Communications to SIMAI Congress,
ISSN 1827-9015, Vol. 1 (2006)

DOI: 10.1685/CSC06079

Open Access Tools[1] and Technology

Nunzio Femmino' and Dario Orselli[2]

The new digital technologies and the world wide web, offered solutions and strategies for the development of electronic publishing and correlated services, also creating an alternate model of scientific literatures circulation.

If new technologies found solutions, formats and standards[3], the Internet has became the main actor to spread the *e-prints*[4]. That allowes to start new parallel models, that would be able to overcome the barriers set to the dissemination of contributes of scientific research results.

As it is well known, the BOAI[5] clearly describes the goals of the Open Access movement, i.e. to make freely available, and without any charge, the refereed scientific literature. BOAI also defines the two OA main strategies: the *Open access self-archiving* and the *Open access Publishing*. The first one promotes the self archiving, i.e. the submission of referred research output, the second one to support the creation of new open access journals and the conversion of existing academic and commercial ones.

The first strategy is performed by *Open Archive*[6] implementation, mainly with *open source* software which are distributed under the GNU-GPL[7] license, i.e. the freedom granted by the author to the software users to execute, to copy, to distribute, to modify it, and to redistribute the changes, in respect of the only restriction imposed i.e. every copy or change made, have to inherit the same liberties about the 'open source code'.

In fact the Open Archive software are mainly distributed in open source mode as their support infrastructures (as databases and operating systems)[8] and the very programming languages[9] used for their implementation.

The open archives can be grouped in two great areas according their organizational methods: *institutional OA* ones collect the witness of scientific production or the cultural activities of an Institution or Corporate hosted; *disciplinary OA* ones collect contributes from the same discipline. They can be also divided in *centralized* or *distributed* architecture according to the method they used to deposit the contributes: on one alone server or on many servers remotely connected by unique search interface.

Among the main characteristics of an open archive, both institutional and disciplinary, we can distinguish: standard *Metadata*, MARC[10] and Dublin Core[11] useful to share the contents with the most important web search engines; the OAI-PMH[12] a communication protocol that allows the interoperability and the data exchange with other Open Archives through a harvesting system.

The international organization reference for the Open Archive is the OAI[13] whose intent was to promote the interoperability of heterogeneous archives and dissemination of the e-prints.

The OAI architecture is composed by two different elements: the *Data Providers* and the *Service Providers*. The first ones (DP) are the archives where the full text research jobs (articles, conference presentation, etc.) and relative metadata are deposited, i.e. they are physical data containers (*repository*)[14]; the second ones (SP) are high level systems comparison to the DP as they can offer added value services, like indexing and collecting metadata from others DP (*harvesting*).

The harvester is a client that, according to the OAI-PMH, forwards one of the six commands (*Identify, ListMetadataFormat, ListSets, GetRecord, ListIdentifiers, ListRecord*) to the repository, which can answer with different metadata formats in XML[15] output schema, common to every repository. However, the harvesting process of metadata is not so easy and immediate as it seems. In fact, most of the existing tools, require several manual steps at the system console or the possible creation of some simple scripts of data conversion.

However, through an harvesting activity of metadata, many projects have been realized, such as portals[16] of simultaneous virtual search in different Data Providers which offer to the end user a high level service.

Among the most popular open archive software, the most famous is certainly *EPrints archive software*, developed by Southampton University with more than 200 installations all over the world. Other available softwares, as *CDSware*, *Dspace* or *Fedora*, with as good characteristics, are valid alternatives.

EPrints owes its popularity to different reasons: EPrints was one of the first projects to develop an open archive software. Its great ease of installation and personalization, the presence of a vast and active community which has developed around it, supported by the punctual and direct assistance of the development group, have contributed to its propagation and have created, indeed, a 'standard' in choosing of the software to be implemented.

Dspace was developed in 2000, within a joined project between MIT (Massachussetts Institute of Technology) and Hewlett-Packard Company. Thanks to its structured architecture in community, sub community, collections and items, it is particularly suitable for the dissemination and management of the didactic and multimedia materials and, naturally, of the traditional e-prints.

Some open archive software have some important peculiarities: they are repository and service provider at the same time. One of this is CERN Document Server Software (*CDSware*), a portal developed by CERN of Geneva that operates, through a single advanced search interface, as main collector of electronic resources and bibliographical records of its digital library. Through different tools available with CDSware (*BibHarvest, BibConvert, BibFormat* and *BibUpload*) it is possible to integrate different data collections, even completely different in their typology and contemporarily make them accessible to the end user.

