The *Ecumene* Experience to Data Integration in Cultural Heritage Web Information Systems

Alessio Bechini[1], Andrea Tomasi[1], and Gianni Ceccarelli[2]

[1] University of Pisa - Dept. of Information Engineering, via Diotisalvi 2
I-56126 Pisa, Italy
{a.bechini, a.tomasi}@ing.unipi.it
[2] Hyperborea scrl,
Navacchio (PI) Italy

Abstract. A number of diverse data sources on cultural/artistic heritage, hosted by local institutions, may provide a rich collection of study material for experts in this field. Anyway, a wide exploitation of such a huge amount of information is hampered by the heterogeneity and the physical distribution of data sources. The experience presented here relates to problems and solutions found in making use of this kind of data sources within a single Web Information System that offers a framework for accessing the artistic/cultural information. In particular, design issues in dealing with metadata are discussed, as well as architectural aspects from the metadata harvesting to the presentation tier. All the system leverages the web infrastructure to cope with both organizational and technological diversity among the composing modules.

1 Introduction

An effective dissemination of knowledge about Cultural Heritage can be achieved through the exploitation of modern Web Information Systems (WIS), able to expose large amount of data in a well-structured, accurate way to an ever-increasing community of users. The richness of information in this particular field is settled on by a large number of data repositories managed by different kinds of institutions operating on restricted geographical areas. The physical distribution and the heterogeneity of local data sources have so far hampered a rational utilization of all the potentially available information in the cultural/artistic heritage field. In this paper we present the design of the data-handling portion of a specific WIS, aimed at providing tools for the integration, organization, and search of Cultural Heritage information mainly from the Italian Catholic Church, spread over local data repositories. The services offered by the WIS are oriented to the collection of a consistent view of the different objects involved in a particular search.

In recent years, several experiences in the field of cultural heritage digital libraries have been carried out (among the others, the Perseus Project at Tufts University [6] represents a long-term study). In these occasions, the needs and possibilities facing cultural heritage digital libraries have been pointed out: E.g., people studying ancient world usually work on small, fragmentary data sets; on the other hand, a flood of

documentation may cause serious problems in the comprehension of facts happened in the recent past. Extrapolation and reduction are complementary processes that play crucial roles in the humanities [6]. Beyond these considerations, the starting point for any kind of analysis is the availability of data and the possibility to retrieve the required information. These tasks cannot often be accomplished in a trivial way, especially whenever the desired data is owned by numerous organizations at different geographical locations. This is the case of the system presented in this paper, developed within the "Ecumene Project", and aimed at providing a framework to facilitate the access to different kinds of data in the cultural/artistic heritage domain: the involved institutions are the dioceses of the Italian Catholic Church. Thus we are dealing with a huge amount of data, spanning a very long period of time, and relating to any kind of artwork, architectural goods, and estates, as well as many associated historical archives.

It is clear that the data-intensive nature of this kind of information systems pushes the adoption of an accurate development process. A fundamental step is the design of the conceptual model for data-driven WIS: this methodological phase is often referred to as the definition of the *semantic layer*, emphasizing the importance of semantic interoperability in systems that gather information from heterogeneous sources [16].

The adoption of ad hoc standards for metadata representation (such as e.g. Dublin Core [7]) is an important step to increase the chances for a data source to interoperate with other databases/applications. Anyway, it has been clearly pointed out that integration cannot be achieved entirely through standardization [15]. A more flexible approach can rely on the design (possibly model-driven) of the semantic layer. In particular, the definition of a *mediated schema* (or *global schema*, i.e. a purely logical schema for the purpose of query issuing) is the central activity for the data integration within the web information system. Although it is widely known problem in data integration [9], the heterogeneous structures of the data about artistic/cultural heritage make it particularly challenging. The interoperation of the autonomous data sources is provided by a semantic mapping between the schemas of the different sources.

Often, a convenient way to describe mappings between the mediated schema and the local schemas is the employment of proper configuration documents, that can be produced with the support of graphical tools: this is the case e.g. of DDXMI [13], that follows this approach in a simplified scenario. It should be pointed out that general query processing in a data integration system is never a simple task, as serious difficulties arise even with a very simple mediated schema, as integrity constraints are enforced [4].

