OAIsistema, a simple solution for implementing OAI-PMH on small size archives of archaeological documents

CLAUDIO BARCHESI

ISCIMA CNR – ITALY

ABSTRACT

Since 1990 the Istituto di Studi sulle Civiltà Italiche e del Mediterraneo Antico of the Italian National Research Council (CNR) publishes the journal "Archeologia e Calcolatori" (http://soi.cnr.it/archcalc), under the direction of Paola Moscati. Its content focuses on the application of IT to archaeological methods and analyses the relevant theoretical aspects.

While Pdf seems to be the most satisfactory data format to deliver the electronic version of documents, international standards developed within the Open Archives paradigm can surely offer fascinating solutions to disseminate metadata describing their content. Especially for archeological and historical resources, Open Archives can play a main role, representing a valid base on which to establish projects aimed at a better visibility and wider diffusion of information.

For small research institutions and university departments that lack of expert Unix System Managers the most part of the existing OAI software is not simple to implement.

In this paper a simplified approach to OAI implementation for small and medium size archives will be discussed. This project is based on a OAI Static Repository file (OAI-SR, i.e. an XML file based on a schema whose guidelines has been recently issued by OAI), Visual Basic, ASP, XML/XSL and Java technologies. This project has been applied to the collection of PDF version and abstracts coming from the articles published in "Archeologia e Calcolatori", but it can easily be adapted to other cultural subjects with small efforts.

1. INTRODUCTION

The Istituto di Studi sulle Civiltà Italiche e del Mediterraneo Antico (ISCIMA) is a research institute of the Italian National Research Council. The Institute collects the personnel, the experiences and the projects of the former Istituto per l'Archeologia Etrusco-Italica (IAEI) and the Istituto per la Civiltà Fenicia e Punica "Sabatino Moscati" (IFP) that joined together in 2001.

ISCIMA inherits a strong tradition in the application of new methodological approaches to archaeology, as well as an established experience in the diffusion of scientific results and archaeological data collection by means of new media and IT standards (Barchesi, C. 2004). This experience, which began in the nineteen-eighties with the pioneering studies of Paola Moscati, has been important in focusing the recent researches, which has been carried out with respect to the relevant standards of scholarly communication and have followed a methodological approach concerning the applicability to other contexts and diverse documental collections (Moscati, P. 1999a, 1999b, 2001, 2002a, 2002b).

The ISCIMA publishes the international journal Archeologia e Calcolatori (http://soi.cnr.it/archcalc), founded in nineteen ninety by M. Cristofani e R. Francovich which is edited by Paola Moscati. The journal has an international Scientific Committee, which assures a rigorous and authoritative editorial line, and offers an extraordinarily broad range of topics covering all aspects regarding computer applications in archaeology. The index of the 15 volumes already published consists of about 400 articles in French, English, German, Italian and Spanish, written by scholars of the major Italian and foreign Institutes interested in computing in archaeology.

The proceedings of international conferences and special issues, devoted to themes of current interest in the field of archaeological computing, have also been regularly published.

In two thousand and four, an important agreement was established with the editors of the English journal "Archaeological Computing Newsletter" (http://www.gla.ac.uk/Acad/Archaeology/acn/index.html), which – starting from the 61st issue (December 2004) – has resumed regular publication as a supplement to Archeologia e Calcolatori.

This new international initiative has led the editorial board of Archeologia e Calcolatori to promote a project to increase the visibility of the journal and the diffusion of its scientific content. Some aspects of this plan are related to the website of the journal; in particular: increase the Internet visibility of Archeologia e Calcolatori, create a bibliographic index conform to international standards enriched with abstracts of the articles and related information, create a web tool to search the index, allow Internet search engine robots to index the collected records, permit some post-print delivery of relevant papers in PDF format.

All these requirements have been fulfilled through the OAIsistema project. This solution for implementing OAI-PMH (Open Archive Initiative-Protocol for Metadata Harvesting) (http://www.openarchives.org/OAI/openarchivesprotocol.html) on small size archives has been completely developed by the Archaeological Computing Service of the ISCIMA, on the basis of the guidelines of OAI-PMH, and some further technologies.

