PLEIADI, a portal solution for scholarly literature

Paola Gargiulo*, Susanna Mornati**, Ugo Contino*, Zeno Tajoli**

* CASPUR
Via dei Tizii 6 b 00185 Roma Italy
contino@caspur.it
gargiulo@caspur.it

** CILEA
Via R. Sanzio 4 20090 Segrate (MI)- Italy
mornati@cilea.it
tajoli@cilea.it

Abstract

The PLEIADI Project (acronym for "Portale per la Letteratura scientifica Elettronica Italiana su Archivi aperti e Depositi Istituzionali", a portal for Italian scholarly e-literature in open archives and institutional repositories) originated from the collaboration between two major Italian university consortia, CASPUR and CILEA, within the framework of the AEPIC project. PLEIADI aims at building a national platform that offers centralized access to the scholarly literature archived in Italian repositories.

1   Introduction: OA context and OA archives

In the last decades we have been witnessing a paradox in the scholarly publishing: on one hand the Internet has changed the communication, the exchange and the distribution of information in the scholarly and research environment, the ICT developments continue to offer more advanced and integrated solutions to improve access and circulation, on the other hand several factors hamper a wide and free dissemination of scholarly content.
Most of these factors can be easily summarized in the so called "periodicals crisis" which is seriously affecting the traditional publishing models and is remarkably contributing to the restriction of access to journal literature.
Among the main causes of this crisis, it is worth mentioning the limits of traditional publishing practices, the increasing subscription cost of these journals, the dominant role of few and powerful commercial publishers which own a significant share of the prestigious high quality academic and research journals, apply quite restrictive rules for copyright assignment and grant electronic access based on highly priced business. Moreover the application of new technologies in traditional commercial and not for profit publishing has not meant a full advantage of these technological improvements in the journal output (in most cases the electronic version is merely the equivalent of the print one) or cost reduction of the journal subscription whose cost increase is far in excess of inflation, while library budgets continue to shrink. In this given context some significant initiatives began to assert the need to open access to scholarly publishing (Budapest Open Access Initiative)[1] and to suggest that open access can be achieved through OAI standard compliant open archives, digital repositories where authors can self-archive an electronic copy of their published papers (OAI-Open Archives Initiatives)[2]. These repositories can be institution or discipline based, the former ones collect pre-print or post print scholarly and research content and other type of documents produced by an institution, the latter ones collect high quality scholarly or research content in a certain discipline or in a set of disciplines regardless of the author's institutional affiliation.
Within the OAI framework several tools have been developed/implemented to allow the creation, management, and interoperability of open archives via the OAI-PMH protocol [3]

More recently in various countries academic and research institutions committed themselves by endorsing official statements in support of open access (Berlin Declaration,[4] Scottish Declaration [5]), by implementing digital repositories, by encouraging their authors to publish in open access journals and by mandating the deposition of published material in the institutional archives via enforced institutional policies.

All these efforts result in a growing number of operational digital repositories, in a constant increase of open access content and consequently in a wider dissemination of scholarly content. This situation generates a higher impact of research results and therefore produces vast benefits for authors and for their institutions, from individual career advancing, to further research funding, from increasing each institution's visibility, to promoting general research progress and contributing to the creation of new and unpredictable ways of communication and doing research.

2   Italian Archives

The implementation of Italian digital repositories is still at an early stage.[6] Only a few institutions adopted a policy to encourage self-archiving and strategies for open access to their intellectual output. Three archives (or archive systems) are up and running, though not largely populated yet: at the University of Bologna, Firenze, and Trento. Other installations are still at a test phase, or are devoted to a specific portion of academic output (e.g. doctoral dissertations, such as the archives at the University of Padova and Roma "La Sapienza", or didactic material, such as Parma). A few still contain more metadata than accessible full-text papers, though their intentions are towards a wider accessibility of their materials (University of Messina and SISSA). The European University Institute maintains an archive located in Italy, though strictly speaking it is not an Italian institution.

In the latest months, several initiatives have started, among which 4 new open archives, and there are almost 20 known projects dealing with repositories, both in public universities and research centres. With the support of the Council of Rectors of Italian Universities (CRUI), a national event was held in November 2004 in Messina, where more than 30 institutions agreed on the principle of Open Access to scholarly literature, signing the "Messina Declaration".[7] This event has spread the awareness of Open Access issues and boosted the interest around Open Archives and related technologies for a more effective dissemination of scholarly content. To fulfil this need, and to encourage authors to self-archive their work, a range of services should be available that add value and increase the visibility and usability of this open access literature.

3   PLEIADI – End users services

The PLEIADI Project (acronym for "Portale per la Letteratura scientifica Elettronica Italiana su Archivi aperti e Depositi Istituzionali", a portal for Italian scholarly e-literature in open archives and institutional repositories) [8] originated from the collaboration between two major Italian university supercomputing consortia, CASPUR and CILEA, within the framework of the AEPIC project [9]. PLEIADI aims at building a national platform that offers centralized access to the scholarly literature archived in Italian repositories (figure 1). The service was launched in November 2004, during the Messina event.
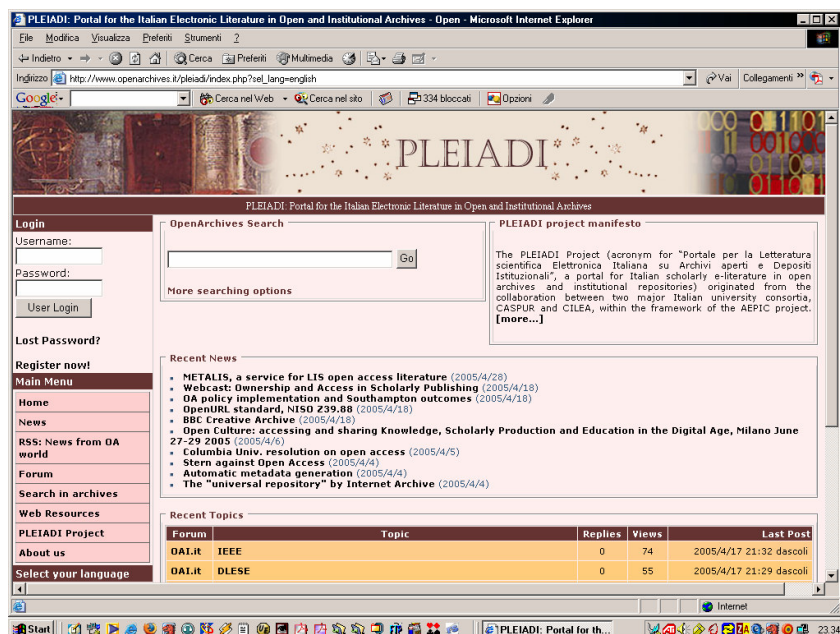
Fig.1 - PLEIADI home page