Several approaches have been used in practical implementation of metadata integration within information systems: for a brief but comprehensive overview, refer to the related works section of [13].

The main obstacles in the integration process are the following:

- The heterogeneity of the involved information systems that are in charge of providing raw data;
- The assortment of structures for the involved data (e.g. archives, information on architectural assets, descriptions of artworks, etc.).

In the Ecumene project these problems have been overcome *from the technological standpoint* by choosing the following strategies:

- To leverage the web infrastructure for the implementation of the communication facilities within the whole information system;
- To make use of XML as a sort of unifying lingua franca in dealing with a wide variety of data sets, *and of XML-related technologies* to manage metadata representations.

These are the two main guidelines adopted throughout the implementation process. Although the web has been recognized as an appropriate data transport channel, it is important to select the level of abstraction to use in the communication framework. Moreover, different levels may be appropriate for interconnecting different application blocks. Typically, such a decision is driven by requirements about application flexibility (leading to increasing the inter-block decoupling), performance, and scalability.

The following sections describe the architectural structure resulting from the Ecumene design process. Particular attention is paid to the role played by the metadata management, which practically summarizes the main aspects of the data integration approach. Finally, conclusions are properly drawn.

2 An architectural view of the overall system

The separation of roles is one of the main concerns in WIS design, and Ecumene follows this approach. The system is structured in several different functional blocks, as we can see in Fig. 1:

- A number of assorted local sites, containing the raw data sources;
- An *access server*, i.e. a block dedicated to the data management, and which is thus able to export data access/search services towards other blocks;
- A *front-end web application*, which represent the interface employed by users to interact with the information system.

The web infrastructure is used throughout the whole system to support the integration of different, separated hosts at the data transport level.

Our basic idea is to strongly separate the navigational issues from the data organization. The main risk in pursuing a navigation-driven organization of metadata is to loose some information that, although not asked for in a basic utilization of the system, could be proficiently pointed out in a more sophisticated interface dedicated to expert end-users. Moreover, the derived structure of the integrated data, their internal relationships, and the associated context information have not to be driven by the architecture of the presentation tier, but instead they are worth being dealt with regardless of specific requirements from the web user interface. Thus, the main

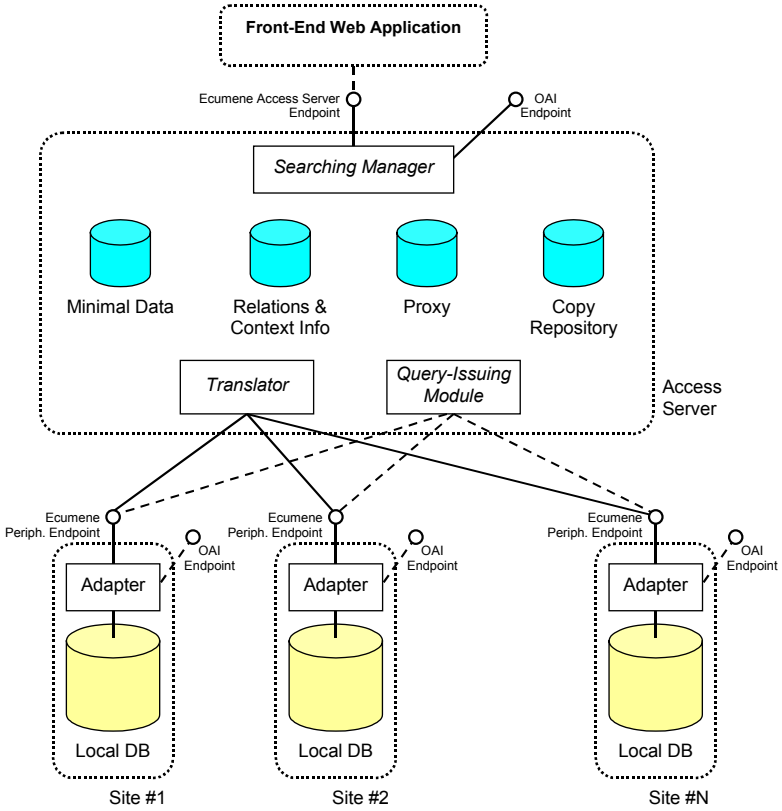

Fig. 1. Outline of the main architectural features of the Ecumene system. Dotted rectangles represent system blocks, and they contain modules (regular rectangles) supporting specific internal functionality. The lollypop symbols represent communication endpoints for accessing services provided by the system blocks. The Ecumene Peripheral Endpoints usually receive incoming requests from the *query-issuing module*, and reply back to the *translator*.

