The SciELO Brazilian Scientific Journal Gateway and Open Archives

A Report on the Development of the SciELO-Open Archives Data Provider Server

Carlos Henrique Marcondes
Information Science Department, Federal Fluminense University, Brazil
<marcon@vm.uff.br>

Luís Fernando Sayão
Nuclear Information Center, Nuclear Energy National Commission, Brazil
<lsayao@cnen.gov.br>

Abstract

SciELO, the Scientific Electronic Library Online, uses a methodology developed by BIREME/PAHO/WHO [1] that enables the implementation of web digital libraries of scientific journal collections of full text articles. Various SciELO gateways are now in operation, providing access to academic journals from Brazil and other countries in Latin America and the Caribbean. SciELO plays a very important role in the worldwide dissemination of the technical and scientific literature published in developing countries, thereby increasing visibility of this literature that otherwise would be accessible only within the borders of those developing countries.

The SciELO methodology utilizes ISIS software for formatting and maintaining the SciELO metadata database. UNESCO developed this software as well as other associated software, which serve as the bases for Science, Technology and Medicine (STM) information systems, databases and networks in several developing countries. This article reports on the results of a pilot-project in which a generic ISIS metadata database interface — used by SciELO — compliant with the Open Archives Initiative (OAI) was developed.

1. Introduction

"Other than permanent peace, it is possible that nothing could contribute more to achieving the goals of developing societies, the less industrialized countries, and to make this world a better world to live in than an effective worldwide transfer of scientific and technological information" (Pauline Atherton, UNESCO's Handbook for Information Systems and Services, 1979).
Science and technology in developing countries face several barriers to effectively playing a strong role in the development of those societies. These barriers include: inadequate financial and human resources, insufficient commercial sector involvement, and the lack of a nationwide development project. These barriers all affect the ability to access and transfer scientific and technological information.

The focus of scientific research in developing countries frequently concerns the specific problems faced by those societies — problems not generally faced by developed countries. For example, EMBRAPA, the Brazilian Agricultural and Cattle Raising Research Corporation, has been focusing its research on tropical crops, and FIOCRUZ, the Oswaldo Cruz Foundation, researches tropical diseases and public health care. In developing countries, researchers use native idioms to communicate their research results, and their communication channels are generally local and irregular. This scenario and the accompanying barriers faced by developing countries to publishing in international scientific journals mean that the results of their STM research is visible only to a few [Sayão, 1996]. Furthermore, because few of the articles from developing countries are published internationally, it follows that few of the articles are included in international databases and other information systems. Even among other developing countries that could benefit from such research, awareness of and access to the research is limited. Since the 1960s, great effort has been made by international organizations like UNESCO, to change this situation.

With the advent of the Internet, in recent years the international scientific community has been engaged in the development of new alternatives for scientific and technical publishing, exchange of information, and access to the results of scientific research. A new paradigm for scientific communication is being created: direct publishing of full-text scientific papers in electronic archives freely accessible via the web.

An example of this new paradigm is ArXiv, the preprint electronic repository at Los Alamos National Laboratory, which was created in 1991 by an American physicist named Paul Ginsparg [Van de Sompel, 2000]. Following Ginsparg's successful eprint archive effort, the international scientific community has been engaged in developing practical alternatives for "free" publishing and open access to academic works using the Internet. As part of this development, new initiatives such as PubMed Central [2], the Public Library of Science [3], BioMed Central [4] and the Open Archives Initiative (OAI) [5] are providing open access to full-text Scientific, Technical, and Medical (STM) documents. The worldwide dimension of open access electronic archives is evidenced by the extensive list of available resources at the OSTI web site [6]. The aim of the OAI is to provide interoperability between web archives of full-text STM documents. The OAI is having an increasing impact on scientific communication and access to scientific documents.