2. SOME CONCEPTS ABOUT OAI AND OAI-PMH

OAI (Open Archive Inititative) (http://www.openarchives.org/) was conceived by Herbert Van de Sompel, of Los Alamos National Laboratory, Research Library, New Mexico, US. The initial goal of OAI was to contribute in a concrete manner to the transformation of scholarly communication. The first meeting of OAI was held in 1999, in Santa Fe, New Mexico (Van de Sompel, H.; LAGOZE, C., 2000). The central theme of the meeting was the establishment of recommendations and mechanisms to facilitate cross-archive value-added services. (Lynch, C. A., 2001). OAI philosophy is based on the Data provider/Service provider distinction (lagoze, C.; Van de Sompel, H. 2001): a data provider is an institution that produces metadata for a class of resources, optionally offering a full text or a digital version of the documents. A service provider is a computer that harvests (collects and stores in a local database) metadata from data providers, and offers (value-added) service to web users, like cross-archive search interface, indexing, etc.
As for the technologies, OAI promotes the use of OAI-PMH that is based on: HTTP protocol, XML language and Dublin Core metadata schema, that is mandatory (also if other metadata schemata can be used). OAI-PMH is then a protocol based on metadata harvesting. It aims at creating the basis for cross-archive search by centralised services.

3. DUBLIN CORE METADATA STANDARD

The Dublin Core metadata standard (http://dublincore.org/) is a simple yet effective element set for describing a wide range of resources. It has 15 elements. Each element is optional and may be repeated. Although the Dublin Core was originally developed with an eye to describing document-like objects, DC metadata can be applied to other resources as well. Content data for some elements may be selected from "controlled vocabulary" (Hillmann, D. 2003).

4. EXISTING SOFTWARE FOR IMPLEMENTING OAI REPOSITORIES

In order to enter in the OA world as data provider much software exists. Amongst the most recognised solutions are Dspace (http://dspace.org/index.html) and EPrints (http://www.eprints.org/software.php) software. These are very powerful programs, which are also very complex to set-up. They need, to be correctly implemented, for Unix/Linux platform and also for expert web servers administrators. This is a pressing problem for many data providers who hold small collections of data but lack such a professional skill.
Lagoze and Van de Sompel, the 'fathers' of OAI-PMH, admitted in 2004 that the implementation of OAI is not trivial and that the size of some data collections often does not justify the investement.

5. A NEW OAI-PMH IMPLEMENTATION MODEL

A possible solution has been recently suggested: the model "Static repository and Static repository gateway" (see Fig. 1), already presented in 2003 (Hochstenbach, P. *et. al.*, 2003), has recently been raised to the degree of OAI implementation guideline. The Static Repository model builds on two types of components:

- the Static Repository itself: an XML file that is made accessible by a data provider at a persistent network-location; it contains metadata expressed on a specific OAI schema.
- the Static Repository Gateway: a network accessible server, operated by a third party that makes one or more Static Repositories harvestable through the OAI-PMH.

ISCIMA adopted such a model. We designed our OAI solution on Windows 2000 platform, which is universally well known and shared on the most part of desktop computers. ASP, DOM, Vbscript and XML-XSLT technologies, have been used to supply data providers with acting solutions for the needs of metadata management and static repository administrations.
Our implementation of the Static Repository model validated well on the OAI tool Repository Explorer (Hussein, S. 2001) and has been successfully registered on the OAI official data provider list. It was the first to appear in the OAI circuit as direct static repository implementation, and has been positively signalled by OAI managers in the OAI implementers mailing list (http://www.openarchives.org/pipermail/oai-implementers/2005-February/001433.html). At present the repository of Archeologia e Calcolatori is on-line, showing all metadata of published articles, as well as a starting collection of PDF resources (http://soi.cnr.it/archcalc/Bibliografia_new.htm).