component of the overall system is the *Access Server* block. It clear that the services that it provides must be accurate enough to fit a wide rage of data requests.

The basic functional blocks are indicated in Fig. 1 by dotted rectangles. Each block is an independent logic unit, which exposes *endpoints* for accepting incoming service requests. Each endpoint is associated with a precise schema (or set of schemas) for the exchanged data. The Access Server presents a main endpoint, able to deliver the withdrawn data with the richest information content. The front-end web application is the client this endpoint has been designed for. Otherwise, it may be particularly important to provide another way, likely a standardized one, to export data for integration with other applications: for this reason the Ecumene Access Server supports an endpoint which is OAI-compliant [14] (OAI stands for Open Archive Initiative). Besides, the OAI endpoint cannot be able to give access to the whole information capability of Ecumene, as the protocol OAI-PMH (i.e. OAI Protocol for

Metadata Harvesting) has not been designed to fit the needs of the cultural heritage data.

Within each block, several modules take care of carrying out the supporting activities. The Access Server encompasses three crucial ones: the *searching manager*, the *query-issuing module,* and the *translator*; other minor helper components are presents, but do not deserve particular attention in our architectural description of the system.

The new developed system cannot be considered as a mere tool for accessing data spread over a number of different places. The system should help the end user to take full advantage of the accessed data, giving what required understanding the *context* for data, to deliver trustable, un-ambiguous data, and to relate different pieces of information within a unique picture. All these requirements can be fulfilled only by adding new capabilities to the system, which must thus rely on an *internal database* with proper ancillary information. Such a database is logically placed within the access server block, and it can participate in the data-searching task.

The purpose of the *searching manager* is to provide an interface for requesting information from the access server. The searching manager, once a query is posed over the mediated schema, is in charge of answering making use of data stored both within the access server, and within the local sources. A specific problem addressed at this level is the reformulation of the incoming queries into queries that refers to the schema of the local sources as well [9]. The searching manager is responsible, of course, for providing and handling both the Ecumene Access Server Endpoint, and the OAI endpoint; thus, it is in charge also of any kind of authentication procedures and security issues in the communication with different categories of users.

The *query-issuing module* is designed to pass sub-queries generated by the searching manager to the peripheral databases. This task can be accomplished only if complete information on each local database is known. Otherwise, a support on the specific peripheral site is required.

The *translator* is aimed at getting data from the external sources, and at re-organizing them according to the Ecumene mediated schema. Whenever the incoming data is very structured, this functionality is quite simple, as it can be modeled by a mapping function between elements (fields) belonging to the two different schemas (notice that a single mapping relation may involve several elements, as practically shown in [13]. The outcome from the translator is the following set of information:

- All the available raw data, organized according to the mediated schema, about specific objects belonging to the cultural heritage domain;
- Minimal data to be used for searching purposes;
- Context information;
- Relations with other data.

The adjustment of local data upon the mediated schema may be achieved in one or more steps, corresponding to the introduction of one or more modules in the system architecture. Such an adjustment is logically related to the data transportation from the original repository to the Access Server block. The Ecumene system takes advantage of several local data sources, and a large number of them belong to the

system itself: thus, in designing the complete WIS, we can suppose to operate also on most of the local sites (i.e. the several data repositories of the Italian Church dioceses). In this context, the introduction of an *adapter* module for supporting data transmission from each local source may be regarded as a viable solution to:

- Make easier (and more flexible) the data exchange;
- Possibly provide further access endpoints (possibly standard ones, e.g. OAI-compliant [14]) for the local data source towards external client applications.

