Web Digital Archives Integrated Architecture

Carmen Costilla1, Juan P. Palacios2, M.ª José Rodriguez2, Raúl Fernández2, José Cremades3 and Antonio Calleja3

1 Titular Professor, Information Systems and Databases (SINBAD) Research Group
Dept. Ingeniería de Sistemas Telemáticos (DIT), ETS Ingenieros de Telecomunicación (ETSIIT),
Technical University of Madrid (UPM), www.sinbad.dit.upm.es, costilla@dit.upm.es

2 Ph. Researchers at the SINBAD-DIT-ETSIT-UPM
www.sinbad.dit.upm.es, jppalacios@sinbad.dit.upm.es, mjrodriguez@sinbad.dit.upm.es

3 Ph. Researchers at the SINBAD-DIT-ETSIT-UPM and
Founder Partners of CRC Information Technologies
(www.crcit.es), {rfernandez, jcremades, acalleja}@crcit.es

Abstract. The exponential growth of web published information, semantic heterogeneity of data sources and the desire of having proper and useful information accesses have arisen the challenge to offer a semantic web for this so diverse world. Integration of digital archives, coming from multiple independent sources, needs to deal with the hard topic of data sources heterogeneity. In order to solve this problem, this paper proposes a virtual and dynamic web architecture based on Mediator and Wrappers tiers. The Mediator includes an Ontological Kernel in order to provide a universal unification of concepts related to Web Digital Archive Integration.

Key words: Digital Archives (DA), Semantic Web, Ontologies, Information Integration on the Web, Integrated web architecture (Mediator & Wrappers).

1 Introduction

Currently, the web is an information space where daily are published a huge amount of documents with heterogeneous formats. This exponential growth of the present information on the web has taken place in an anarchic way, complicating search processes and useful information retrieval.

Web data management has become an important topic for the governments, libraries, museums, firms, etc. where proper information ubiquity is an intrinsic and main characteristic. For this reason, on-line analytical process and data mining require new ways for heterogeneous data sources integration, allowing an efficient information retrieval.

We have today new web technologies related to interoperability, such as XML and web services [Brit01]. However, web data integration has much to improve [DoMc03], because Web Information Systems (here after, WIS) are being still integrated by hand -with tedious processes and ad hoc works- that produces highly vulnerable results. Today, solving the semantic heterogeneity problem for the future web intelligence [ChJB99] is a challenge; current research walks quickly in this direction [ZhLY03]. By this, we hope drastic improvements in the web [Kim02].

In this context, archives are an interesting and representative WIS research case. Currently, archives are living a digitising and standardization process. Digital Archive (here after, DA) is a collection of documentary data in digital format joined to descriptive metadata stored in a repository.

This paper describes some aspects of web DA integration, thanks to the acquired experience in two recent research lines. A line just finished is due to a Parliamentary Integrated Management System, -SIAP-, we have built. Its first product (called SGP) is successfully running at the Asamblea of Madrid since 1999 (please, see www.crcit.es/SIAP for details). The other line, DAWIS-UPM, is focused on heterogeneous web data sources integration and semantic web, applied to web DA virtual integration. Both researches are granted by Spanish National and Madrilenian Autonomous governments.

Building a web DA integrated architecture, with dispersed components and services, is a paradigmatic application [Shir02]. Additionally to the services, the overall functionality depends on the capability for the semantic specification and implementation each web query requires [MeMM97].

This paper proposes a virtual web architecture based on mediator (composed by ontologies, mappings and repositories) and wrappers (coupling heterogeneity between data sources, using XML) tiers.

Section 2 briefly introduces the international standard archival descriptions: ISAD(G) and ISAAR(CPF). Point 2.1 describes what a DA is. Section 3 summarizes some general aspects about web digital archive integration. After this, section 4 describes the DA Web Integrated Architecture we propose for DAWIS-UPM. Point 4.1 remarks the ontological aspects we have considered; applied to some existing archival standards (DCMI, ISAD, ISAAR) and parliamentary archive of SIAP system (we have built); and point 4.2 deals with an ontological integration kernel we suggest for getting a semantic archival unification. Conclusions and references are finally given.