SciELO, the Scientific Electronic Library Online, uses a methodology that enables the implementation of web digital libraries of scientific journal collections of full-text articles. SciELO was developed by BIREME [7] (the Latin America and Caribbean Center on Health Sciences Information), an organization belonging to PAHO (the PanAmerican Health Organization) and to WHO (the World Health Organization). Several SciELO gateways are now in operation, providing access to academic journals from Brazil and other Latin American and Caribbean countries, as well as to academic journals from Spain and Portugal. SciELO plays a very important role in the worldwide dissemination of the technical and scientific literature published in developing countries, thereby increasing visibility for the literature that otherwise would be accessible only within the borders of those developing countries.
The SciELO methodology utilizes ISIS software for formatting and maintaining the SciELO metadata database. UNESCO developed this software and other associated software, which serve as the bases for Science, Technology and Medicine (STM) information systems, databases and networks in several developing countries. This article reports on the results of a pilot-project in which a generic ISIS metadata database interface — used by SciELO — compliant with the Open Archives Initiative (OAI) was developed.

The Open Archives Protocol for Metadata Harvesting (OAI-PMH) is a protocol for automatically gathering the metadata of electronic documents maintained in web archives. The SciELO-Open Archives Data Provider interface could easily be applied to other ISIS database format information systems within developing countries, making them all interoperable and compliant with OAI.

2. What is SciELO

SciELO comprises a set of methodologies for electronic publication, access and preservation of STM full-text journals, using the web. SciELO is one of the most important services provided by BIREME to, among others, the Latin American and Caribbean scientific communities.

SciELO is the product of a partnership among FAPESP [8] (the State of São Paulo Science Foundation), BIREME, and national and international institutions related to scientific communication. With the aim of developing and evaluating an adequate methodology for electronic publishing on the Internet, a pilot project involving 10 Brazilian journals representing different subject areas was successfully carried out from March 1997 to May 1998. Following the pilot project (since June 1998), SciELO has been operating regularly, progressively incorporating new journal titles and expanding its operation to other countries.

Today, the SciELO gateway in Brazil has holdings of about 10,000 articles from more than 90 academic journals. SciELO's aim is to increase the quality, visibility and impact of research from Brazil and other developing countries. Recent surveys confirm that electronic articles are cited much more often than articles published in paper format. In fact, according to Meneghini's presentation [Meneghini, 2002] in the International Conference on Scientific Electronic Publishing in Developing Countries (ICSEP) [9], the ISI impact factor of SciELO journals has increased an average of 42% between 1998 — the beginning of SciELO operation — and 2001.

The SciELO methodology encompasses various components. The first component is the electronic publishing of complete editions of scientific journals and the organization of full-text databases capable of being searched and producing statistical indicators of their usage and impact. Digital formatting of full-text articles using mark-up languages such as SGML aids in the digital preservation of the articles. The SciELO methodology also includes journal evaluation criteria based on international scientific communication standards.

Another important component of the SciELO methodology is its set of tools for operating web digital libraries of full-text scientific articles. Presently there are various SciELO web sites in operation, including national sites as well as thematic sites. The first, pioneering SciELO web site was the Brazil site [10]. It has been joined by web sites in Chile [11] and Cuba [12] which are also in regular operation. Several other countries are evaluating SciELO and/or are in the process of being trained on the SciELO methodology. SciELO
Public Health [13], a regional thematic library covering Public Health scientific journals from Latin America and Spain, was launched in December 1999. A portal to integrate and provide access to the network of SciELO sites is available [14].

The SciELO methodology also incorporates new value-added services by adding links from article records [Santana, 2001] to bibliographic databases such as LILACS and MEDLINE and, recently, to the CVLattes system (the curriculum vitae of Brazilian researchers) maintained by the Brazilian Council for Scientific and Technological Development [15] (CNPq) and to the Brazilian Digital Library of Thesis and Dissertations (BDTD) [Marcondes, 2002a] maintained by the Brazilian Institute for Scientific and Technological Information [16] (IBICT). Cross-linking and cross-navigation among all these systems provide strong integration.