The adapter provides a specific endpoint for the communication with the access server; the local data are presented at the endpoint according to a given schema, used for the data transportation. Using this architectural arrangement, the adapter becomes the only module in the whole system that has to care about the local representation/schema of peripheral data. The practical integration of a new kind of local site in the system requires the development of another particular adapter. The owner of the local site can decide about the adoption of a possible OAI endpoint integrated in the adapter (other standard endpoints are conceivable as well).

A particularly important choice in the system implementation involves the technologies for supporting data communications between block endpoints. It is crucial to identify the system requirements related to the communication features. In the specific case of the Ecumene project, interoperability among different blocks is the prime need, especially between the access server and the local databases: in fact, the system must be open to include any kind of data source that is able to supply a valuable information content. Although high performance is not theoretically a concern in Ecumene, the system response time towards end users must be quick enough to guarantee an acceptable level of usability. Finally, scalability issues must be taken into account: performance should no be severely degraded whenever an increase is experienced in the number of end users, of integrated peripheral sites and in the volume of accessed data. In distributed information systems, communication overhead may give a considerable contribution to the latency in accessing a service [2], and this mostly applies to high-level middleware solutions. The autonomous nature of the composing blocks of Ecumene (and their physical distribution) is certainly suitable to the application of asynchronous solutions, e.g. based on mobile agents [3]. Unfortunately, most of them are currently too heavyweight, and require a supporting platform on the involved host machines: thus, a more lightweight and seamless communication framework is preferable. JMS (Java Messaging Service) can be a possible choice for asynchronous messaging, and XML-RPC (XML Remote Procedure Call) can be viewed as a candidate support for data transmission. Different blocks in Ecumene present different coupling degrees: the front-end web application is much more coupled with the access server, than the access server with the peripheral sites. Thus in this particular settlement a lower-level communication middleware can be proficiently used (e.g. Corba or RMI, in Java environments).

Table 1. Different categories of data used by the Ecumene system, and accessed at local data sources.

Data Category	Description
OA	Artworks (paintings, sculptures, etc.)
A	Architectural assets
Historical Archives	Archival data
Digital Libraries	Typical data dealt with through OAI
Authority Lists	Authority list information

3 Dealing with Metadata in the Access Server

There exist various standards for data modeling and representation of artworks, historical buildings, archive's documents, books, and other different types of cultural heritage data items. Usually each standard description focuses on a particular category of items. In the Ecumene Project the approach to the generic item representations tries to exploit their semantic interconnections, and their description as well. The main categories of data items that our WIS is asked to deal with is briefly shown in Table I.

The mediated schema in Ecumene has been intentionally modeled as a subset of EAD [8][1]: this choice of making the Ecumene schemas fit into EAD has been inspired by the possibility to better share search results from Ecumene with other systems in different applications fields, as the archival ones. The richness of EAD schema has allowed us to practically setup this kind of schema modeling in a straightforward way.

Another important point in dealing with data and metadata at the access server level is to mark out a set of common information (*minimal metadata*) that all the data sources are required to provide. If a data source is not able to offer such a minimal set of information, its contribution cannot be proficiently exploited by the entire information system. The schema for minimal data has thus been designed as a sub-set of the complete Ecumene schema. Minimal data may be also of significant help to support scalability in the number of integrated sources. In fact, inside the access server the inspection of such data can be used to avoid issuing useless and time-consuming sub-queries to peripheral data sources. It is important to underline that in Ecumene the role of metadata is slightly different than in other approaches like Dublin Core [7] and Z39.50 [17]: in fact, metadata are not simply thought as a tool to retrieve actual data, but instead as a crucial step towards the integrated fruition of heterogeneous data within a common description framework.

During the Ecumene project advancement, we have experienced that a plain integration of descriptions pertaining to different topical areas (e.g. records about

[1] The EAD (Encoded Archival Description) Document Type Definition is a standard for encoding archival finding aids, which is maintained by the Library of Congress in partnership with the Society of American Archivists.