6. "ARCHEOLOGIA E CALCOLATORI" OAI REPOSITORY

An appropriate set of metadata has been created in accordance with the Dublin Core specifications to describe each article published in Archeologia e Calcolatori. It contains: title, authors, source, description about page number, year of publication, curator, publisher, relation with other articles, identifier – i.e the URL for the full text digital version (if present) – format and type of the document, the abstract in English, original language, ISBN, and finally four subject

terms: one for the archaeological issues, one for the chronology related to the arguments described in the article, one for the topographic identification of the archaeological area, and lastly for the type of computer application. The metadata are stored in a database using the Microsoft Access software.
Visual Basic was used to create a script that performs the automatic export of the metadata table as an XML file conforming to the OAI schema. The exported file is saved in the root of a website and thus represents the XML OAI static repository. The process of exporting is automatic and therefore the metadata creator is not required to have any knowledge of XML and OAI-PMH.
In order to be accepted on an OAI circuit, the static repository requires a gateway. For this, we have been supported by the University of Ghent, which administers the gateway developed by Patrich Hochstenbach. The process for connecting to the gateway has been implemented by a script embedded in the software module of our database.
Through the Gateway of Ghent, Archeologia e Calcolatori can be harvested by OAI service provider. The static repository has been registered in the OAI official data provider register. OAISTER(http://oaister.umdl.umich.edu/o/oaister/) and SAIL of Italian CNR (http://eprints.bo.cnr.it/), harvest our data collection. In particular OAISTER (a cross archive search service) added Archeologia e Calcolatori to its 458 institutional data providers that offers 5,270,000 records.

7. ADDICTIONAL WEB SERVICES FOR "ARCHEOLOGIA E CALCOLATORI" REPOSITORY

Besides entering OAI, we have also used our data collection to produce two further services:
The first is based on IIS of Microsoft Windows that allows for the use of ASP technology to perform dynamic server side operations on data; by the use of ASP, ODBC and ADO a local search engine for the website of "Archaeologia e Calcolatori" has been created.
Using ASP, Vbscript and the MSXML parser for DOM, we have added a second service, able to offer to the Google spider (also called robot or crawler) a persistent URL for each record contained in the repository, which may be in this way easily indexed on the web. This service is similar to DP9 (Liu X. *et al.*, 2002), but it is focused for a personal use (see Fig. 2). Google spider finds on the homepage of "Archeologia e Calcolatori" a hyperlink to an HTML page listing the journal issues ordered by date; each element of the list has a parameterised hyperlink. Hyperlinks point to an ASP page to which the date parameter is passed. The ASP page forwards this parameter to an XSL stylesheet, which selects the matching elements from the XML static repository file and returns an HTML document listing all the articles published in that year. Each elements of the produced list has a parameterised hyperlink with the ID of the article. A further ASP page forwards this parameter to a last XSL stylesheet which selects the specified record from static repository file and returns the final HTML document. The HTML restitution contains metadata of the article inside the body section of the page as well as in the head section, following the indications of Dublin Core Initiative (Kunze J.). Following the hyperlinks chain Google spider can index in such a way the entire repository.
As a consequence of the OAIsistema implementation Internet users have three different paths to recover "Archeologia e Calcolatori" articles on web (see Fig. 3):

- firstly through Google, the most popular search engine tool on the web. Google sees each record of the repository as a single HTML page accessible at a persistent URL;
- the second path is accessible through the journal's website, where a user can perform a local search, combining authors, subjects and date;
- the third path is accessible through the OAI cross-archives search engines.

CONCLUSIONS

The OAIsistema project may be easily applied to other kinds of small archaeological archives. This can contribute to lowering barriers in the adoption of Open Archive protocol for archaeological metadata documents, increasing the visibility of scientific research and promoting the diffusion of results towards a wider audience.

REFERENCES

BARCHESI, C. (2004) – Il settore informatico dell'ISCIMA. Un percorso ventennale nell'informatica archeologica. "Proceedings of CAA2004-Italia, Aprile 2004, Prato", *Journal for the Intercultural and Interdisciplinary Archaeology*, to be published.

HOCHSTENBACH, P.; JEREZ, H.; VAN DE SOMPEL, H. (2003) – The OAI-PMH Static Repository and Static Repository Gateway. – http://lib-www.lanl.gov/~herbertv/papers/jcdl2003-submitted-draft.pdf (visited on 15/04/2005).

HILLMANN, D. (2003) – Using Dublin Core – http://dublincore.org/documents/usageguide/ (visited on 15/4/2005).