The success of SciELO depends heavily on the development of strong partnerships among national and international scientific communication players — authors, editors, scientific and technological institutions, funding agencies, universities, libraries, etc. — aiming at the dissemination, improvement and sustainability of the SciELO model.

In addition, an important SciELO methodology component is the use of the CDS/ISIS software format for maintaining the databases in BIREME and STM information systems in developing countries.

3. ISIS software and STM information systems in developing countries

The CDS/ISIS software plays a very important role in STM information systems in developing countries. CDS/ISIS was originally conceived at the International Labour Organization (ILO) as a database management system for textual data running on mainframes. With the rise in usage of microcomputers in the 1980s, UNESCO decided to adapt the mainframe version of CDS/ISIS to run on personal computers (PCs) under the DOS operating system. The Micro CDS/ISIS version 1.0 was released in 1985 and has evolved to the current version 3.8. At the same time, UNESCO developed a minicomputer version that runs on an HP 3000 platform and, more recently, developed a UNIX version. In 1995 UNESCO released the first version for the WINDOWS platform [17].

CDS/ISIS software is a database management system with especially strong support for textual data, typically used in document management systems, libraries, documentation centers and archives. CDS/ISIS software has facilities to easily manage variable size fields and records, optional fields, and repeatable fields used in document management systems. In addition to these features, CDS/ISIS software has a powerful formatting language that provides facilities for programming various layouts of output on screen or printer. The software's indexing and retrieval facilities use a powerful search language, which allows great precision in specifying search criteria necessary for document retrieval and provides field-level and proximity search operators in addition to the traditional Boolean operators and/or/not. The software enables free-text searching as well. In addition to these search functions, CDS/ISIS has facilities for importing and exporting data according to standards, such as ISO2709, enabling its use as the basis for distributed networks and information systems. CDS/ISIS was created as multi-lingual software, providing integrated facilities for the development of local linguistic versions as well.

These features make CDS/ISIS a powerful tool for use in STM information systems located throughout the world, but an especially important tool for use in developing countries. The CDS/ISIS software is being used by some of the most important STM information systems and networks in developing countries, including LILACS, REPIDISCA, CEPAL, etc. At
the same time, the use of CDS/ISIS software creates an international network of distributors, user groups, independent developers, events and publications that provides distribution, overall support, training and software tools worldwide.

BIREME/PAHO/WHO uses and develops tools for adding more features to the CDS/ISIS software as well as for providing information services to Latin American and Caribbean scientific communities. Originally, BIREME used CDS/ISIS as the basic software for all its network and system applications for data entry, information retrieval, and bibliographic data interchange.

The need for new features and the desire to address ISIS software limitations has led BIREME to develop software tools, libraries and utilities that significantly extend the original features CDS/ISIS [18]. At the same time, however, BIREME's development strategy was to maintain the CDS/ISIS database format across all these newly developed tools. Among the tools are: the CISIS utilities library, a set of utilities for manipulating databases in the CDS/ISIS format (available in versions for the PC environment under the DOS/WINDOWS operating system and available for different UNIX operating systems); WWWIsis, a tool for web management of CDS/ISIS databases; and IAH, a powerful web information retrieval interface. BIREME developed these technologies with the aim of providing enhanced facilities for STM information systems in developing countries.

4. SciELO and Open Archives

The arrival of the Internet and the corresponding opportunity for direct web publishing gave developing countries access to free STM information maintained in web archives [Ginsparg, 1996]. Currently, a lively, ongoing debate is taking place between the scientific community and the publishing community about free access to academic articles. The scientific community considers access to electronic academic publications as the means to increase the visibility of its research and to accelerate the pace of scientific development as well as to disseminate more widely the results of research, which is considered by many as the heritage of all humanity [Harnad]. According to some, the imposition of fees for access to scientific information by international commercial publishers hinders the free flow of research results and slows scientific development. However, even if STM information is free, developing countries can only take advantage of its availability if they have the adequate tools. The scientific community's efforts towards freeing STM information is, at present, centered around the Open Archives Initiative (OAI). The OAI proposes standards to enable interoperability between different web archives of STM full-text documents. This interoperability problem arises when, for example, one paper is published in different electronic archives or when different archives focus in the same subject.