1 SIAP Parliamentary Information Management System, a CRC Information Technologies product (www.crcit.es)
2 Archives

From a technical point of view, archives constitute a worthy and representative application case to consider in WIS research. They contain a huge documentary information above human activity. In real world, archives hold as different contents as kind of culture, business, living and relaxing we know. However, we could summarize that their main goal is: propitiating an easy access to the information content, guarantying the safe and custody of what they hold. Consequently, when you are looking for providing access to an archive content (as a catalyze of cultural and historical heritage), undoubtedly the web is the optimal broadcasting via.

The General International Standard Archival Description ISAD(G) (2nd ed.) [ISAD00] is the most important one for archivists. Its main goal is the identification and description of the content and context of the archival material, in order to promote its management and accessibility. ISAD(G) is developed by the International Council Archives ICA, a worldwide organization of archival community. ICA is in charge of the promotion, preservation, expansion and use of the cultural heritage. ISAD(G) defines a set of twenty-six elements (metadata terms) for archival description [Land00] that may be applied regardless the format or medium of the archival material. Some combinations of these elements constitute the unit of description.

Figure 1 represents ISAD hierarchical unit of description levels, from a general to a more specific content. Every level represents a unit of description with different degrees of detail. Additionally, there is another standard called ISAAR(CPF), the International Standard Archival Authority Record for Corporate Bodies, Persons and Families (please, see http://www.ica.org for ISAD and ISAAR details). Section 4 describes how we use these two standard concepts.

In figure 2 we can see the many different contents a DA could store. It's true that, in some archives, the process of digitizing could be hard, slow and may be impossible to get in a complete way. However, today more and more institutions are trying it in order to offer a higher accessibility. Hence, the web DA is disseminating certain contents that, generally, consist of metadata like: descriptive data for indexing, thesaurus of Information Retrieval, summaries, index, key words, synonymous and antonyms lists (broader / narrow terms, …) and other related terms [BaBe99].

3 Web Digital Archives Integration

Public DA integration should allow access to inherited cultural content without distances, languages and cultures barriers, using multiple devices that interconnect the web and hiding source information heterogeneity.

Currently, many important initiatives based on DA exist (OAI, DCMI, EAD, etc.) in order to put them accessible to the user (via browser and portals). The systematic and (semi-) automatic generation of web DA allows to improve the management of huge documentary inheritance held into Archives, as well as facilitates the creation of new DA.

However, we don't still have a good integrated web solution for making possible that any DA in the world could be ‘added’ to a web environment, independently of data source content and its description level.

3.1 Virtual Integration of Web Heterogeneous Data Sources

Dynamic and virtual web integration of multiple heterogeneous DA represents a very different challenge to the classic database static schema integration. This last is obtained by using a fixed federation of local participant databases schemata in a global static schema [OzVa99].
In the virtual integration of web heterogeneous data sources there is not centralized information management, neither federated information, neither integration information systems dealing with materialized data [DoDi99]. In the virtual integrated systems, data sources remain being autonomous data inside every local source. Integration is only virtual (no materialized), as we describe later.

In spite of having many standards in databases and static schema integration reference models, a systematic solution for web dynamic integration, requires a standard that doesn't exist yet (solutions are often static and *ad hoc*).

However, the subject of the non-existence of a standard web is old [Flor98]. Since the first attempts to define data exchange protocols in Internet (EDI) until the proliferation propitiated by the web (http, html, xml, obi, cxml, etc.), a proper model for the interactions among system components has not been possible yet [Leym01]. We are dealing with independent WIS, whose first requirement to be part of any integration is to guarantee a total autonomy to each participant local individual. In our case, we are referring to integrate DA without affecting the independence each DA owns.

Therefore, the DA doesn't federate in an integrated static global schema. On the contrary, it is produced by a **Web Mediator System** [WiGe97]. Doing so, data remains in native local data sources, and data integration takes place *on-the-flight*, during the web query processing in the mediator [MeMM97]. That is the **virtual** adjective meaning here used.

## 4 DAWIS-UPM

DAWIS-UPM is a national project which global objective is the definition of a web integrated architecture, providing a virtual and dynamic access to many different DA [CEMS02], [SCMC03], [ECMA03]. DAWIS-UPM is successor of SIAP, looking now for the web integration of multiple DA. In this sense, we find very valuable the acquired experience in the previous project we have done for the Parliament of Madrid (SIAP).