The experience of the CERN which integrates in its own portal articles collections, pre- and post-print, books, videos, photos and many other and Antonella De Robbio's[17] promptings for the same experimentation in Italy, have contributed to the creation, at the University of Messina, of an institutional portal with CDSware. The Messina portal integrated in several collections, different bibliographical catalogues (among them, some with no standard), its own institutional open archive, already implemented with EPrints archive software and, through a hard-working of harvesting activity, some institutional Italian open archives and several international disciplinary open archives, among which the famous BioMed Central and RePEc.

The external sophisticated tools, together with metadata standard (MARC21) and the double and simultaneous DP and SP function, make CDSware an unique international open archive software. The external tools are particularly ultimate contributions. *BibConvert*, is a module through which it is possible to convert any OAI or not OAI compatible sequential text files to the XML MARC21 format that's necessary to be uploaded into CDSware. It's a very flexible and powerful tool written in *Python* that can perform complex conversion operations through a configuration template.

The output format of the search results is managed by *BibFormat* that allows to customize fonts, dimensions and colours. The output file will be an XML file that will contain the data information of the input file together with their formatting data. The two files so obtained, joined by the *sysnumber*, will be imported both into the CDSware database, creating a duplication of information. This allows to realize in a simple and immediate way, different kinds of visualization according to the data collections. In addition, through the *link Rules* function it's possible to use *BibFormat* as a automatic *links constructor* and create, for example, a link to a journal site from its title.

The harvesting process is performed by *BibHarvest*, another module of CDSware software, even if it's possible to execute it also through a web browser through above-mentioned commands.

There are different technological tools and different methods of approach about the second strategy, that's the *Open Access Publishing*. The implementation of an open access electronic journal needs to perform all required steps: one will need a web domain and a web space[18] that assure the persistence of the data; to get the identifiable International Standard Serial Number (ISSN)[19] of the journal; to undersign the Digital Object Identifier (DOI)[20], an alphanumeric string that permits to identify on-line, in univocal way, digital objects as documents of text, images or audio file. The inserting into the Directory of Open Access journals (DOAJ) [21] (nowadays over 2100 journals) assures a wide visibility.

Amongst the most known software tools for the management open access electronic journals can be numbered Open Journal System (OJS)[22] and Hyper Journal (HJ)[23] that can contextually manage the web/portal of the journal (descriptive pages, organization, committees, policy, search articles) and the whole work-flow of the jobs from the acceptance of articles till the definitive publication through the web interface. Besides these important functionalities, these softwares are OAI-PMH compatible and they can be managed as real open archive softwares.

The main characteristic of Hyper Journal, an Italian project, is that it allows the end user to have a dynamic contextual visualization of the quoted articles and all those quoting the one that is being read. This is possible thanks to the *sesame*[24] RDF database that realizes a semantic network capable to carry out bibliometrical calculations such as: the number of quotations received by an article or by an author, citation source groupings by journal, by topic, by period.

However, as the present experience of the *Centro di Ateneo per le Biblioteche* (CAB) of Messina, which provides technical support for open access publishing of the journal “Atti della Accademia Peloritana dei Pericolanti – Classe di Scienze MM.FF.NN.”[25] the whole work-flow can be realized manually, without any specific tool, through non automatic updating of the journal web pages and external indexing tools of metadata.

Another supporting strategy to open access is the possibility to host events (conference, workshops) and publish their acts by OAI-PMH compatible softwares. These as Open Conference System (OCS)[26], are able to manage the whole work-flow, the web page/portal, the call for paper, the deposit of contributions and presentations, as OA publishing softwares.

A capital source of information about available resources on the net offering a wide range of softwares, tools, support documentation and open access advocacy, links to Italian and international projects is AEPIC[27] site, a CILEA and CASPUR collaborative project with the aim of promoting open access and planning and implementation e-publishing services for universities and consortium.

Acknowledgment: Ernesto Modica
Thanks to Ernesto Modica for his help in producing the English version of this paper.

References

[1] A detailed and useful report realized by Antonella De Robbio
URL: < http://eprints.rclis.org/archive/00000393/02/nuove_tecnologie_03.11.05.html >.

[2] Respectively in *Centro di Ateneo per le Biblioteche* and *Centro di Calcolo elettronico* of the *Università di Messina*.

