in S; more specifically, for each multivalued structure in S, the relation is nested on the corresponding attributes; with respect to the example, a new attribute `YearList` is introduced by nesting with respect to `Year` and `ToYearPage`; the result is shown is table *(c)* in Figure 9;

[7] To avoid inconsistencies, we have to impose some constraints for unique page-schemes; for example, the corresponding local URL must be a constant string; second, the associated page structure must contain a single attribute, which has to be multivalued.

URL	Authors	Title	Reference	Year	ToRefPage	ToYearPage
result.html	Leonardo Da Vinci, N. Copernico, ...	The Hammer Code: A Study on the Nature ...	PODS 1485: 200-215	1485	http:\\... /pods85.html	1485.html
...	...	...	...	...	...	...
result.html	N. Copernico, ... Leonardo Da Vinci, ...	Gas Dynamics and Possible ...	VLDB 1489: 159-166	1489	http:\\... /vldb89.html	1489.html
result.html	Leonardo Da Vinci, N. Copernico	Querying and Updating Relational Databases	PODS 1489: 101-109	1489	http:\\... /pods89.html	1489.html
...	...	...	...	...	...	...
result.html	M. Buonarroti, ... Leonardo Da Vinci, ...	Monalisa and the Management ...	TCS 75(3): 291-308 (1490)	1490	http:\\... /tcs75.html	1490.html

(a)

URL	Year	ToYearPage
result.html	1485	1485.html
...	...	...
result.html	1489	1489.html
result.html	1490	1490.html

(b)

URL	YearList	
	Year	ToYearPage
result.html	1485	1485.html
	...	...
	1489	1489.html
	1490	1490.html

(c)

Figure 9: Intermediate steps to generate derived pages

- finally, a local page is generated for each tuple in the resulting relation. Nested attributes are translated using HTML list environments. Note that, coherently with the fact that the page-scheme has been defined as being unique, the nested relation for `DaVinciYearsPage` contains a single tuple, and thus a single page is generated. The page is reported in Figure 10.

The same algorithm can be used to generate the pages for `YearPage`. In this case, the final nested table has several tuples, one for each year, so that different pages are generated. See Figure 11 for an example. Note that the ADM schema describing the new hypertext can be easily produced starting from the structures specified in `DEFINE PAGE` statements.

We have done several experiments using PENELOPE to integrate different sites from the same domain. The system effectively allows to generate an integrated view over the existing sites. Moreover, it is possible to choose to which extent the derived hypertext is to be locally materialized. This has proven very useful when dealing with multimedia data; as an example, we have generated an integrated view over the Louvre [4] and the Uffizi [7] virtual museums; in this new site, an integrated list of works is reported for each artist, and an image of the work is accessible. In order to reduce the disk storage, images have not been downloaded, and they are accessed on the original sites. In this way, integrated Web views can be generated by re-using portions of existing sites.

Acknowledgments

We would like to thank Elena Tabet, Alessandro Masci and Salvatore Labonia for stimulating discussions on the subject of the paper and for their contributions to the development of the prototype. Thanks also go to Stephane Grumbach, who provided interesting comments on an early draft of the paper.

Figure 10: Unique instance of page-scheme `DaVinciYearList`

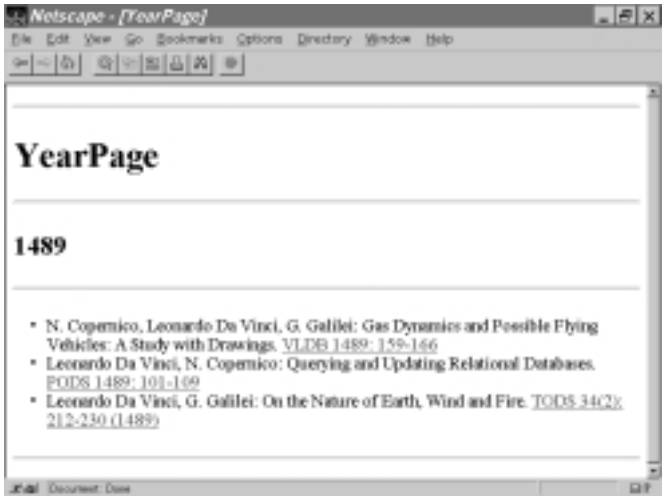

Figure 11: One instance of page-scheme `YearPage`

References

[1] AltaVista Technology Home Page. `http://www.altavista.com`.

[2] The ARANEUS System Home Page. `http://poincare.inf.uniroma3.it:8080/Araneus`.

[3] Informix Home Page. `http://www.informix.com`.