The Open Archives Initiative Protocol aims to standardize the dialogue between two types of institutional OAI partners: Data Providers and Service Providers:

- **Data Providers** are Internet servers with electronic archives holding digital documents and metadata about those documents. Data Providers expose metadata about their documents. Responses from Data Providers to Service Provider requests are all formatted in XML [19].
**Service Providers** are services that use metadata collected automatically from Data Providers, to develop value-added information services such as unified access, peer-review, qualified databases, etc.

The Open Archives Protocol for Metadata Harvesting (OAI-PMH) is a standard encompassing a common metadata format, the Dublin Core Metadata Element Set [Dublin Core, 1999], and procedures enabling the automatic gathering (by a robot program called a harvester) of document metadata stored in an electronic archive. A Service Provider's harvester sends requests for metadata to the Data Providers' OAI-PMH servers, which respond to the requests by sending document metadata held on the Data Providers' servers to the Service Provider. The request commands, or OAI-PMH verbs, include the following six:

1. **Identify**, which returns information about the electronic archive and its policies;
2. **ListSets**, which returns the set schema of the electronic archive;
3. **ListMetadataFormat**, which returns the identification of the metadata format through which metadata about the archive's holdings can be disseminated;
4. **ListIdentifiers**, which returns the unique identifiers of metadata records maintained in the electronic archive;
5. **GetRecord**, which returns the metadata of a single record; and
6. **ListRecords**, which returns the metadata of various records.

Once the metadata from different electronic archives is gathered, a Service Provider can develop, for example, a database with metadata from various electronic archives and, thus, can provide a unified search service. One of the metadata elements disseminated by OAI-PMH is the **Identifier**, the URI of the electronic document that enables access to a document's full text. The most common type of Service Provider is one that provides unified search services, such as those provided by Arc [20], Citebase [21], OAIster [22], etc.

The Brazilian Digital Library in Science and Technology Project [Marcondes, 2002b] conducted by IBICT first identified the problem of interoperability and unified access among different Brazilian digital libraries, archives, electronic theses and dissertation (ETD) collections, and STM full-text document databases. The Brazilian Digital Library (BDL) proposed a web gateway to provide unified access to metadata databases collected from different Brazilian STM systems through OAI-PMH. BDL's first subproject was a union catalog of Brazilian electronic theses and dissertations, encompassing ETDs from universities like USP, PUC-RIO, UFSC and ENSP/FIOCRUZ. As one of the principal partners of the BDL project, recently BIREME decided to enhance the services and facilities provided by SciELO and to increase the visibility of journal articles held in SciELO, an OAI-PMH Data Provider Server, through an interface compliant with the Open Archives Protocol for Metadata Harvesting. This would enable metadata about SciELO journal articles to be exposed and collected by any Service Provider, including BDL.

The SciELO-Open Archives project began in 2002 with a feasibility study. The project implemented, in a local computer, two of the OAI-PMH verbs: **ListIdentifier** and **GetRecord**. The feasibility study used BIREME's product WWWIsis XML IsisScript Server, a tool for web manipulation of CDS/ISIS databases.
After evaluation of the results of the feasibility study, BIREME's computer team began a pilot-project. In order to favor generality, the pilot-project used the PHP language instead of WWWIsis. All six OAI-PMH verbs were implemented. The SciELO-Open Archives server [23] is compliant with version 2.0 of OAI-PMH. The screens presented in the Appendix to this article illustrate the interaction with the SciELO Service Provider server through the OAI Repository Explorer.

The project implemented the OAI concept of "Sets", which can be associated with a classification schema or with different collections. In the SciELO-Open Archives project, the Sets are associated, at present, with each electronic journal collection, and the SetSpec of each journal is its ISSN. The concept of resumptiontoken, a mechanism to properly retrieve different segments of a long list of record identifiers or records was implemented as well.