The main goal is to provide a semantic unification for web knowledge enrichment. For getting this, we don’t start from scratch. There are valuable proposals like FEDORA and OAI, that offer interoperable architectures, limited to provide remote access and C/S, to a determined DA [Fed02], [OA102], [OA103], [PrHa02]. Undoubtedly, they are an important first step.

DAWIS-UPM needs to provide a systematic and (if possible) automatic definition of the underlying **reference architecture** [CEMS02], through which web user could access to a lot of heterogeneous DA [DoMc03] in a transparent way. That means the consideration of the following aspects: web architecture, DA model specification, query processing, wrappers, and ‘mediators’ with ontologies, repositories and mappings. Figure 3 represents the proposed architecture, that will be service-based and compliant to current web-centric architectures (J2EE).

For each original archive and for each integration, the corresponding DA is expressed in terms of XML schema and mapped to a specific ontology.

The final idea is the implementation of this architecture by means of Java libraries based on service-levels [ECMA03], where ‘mediator’ and ‘wrappers’ will be dynamically generated starting from web user query. For this, the following architectural topics should be consider: flexibility, integration (and its related semantic problem), interoperability, cooperation and efficiency, as we have discussed in [CEMS02].

We want to remark a web services architecture doesn’t solve the semantic integration problem [Kreg01], [Brit01]. Web services allow to exchange data among remote applications, but they don’t guarantee that the receiver application will be able to understand them. As it is pointed out in [Shir02], the problem is the lack of a global ontology.

In **DAWIS**, Servlets are of special interest, because they have a certain semantic orientation. Servlets are Java classes for request processing and corresponding answers generation, supporting part of the application logic (although it will be only in an explicit way) [Kreg01]. Moreover, CORBA is another standard that needs a wide fan of services providing design and development of current distributed systems [Brit01].

---

**Fig.3. Web Digital Archive Integrated Architecture through Mediator and Wrappers.**
Thus, in WIS we consider ontologies notably contribute to reinforce semantics lacks, what is different to current distributed architecture approaches. They only consider defined static interactions among software components (during the compilation) and between programs and users [PaLa98].

Thus, in order to solve dynamic and virtual web integrated access problem, semantics of the systems is an important and preliminary characteristic must be considered.

An ontology describes shared knowledge about a specific domain in order to establish a communication via between humans and programs [CHIJO2], [BaBe99]. Each ontology provides a precise definition of the terms (concepts, properties) of a specific domain; that is to say, it facilitates a common language to share and to re-use knowledge of some phenomenon of the domain of interest [Grub93], [ZhLY03]. The goal is to achieve a common understanding consensus that could be able to re-use and share it between applications and groups. Consequently, applications can speak to each other, and interpret without ambiguity the information they exchange. For this reason, they have been accepted as powerful description tools that allow to make explicit the semantic web. Since last decade, many researches [CaGL01], [ChJB99], [BeHL01], [ZhLY03], [HaH97], [GuWe02] are dealing with these semantic problems, where ontologies play a fundamental role.

DAWIS-UPM needs to build ontologies [WVVS01] because DA domain is not simple and has a wide audience. We adopt ontologies as a paradigm that reduces complexity, and XML is the used technology for the interoperability of DA [Kim02].

Similarly to the architectural aspects we have mentioned, the archival conceptual modeling neither starts from scratch. On one hand, we have the well-known and very general DCMI (Dublin Core Metadata Initiative) ontology. On the other hand, ISAD(G) and ISAAR(CPF) are international standards specifying how archival content should be described. These standards are widely followed by archivists. Finally, a fourth special Parliamentary ontology -coming from SIAP Parliamentary DA- is also undertaking.

- DCMI: Dublin Core Metadata Initiative [DCMI03] promotes interoperable norms about metadata standards and develops specialized vocabularies for describing very general resources. According to RFC 2396 definition, a resource is understood as ‘something that has identity’.
- ISAD(G) Metamodel: The International Standard Archival Description [ISAD00] specifies the identification, content and context of archival information units. Also, it provides links in order to fix any combination of archival information elements. As it is said in section 2, ISAD(G) defines a set of twenty-six descriptive elements. Only six of these elements are considered essential and must be used in any archival unit of description (reference code, title, creator, date(s), extent of the unit of description and level of description).
- ISAAR(CPF) Metamodel: It can be considered as a complement of ISAD(G) to specifically manage the description of authority records [ISAA03]. It defines twenty-seven descriptive elements organized in four information areas.
- SIAP: Finally, we have also defined an ontology for the parliamentary archive of SIAP. Concerning to ISAAR(CPF) and ISAD(G) elements, in [SCMC03] you can find the UML respective models we have defined. Also, in http://sinbad.dit.upm.es, we provide ISAD, ISAAR and SIAP ontologies, already designed through Protégé tool.