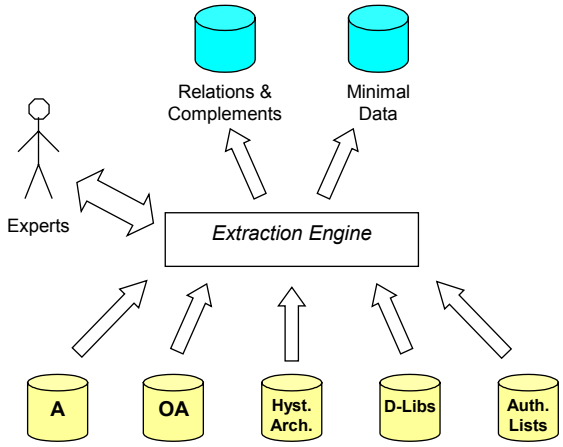

Fig. 2. The extraction process handled within the access server: data pertaining to different topical areas are gathered from peripheral sources; relations about the described data items are extracted and stored in an internal database, as well as context information. A core subset is also cached as *minimal data*, to be used in supporting searches.

artworks and historical data within archives) does not usually produce an information content that is understandable only by experts in that specific areas. We believe that one of the goals of a WIS like Ecumene is to better support the semantic integration of different descriptions, adding its own contribution to the clearness and precision of the search outcome. The prime way to do it is to acquire *context information* from the raw data at the peripheral databases, and to share it among all the data it is related to. This operation may often call for the contribution of experts in different fields, who can definitely validate context information and set up relations that cannot be extracted from the raw data. For the representation of context data, a possible way to be carried out is through the adoption of EAC-compliant schemas, which may be coupled with the sub-EAD schemas [8] employed for the different categories of the treated data. Context data refer to proper ancillary information about the involved places, institutions, families, etc. All this crucial data (as they are often the logical hooks to setup relations among data in different categories) should be validated through references to trustable sources: in other words, the creation of a sort of authority list, or the exploitation of already available ones, is practically mandatory in the presented scenario. A currently ongoing initiative for the investigation of these problems is the LEAF Project [11, 12], which explores the integration of authority files.

The operations performed within the Access Server to generate minimal data and context information are performed inside the Access Server by the *extraction engine* module. Fig. 2 illustrates this particular aspect; we should say that the extraction process is not carried out *only* during setup or maintenance activities, but can also be triggered by queries from the end users.

A query is issued to the Access Server mainly from the Ecumene end-point, and it has to be dealt with by the *Searching Manager*. The query result must be obtained from the joint contributions of data internally available in the Access Server (i.e. context information and relations), and data harvested from the local sources. The Access Server maintains four different databases to be used during search activities:

- A *copy repository*, to keep data belonging to sources that are not permanently connected via the web infrastructure;
- A *proxy*, i.e. a data cache holding data related to the most requested searches;
- A database with the *minimal data*, and
- A repository with *relations* among data items, and *context information*.

These repositories are the practical bases for the semantic integration of all the diverse data gathered at different distributed locations, presented with a coherent and clear organization.

4 Front-end Web Application

The *front-end web application* is the system block that is assumed to take advantage of the data integration provided by the access server block. It encapsulates both the *application layer* and the *presentation layer* as set down by a number of WIS design methodologies [1, 16]. The presentation layer is the portion of the front-end web application that is directly in contact with the system users.

In this paper we do not specifically focus on the architectural description of this block, although it represents a crucial component for the whole information system. The design of this block must take into account navigational aspects as well as adaptation issues [9], and at this level modeling languages that also deal with hypertext models may be of valuable help [5].

5 Conclusions

The integration of data from heterogeneous sources in the domain of cultural/artistic heritage is a complex problem. Another facet of this problem is the coherent organization of diverse gathered data, belonging to completely different application areas and topical fields: this is a crucial point for obtaining a proficient interaction of the end user with the WIS. The Ecumene project has tackled this kind of problems, operating on data repositories from dioceses of the Italian Catholic Church. The result of the design process for the whole system is an open architecture, composed of distributed loosely coupled blocks, whose core component is in charge of dealing with the data and metadata management, operating a logical separation from the navigational issues dealt with in the front-end web application block.