At this time, the implementation supports only the OAI-Dublin Core metadata format, the default format of the Open Archives Initiative. The SciELO "default" metadata format is LILACS (Latin America and Caribbean Health Science Literature), a regional information system with a much richer metadata format, in which every record includes the title in English, Portuguese and Spanish. The thematic description of a document uses a classification schema and Health Science Descriptors [24] (DeCS). A trilingual and structured vocabulary [25], DeCS was created by BIREME for use in indexing articles from scientific journals, books, conference proceedings, technical reports, and other similar types of materials, as well as for searching and retrieving subjects from scientific literature in LILACS, MEDLINE and other databases. In LILACS methodology, all descriptors assigned to a record are automatically translated to English, Portuguese and Spanish.

The LILACS metadata format was developed from MeSH (Medical Subject Headings) of the U.S. National Library of Medicine with the aim of permitting the use of common terminology for searching in three languages, thus providing a consistent and unique environment for the retrieval of information regardless of the language.

A natural step forward for the SciELO-Open Archives project is the development of an XML schema for the LILACS format.

Because the metadata for journal articles, once published, is included in the SciELO database and never changes — as do other common Open Archives documents like preprints and research reports — in the SciELO-Open Archives project the OAI concept of datatestamp is associated with the article's date of publication.

The SciELO-Open Archives project took advantage of the existence of the OAI Repository Explorer, which is a site where it is possible to test repository compliance with OAI-PMH. The OAI Repository Explorer sends OAI-PMH requests to a specified URL and reports any errors that are detected.

Now that the initial implementation is complete, the SciELO-Open Archives project has begun an evaluation phase. The SciELO OAI-PMH server performance must be observed under real access conditions. Access logs must be examined to find out which Service Providers accessed SciELO, and metadata from journal articles held in SciELO must be traced. The provenance statements of the about container will be included in each record delivered by SciELO to meet this objective. The aim of this evaluation phase is to determine to what extent the adoption of the SciELO-OAI interface improved the visibility and impact of journal articles held in SciELO.
5. Conclusions

The SciELO-Open Archives project was conceived with the aim of extending the SciELO framework so that it could be applied to CDS/ISIS metadata databases. After the pilot-project, it seemed feasible that the lessons learned in the pilot would enable us to meet this objective. The availability of a routine that could extend CDS/ISIS capabilities to make a CDS/ISIS metadata database compliant with OAI-PMH as a Data Provider would be a strong tool to increase the visibility of any information system using CDS/ISIS software.

The development of the SciELO-Open Archives server benefited from a significant effort by BIREME to keep SciELO methodology up to date and to improve, with the use of new technological tools, the visibility and impact of STM information from developing countries.

The recent increase of freely available STM literature and the existence of a simple but powerful technological framework for access and interoperability like the Open Archives Initiative is of great significance to developing countries [Chan, 2002]: it provides the promise of a more equitable distribution of global knowledge. However, the availability of free STM information is far from enough. Developing countries must correctly identify and evaluate information appropriate for their specific needs. They need technological tools, like the SciELO-Open Archives Data Provider Server, to take advantage of open access to STM literature. Furthermore, making the effort to provide such access is not a task to be taken by each library or information system alone; it is a task that relies on the cooperative efforts of many such institutions.

Notes

[1] BIREME/PAHO/WHO: BIREME (Latin America and Caribbean Center on Health Sciences Information); PAHO (PanAmerican Health Organization); and WHO (World Health Organization).
[8] FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) <http://www.fapesp.br>.


Bibliography


Appendix

Screen shots illustrating interaction with the SciELO Service Provider Server through the OAI Repository Explorer:
Figure 1: Archive Self Description page

Figure 2: List of Fields
Figure 3: Web page for the Brazilian Journal of Biology

Copyright © Carlos Henrique Marcondes and Luís Fernando Sayão

DOI: 10.1045/march2003-marcondes