DAIP DA has a special nature (parliamentary). The interest about this ontology goes further on, because - in the future- the integration way, here considered, will allow us to conceive a system able to offer rules to integrate this specialized concepts of diverse nature (police, government, etc.). Ontologies are complex to build as a hole in one step. So, building them step-by-step, empirically testing each new (kind of) addition and developing appropriated learning techniques for each step, you may automate the process; and, next time, you may build a new one in a more systematic way.

4.2 Ontological Integration Kernel

This section discusses our preliminary approach concerning to the definition and development of a global ontology [ZhLY03], unifying the semantics of the diverse existent ontologies related to DA domain [CaGL01].

Regarding this, we consider two important previous works: Observer [Mell01], [Mena98], [MIKS00], (http://sol1.cps.unizar.es:5080/OBSERVER) and Harmony [LaHu02], [DoHL03], (http://metadata.net/harmony). They constitute an important contribution to our current research.

Observer considers ontologies as if they were on only one horizontal plain. Then, inter-ontological mappings could be also considered in a parallel plain. For this reason, we call this kind of relationships ‘inter-ontological-horizontal-mappings’. Besides, Observer provides the capability to query many specific ontologies from different domains. Only if
the user wants, the system allows a new query formulation referring other ontologies concepts. In this sense, the answer enrichment is provided. We recognize Observer great improvement concerning query processing against many ontologies.

On the other hand, Harmony-a digital library project-defines a common ‘ABC metamodel’ for metadata integration coming from many other ontologies. Harmony considers ABC model as if it was on a higher plain allowing to attach other more specific ontologies on a lower plain (CIDOC/CRM, MPEG-7, IMS, …). It acts as an umbrella containing the most general concepts. Then, inter-ontological mappings are always provided in a vertical plain. For this reason, we call this kind of relationships "inter-ontological-vertical-mappings". Newly, we want to appreciate this valuable contribution.

Considering these two projects results as pillars, we propose to develop specialized techniques and methodologies that can guarantee a right shared and common semantic integrity among current DA ontologies. Our ontological integration kernel implementation will be performed over the definition of an ontological unification approach, operating as a kernel in which current (more or less) specialized ontologies could be semantically connected, as figure 4 represents.

The distinctive feature DAWIS presents is the semi-automatic ontological kernel generation for a global integration that doesn't have to be defined in a manual neither static way. Thus, this ontological kernel will be in charge of providing a common understanding for fundamental DA concepts. In order to obtain a flexible and scalable integration, ‘bridges or steps’ among ontologies -like an intelligent and generic ‘crosswalk’- need to be provided. This allows knowledge inference from any ontology to another.

We are aware of inter-ontological semi-automatic mappings definition can turn out to be a hard and expensive task. So, today, our research is going on this way, applying the step-by-step method previously mentioned.

5 Conclusions

This paper is focused on Web Digital Archives Information Systems looking for a virtual, flexible and dynamic integration by means of a Mediator and Wrappers architecture. We take our acquired experience with SIAP, Parliamentary Archive Management System successfully running at the Asamblea de Madrid since 1999. The paper presents four current ontologies, closely related to DA domain. Applying an empirical step-by-step methody, a high-integrated semantic description level is proposed by means of a global ontological kernel inside the Mediator layer.

Acknowledgements

This work is partially granted by Spanish Ministry of Science and Technology (MCYT-TIC2002-04050-C02-02, DAWIS-UPM project), by Community of Madrid (07T/0056/2003/3, EDAD-UPM project) and by Spanish NoE Databases, RedBD, MCYT-TIC2000-3250-E.

References

[DoHL03] Doerr M, Hunter J and Lagoze C, Towards a Core Ontology for Information Integration, JOIDI 4(1), April 03
[GuWe02] Guarino N and Welty Ch., Evaluating Ontological Decision with ONTOCLEAN. Communications of the ACM. vol.


[Kim02] Kim, H. Predicting how ontologies for the semantic web will evolve. Communications of the ACM. Vol. 45, n. 2, pp. 47