The compliance with existing standards has been pursued only in order to facilitate the information sharing with other possible distinct WIS applications: it has never to

be closely followed when it imposes constrains that hampers the exploitation of the information richness present in the raw data. Otherwise, the accessibility to the system functionality should be provided also through communication endpoints that support standards like OAI-PMH.

The main obstacles in the integration process have been overcome in the actual WIS implementation by choosing to leverage the web infrastructure for providing the required communication capabilities, and by making use of XML-related technologies to manage metadata representations. Such implementation decisions have given a contribution to meet performance and scalability requirements as well. The adoption of adapter modules at peripheral data sources has allowed the use of a common schema for data communication, simplifying the work for data integration within the access server block. Anyway, the possible future treatment of other kinds of semi-structured and unstructured data in the Ecumene system will certainly call for the introduction of state-of-the-art ontology-based techniques/languages, which have not been necessary up to this point in the core block of the described WIS.

The Ecumene Project is partially supported by the Parnaso initiative of the Italian MIUR, Minister for Education, University, and Research.

References

1. Barna, P., Frasincar, F., Houben, G.-J., and Vdovjak, R.: Methodologies for Web Information System Design. In proc. of ITCC'03, IEEE CS Press, 2003, 420-424
2. Bechini, A., Foglia, P., and Prete, C. A. : Use of a CORBA/RMI Gateway: Characterization of Communication Overhead. In Proc. of ACM 3rd Int'l Workshop on Software and Performance, July 2002, ACM Press 2002, 150-157
3. Bellavista, P., Corradi, A., and Tomasi, A.: The Mobile Agent Technology to Support and to Access Museum Information. In Proc. of ACM Symposium on Applied Computing 2000, ACM Press 200, 1006-1013
4. Calì, A., Calvanese, D., De Giacomo, G., and Lenzerini, M.: Data Integration under Integrity Constraints. In A. Banks Pidduck et al. (Eds.); CaiSE 2002, LNCS 2348, Springer-Verlag 2002, 262-279
5. Ceri, S., Fraternali, P., and Matera, M.: Conceptual Modeling of Data-Intensive Web Applications. IEEE Internet computing, 6(4) Jul-Aug 2002, IEEE CS Press, 2002, 20-30
6. Crane, G.: Cultural Heritage Digital Libraries: Needs and Components. In: M. Agosti and C. Thanos (Eds.): ECDL 2002, LNCS 2458, Springer-Verlag 2002, 626-637
7. Dublin Core Metadata Initiative: Metadata Terms Documentation, http://dublincore.org/documents/dcmi-terms/
8. EAD, Encoded Archival Description, http://www.loc.gov/ead/
9. Frasincar, F., Houben, G.-J., Barna, P., and Pau, C.: RDF/XML-based Automatic Generation of Adaptable Hypermedia Presentations. In Proc. of Int'l Conf. on Information Technology: Computers and Communications (ITCC'03)
10. Halevy, A. Y.: Data Integration: A Status Report. In Proc. of BTW 2003, LNI 26 GI 2003, 24-29
11. Kaiser, M., Lieder, H.-J., Majcen, K., and Vallant, H.: New Ways of Sharing and Using Authority Information. D-Lib Magazine, 9(11), nov 2003, ISSN 1082-9873

12. LEAF project (Linking and Exploring Authority Files), http://www.crxnet.com/leaf/index.html
13. Nam, Y.-K., Goguen, J., and Wang, G.: A Metadata Integration Assistant Generator for Heterogeneous Distributed Databases. In R. Meersman, Z. Tari (Eds.): CoopIS/DOA/ODBASE 2002, LNCS 2519, Springer-Verlag 2002, 1332-1344
14. Open Archive Initiative, http://www.openarchives.org/
15. Parent, C., and Spaccapietra, S.: Issues and Approaches of Database Integration. Communication of the ACM, 41 (5), 1998, 166-178
16. Vdovjak, R., and Houben, G.-J.: Providing the Semantic Layer for WIS Design. In Proc. of CaiSE 2002, LNCS 2348, Springer-Verlag 2002, 584-599
17. Z39.50 reference website at Library of Congress: http://www.loc.gov/z3950/agency/