HUSSEIN, S. (2001) – Using the repository explorer to achieve OAI protocol compliance. In JCDL '01, June 24-28, 2001, Roanoke, Virginia, USA. New York, NY: ACM press, p. 459 (http://www.dlib.vt.edu/projects/OAI/reports/jcdl_2001_demo_repository_explorer.pdf) (visited on 15/4/2005).

KUNZE J. – Encoding Dublin Core in HTML, http://www.ietf.org/rfc/rfc2731.txt (visited on 15/4/2005).

LAGOZE, C.; VAN DE SOMPEL, H. (2001) – The Open Archives Initiative:Building a low-barrier interoperability framework. In JCDL '01, June 24-28, 2001, Roanoke, Virginia, USA. New York, NY: ACM press, p. 54-62 (http://www.openarchives.org/documents/jcdl2001-oai.pdf) (visited on 15/4/2005).

LIU, X., *et al.* (2002) – DP9: an OAI gateway service for Web Crawlers. In JCDL '02, July 13-17, 2002, Portland, Oregon, USA, http://www.cs.odu.edu/~liu_x/dp9/dp9.pdf (visited on 15/4/2005).

LYNCH, C. A. (2001) – Metadata harvesting and Open Archives Initiative. ARL Bimonthly Report 217. Washington D.C.: ARL (August 2001), p. 1-9 (http://www.arl.org/newsltr/217/mhp.html) (visited on 15/4/2005).

MOSCATI, P. (1999A) – The "Caere Project": Methodological and Technical Considerations. In GUARINO, A. ed. – Proceedings of the II International Congress "Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin (Paris 1999), Paris: CNR-CNRS, p. 119-128.

MOSCATI, P.; MARIOTTI, S.; LIMATA, B. (1999B) – Il "Progetto Caere": un esempio di informatizzazione dei diari di scavo, *Archeologia e Calcolatori*, Firenze. 10, p. 165-188 (http://www.progettocaere.rm.cnr.it/databasegestione/open_oai_page.asp?id=oai:www.progettocaere.rm.cnr.it/databasegestione/A_C_oai_Archive.xml:257).

MOSCATI, P. (2001) – Progetto Caere: questioni di metodo e sperimentazioni, *Archeologia e Calcolatori*, Firenze, 12, p. 47-53 (http://www.progettocaere.rm.cnr.it/databasegestione/open_oai_page.asp?id=oai:www.progettocaere.rm.cnr.it/databasegestione/A_C_oai_Archive.xml:296).

MOSCATI, P. (2002A) – From an Etruscan town to modern technologies: new advancements in the "Caere Project". In F. DJINDJIAN, F.; MOSCATI, P., eds., XIV Congress of the UISPP, Commission IV Data Management and Mathematical Methods in Archaeology (Liège 2001), *Archeologia e Calcolatori*, Firenze, 13, p. 135-149.

MOSCATI, P. (2002B) – Archeologia informatica: tra tradizione e rinnovamento, *Bollettino IC* n.s. 5, p. 21-27.

VAN DE SOMPEL, H.; LAGOZE, C. (2000) – The Santa Fe Convention of the Open Archives Initiative. D-Lib Magazine, 2000 (http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html) (visited on 15/4/2005).

FIGURES

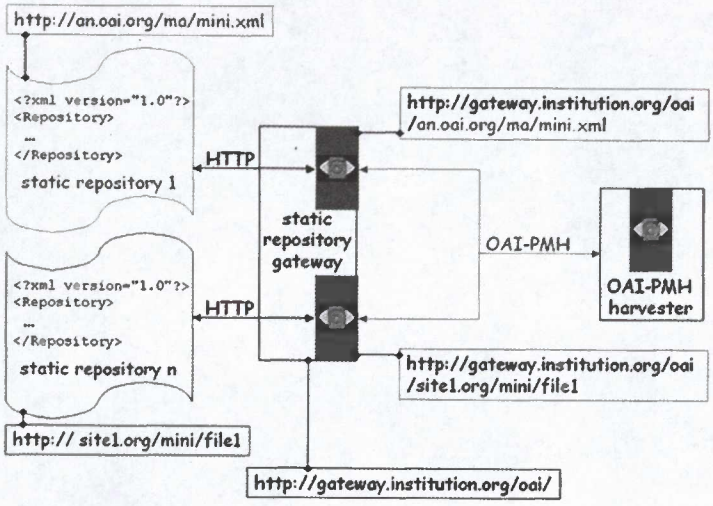

Fig. 1 – OAI static repository model (from HOCHSTENBACH, P. *et al.*, 2003).