[4] The Louvre Web server. `http://www.louvre.fr`.

[5] The Max Planck Institute virtual library. `http://www.mpi-sb.mpg.de/guide/staff/hustadt/-library.html`.

[6] Oracle Home Page. `http://www.oracle.com`.

[7] The Uffizi Web server. `http://www.uffizi.firenze.it`.

[8] S. Abiteboul. Querying semi-structured data. In *Sixth International Conference on Data Base Theory, (ICDT'97), Delphi (Greece), Lecture Notes in Computer Science*, 1997.

[9] S. Abiteboul, S. Cluet, V. Christophides, T. Milo, G. Moerkotte, and J. Siméon. Querying documents in object databases. *Journal of Digital Libraries*, 1997. To Appear. `http://www-db.stanford.edu`.

[10] S. Abiteboul, S. Cluet, and T. Milo. Querying and updating the file. In *International Conf. on Very Large Data Bases (VLDB'93), Dublin*, pages 73–84, 1993.

[11] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of databases.* Addison Wesley Publ. Co., Reading, Massachussetts, 1994.

[12] S. Abiteboul, D. Quass, J. McHugh, J. Widom, and J. Wiener. The Lorel query language for semistructured data. `http://www-db.stanford.edu`, 1996.

[13] P. M. G. Apers. Identifying internet-related database research. In *Second International East-West Database Workshop, Klagenfurt, Workshops in Computing*, pages 183–193. Springer-Verlag, 1994.

[14] P. Atzeni and V. De Antonellis. *Relational Database Theory: A Comprehensive Introduction.* Benjamin and Cummings Publ. Co., Menlo Park, California, 1993.

[15] P. Atzeni and G. Mecca. Cut and Paste. In *Sixteenth ACM SIGMOD Intern. Symposium on Principles of Database Systems (PODS'97), Tucson, Arizona*, 1997. To Appear. `http://poincare.inf.uniroma3.it:8080/Araneus/publications.html`.

[16] T. Berners-Lee, R. Cailliau, A. Lautonen, H. F. Nielsen, and A. Secret. The World Wide Web. *Communications of the ACM*, 37(8):76–82, August 1994.

[17] P. Buneman, S. Davidson, G. Hillebrand, and D. Suciu. A query language and optimization techniques for unstructured data. In *ACM SIGMOD International Conf. on Management of Data (SIGMOD'96), Montreal, Canada*, pages 505–516, 1996.

[18] R. G. G. Cattel. *The Object Database Standard ODMG-93.* Morgan Kaufmann Publishers, San Francisco, CA, 1994.

[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] R. Hull and M. Yoshikawa. ILOG: Declarative creation and manipulation of object identifiers. In *Sixteenth International Conference on Very Large Data Bases, Brisbane (VLDB'90)*, pages 455–468, 1990.

[21] R.B. Hull and R. King. Semantic database modelling: Survey, applications and research issues. *ACM Computing Surveys*, 19(3):201–260, September 1987.

[22] M. Kifer, W. Kim, and Y. Sagiv. Querying object-oriented databases. In *ACM SIGMOD International Conf. on Management of Data*, pages 393–402, 1992.

[23] W. Kim, editor. *Modern Database Systems: the Object Model, Interoperability, and Beyond.* ACM Press and Addison Wesley, 1995.

[24] D. Konopnicki and O. Shmueli. W3QS: A query system for the world-wide web. In *International Conf. on Very Large Data Bases (VLDB'95), Zurich*, pages 54–65, 1995.

[25] L. Lakshmanan, F. Sadri, and I. N. Subramanian. A declarative language for querying and restructuring the Web. In *6th Intern. Workshop on Research Issues in Data Engineering: Interoperability of Nontraditional Database Systems (RIDE-NDS'96)*, 1996.

[26] A. Y. Levy, A. Rajaraman, and J. J. Ordille. Querying heterogeneous information sources using source descriptions. In *International Conf. on Very Large Data Bases (VLDB'96), Mumbai(Bombay)*, 1996.

[27] M. Ley. Database systems and logic programming bibliography server. `http://www.informatik.uni-trier.de/~ley/db/index.html`.

[28] A. Mendelzon, G. Mihaila, and T. Milo. Querying the World Wide Web. In *First Int. Conf. on Parallel and Distributed Information Systems (PDIS'96)*, 1996. `ftp://db.toronto.edu/pub/papers/websql.ps`.

[29] M.A. Roth, H.F. Korth, and A. Silberschatz. Extended algebra and calculus for ¬1NF relational databases. *ACM Transactions on Database Systems*, 13(4):389–417, December 1988.

[30] 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.