[3] As for example the formed postscript, or, more recently, the portable document file (PDF)

[4] «E-prints are the digital texts of peer-reviewed research articles, before and after refereeing. Before refereeing and publication, the draft is called a "preprint". The refereed, accepted, final draft is called a "postprint". The term e-prints include both preprints and postprints.», Antonella De Robbio and Imma Subirats Coll, *E-LIS: an International Open Archive Towards Building Open Digital Librariers,* High Energy Physics Libraries Webzine, 11/2005.
URL: < http://eprints.rclis.org/archive/00004476/01/e-lis.pdf>

[5] BOAI, Budapest Open Access Initiative
URL: < http://www.soros.org/openaccess / >

[6] «Self-archiving can be defined as the deposit of a digital document in a public, free-access repository, for example, an e-print archive. An e-print archive is a collection of digital research documents such as articles, book chapters, conference papers and data sets. », Antonella De Robbio and Imma Subirats Coll, *E-LIS: an International Open Archive Towards Building Open Digital Librariers,* High Energy Physics Libraries Webzine, 11/2005.
URL: < http://eprints.rclis.org/archive/00004476/01/e-lis.pdf>

[7] The GPL, General Public License it is considered from many, if not all, the fundamental license of free software.
URL: < http://www.gnu.org/copyleft/gpl.html >

[8] Primarily LINUX among the operating systems, Apache among the Web Server and MySQL and PostgreSQL among the RDBMS database.

[9] Primarily Java, PHP, Perl and Python.

[10] MARC - MAchine Readable Cataloging, is a standard international servant to facilitate the bibliographical information interchange and the production of bibliographical recordings thanks to the codification of the bibliographical elements and the logical and physics structure of the whole.

[11] Dublin Core Metadata Element Set, is a standard of metadata (still developing and improvement) composed by 15 elements of base also extended to sub elements or qualifiers. Every element is defined using a set of 10 attributes write by the norm ISO 11179.
URL: < http://purl.org/dc >

[12] OAI-PMH, the Open Archives Initiative Protocol for Metadata Harvesting: it furnishes a frame of interoperability independent from the applications and based on the harvesting of the metadata.
URL: < http://www.openarchives.org/OAI/openarchivesprotocol.html >

[13] OAI, Open Archives Initiative: been born and managed by C. Lagoze (Cornell) and H.Van de Sompel (Los Alamos), financed to drawn by various institutions, it develops and it promotes the standards for the interoperability.
URL: < http://www.openarchives.org / >

[14] «A *repository* is a network accessible server that can process the 6 OAI-PMH requests in the manner described in this document. A repository is managed by a data provider to expose metadata to harvesters.», from definition section by protocol description OAI
URL: < http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm #DefinitionsConcepts >

[15] Extensible Markup Language (XML) is to simple, very flexible text format derived from SGML (ISO 8879)
URL: < http://www.w3.org/XML / >.

[16] For example, PLEIADI, Portal for the Literature scientific Italian Electronics on open Archive and Institutional Deposits, realized by CILEA-CASPUR collaboration.
URL: < http://www.openarchives.it/pleiadi >

[17] of the University in Padua. Her scientific production it to E-LIS, E-Prints in Library and Information Science.
URL: < http://eprints.rclis.org / >

[18] The web space is the "physical" space of the publication, nevertheless its different from domain that, can be gotten forwarding an application to the Italian Registration Authority c/o Istituto per le Applicazioni Telematiche del CNR of Pisa.

[19] Forward request to the Centro Italiano ISSN, C/o CNR - DAST II° Sezione di Roma
URL: <http://www.isrds.rm. cnr.it/issn >

[20] It is composed by two parts, a prefix and a suffix. It is assigned by an agency recognized by the DOI Foundation and in Europe the agency is MEDRA where is the Italian association of Publishing, with technical support by CINECA of Bologna. Through the DOI, since it is independent from the URL where are the digital object, it is possible to track the object if update the metadata of the database of MEDRA.
URL: < www.doi.org >, < www.medra.org >, < www.aie.it >

[21] The access to the directory is possible if the journal has a peer-review politics and it is possible through special request available to.
URL: < http://www.doaj.org / >

[22] Open Journal Systems is to journal management and publishing system that has been developed by the Public Knowledge Project through its federally funded efforts to expand and improve access to research
URL: < http://pkp.sfu.ca/ojs / >

[23] HyperJournal software is application that facilitates the administration of academic journals on the Web. HyperJournal Khan be used not only to establish an online version of an existing paper periodical, but also created to an entirely new, solely electronic journal
URL: < http://www.hjournal.org >

[24] Sesame is an open source RDF database with support for RDF Schema inferencing and querying. Originally, it was developed by Gathers as to research prototype for the EU research project On-To-Knowledge.
URL: < http://www.opendf.org >

[25] URL: < http://antonello.unime.it/atti >

[26] Open Conference Systems (OCS) is to free Web publishing tool that will created to complete Web presence for your scholarly conference. It has been developed by the Public Knowledge Project at the University of British Columbia to improve the scholarly and public quality of research online.
URL: < http://pkp.sfu.ca/ocs / >

[27] AEPIC - Academic E-Publishing Infrastructures
URL: < http://www.aepic.it / >