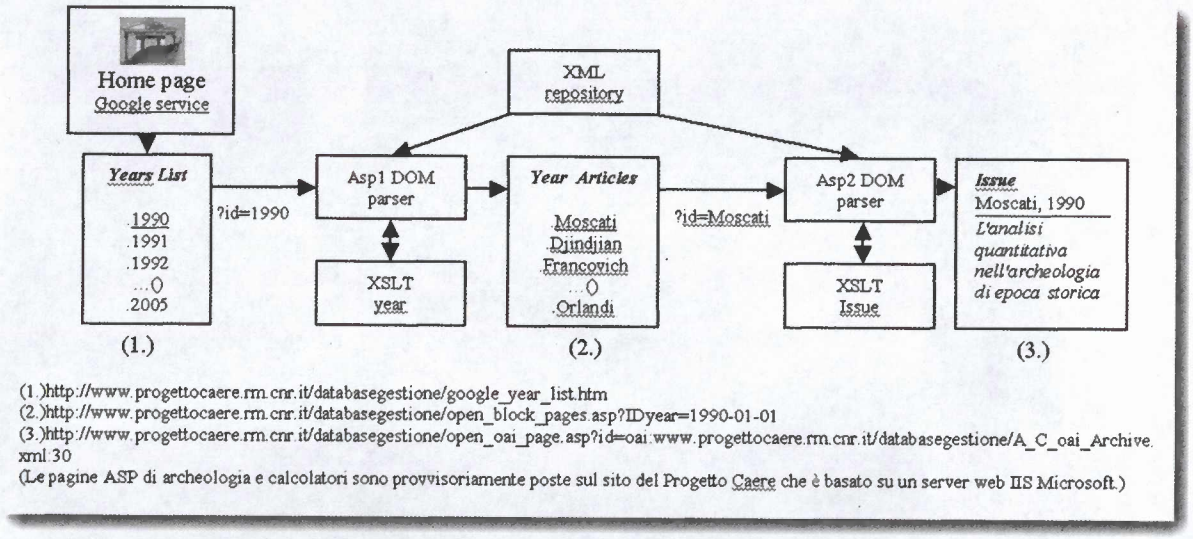

Fig. 2 – OAIsistema solution for the Google indexing of the repository.

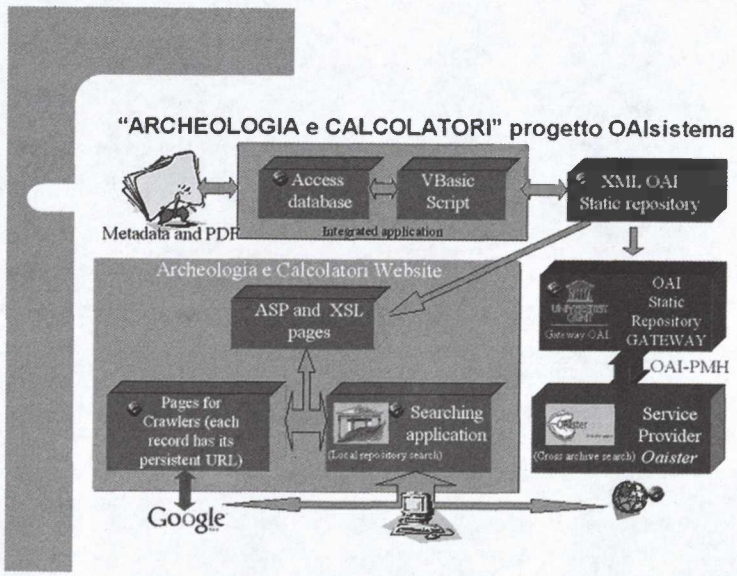

Fig. 3 – OAIsistema block diagram.