The main features of the initiative are: discovery and locate, personalized services, visibility and impact, awareness and dissemination, legal protection, preservation. Some of these features are full operational, others will be implemented in the near future. The project aims at allowing the academic and research community to discover, to access Italian OA scholarly production in a structured way, at increasing its visibility and its impact on the Italian and the International academic and research world. Furthermore PLEIADI provides the end-user with a personalized environment offering several services of high interest to the researcher and to his/her work, and it aims at keeping the end-users abreast with what is going in the OA realm and contributing to his/her awareness and support of OA. The personalized services include users profile creation, alerting services, search and citations savings, personalized news services (news, forum, RSS).

Other services such as citation parsing, statistical information on search results and access to documents, export facilities to build individual and/ or institutional web pages for presentation or assessment and to build bibliographies will be implemented in the near future. Services related to time stamping, copyright protection, legal deposit, and preservation issues will also be developed in the future.

Last but not least PLEIADI wants to promote awareness of OA issues in Italy, to contribute to the debate on OA and to the advancement of technical knowledge and expertise in this field by offering its forum as an arena for exchange information among the OA community members (from university and research centres policymakers, researchers, IT people, librarians, students, press, etc.).

PLEIADI platform is made of a two-level architecture. Under the portal visible layer, which was developed using XOOPS [10], an *open source* framework to build up *content management oriented* portal systems, there is a complex infrastructure of service providers. PLEIADI portal offers a federated search interface and a series of user-centred services that supplement the platform offer. It is possible to identify several functional blocks in the portal architecture:

- a ***user area***, comprising user's front-end, multi-language support and authentication/authorization system. This area represents the personalized environment where the user creates its own profile, activates different features from

language selection, alerting services related to News, FAQ, Forum, search savings, citation savings *information area*, based on News, Forum and RSS blocks, together with FAQ, useful links and document sections. This area aims at collecting and disseminating information on OA issues

- a *search & retrieval area*, encompassing the Z39.50 connection with the *harvester* and a sorter module for retrieved documents. This area manages search performance in the Italian archives, two search modules have been designed, a simple one and an advanced one, the latter encompasses different search fields, filtering functionalities, search results sorting, etc.
- a *management area*, with an intuitive *content management engine*, this area handles the all the website contents and web presentations through an intuitive and simple to use graphic interface.
- a *log & stats area*, for useful statistical information on search activities and access to documents in open archives. This area has not been implemented yet.

The technological side of the portal services is constituted by several PHP [11] modules developed within the Xoops community. Their core code was adapted to meet accessibility needs that characterize PLEIADI web site. A web GUI is available for the administration and management of most modules.

XOOPS is an open source software designed for web portals. As several other similar products, it is based on PHP language, MySQL[12] as back-end database, and Apache technology [13] for the web server. PLEIADI developers chose XOOPS to build the web site due to several reasons:

- availability of high-level portal management functions
- use of style-sheets technology to implement dynamic web pages: XOOPS is based on an open-source template generator called SMARTY [14]
- complete reciprocal independence of the different PHP modules from their HTML mark-up, consequential to the use of style sheets

4   PLEIADI – Service provider features

PLEIADI as a service provider [15] performs a series of back-end activities designed to discover, retrieve, access Open Access content from all Italian Open Archives. Metadata related to OA content are harvested from institutional or disciplinary repositories (also called data providers), where they are deposited by authors according to their affiliation or scientific interest.

Once harvesting, filtering, harmonization and indexing activities have been performed, a new collective database becomes available and searchable via HTTP and Z39.50. Its availability through the Z39.50 interface enhances the opportunity for the OA content to be exposed and searched by the various federated searching tools.

To allow easy search and display, a WEB interface has been built embedding the Z39.50 communication channel with the harvester, using PHP as scripting language and the related YAZ functions library [16]. Predefined query templates and query results parsing, filtering and ordering functionalities have been added to the interface in order to get full benefits from the harvesting subsystem.

The metadata are harvested via the OAI-PMH protocol, through the harvester Celestial.[17] Some modifications have been performed on the standard installation of Celestial. A feature for selective harvesting by sets and a web interface to manage it have been added. Thanks to

this feature it is possible to harvest only Italian institutional works from repositories that collect content from other countries and other languages.

As new archives are launched they are added to the list. Metadata are harvested in the OAI Dublin Core format [18]. This schema has the same semantics of 'unqualified Dublin Core'.[19]

PLEIADI also uses metadata present in the 'header' and in the 'about' sections (as described in the OAI-PMH standard) of the harvested XML format.

Another added feature is the integration of crosswalks in Celestial. This feature improves Celestial function as a cache of metadata. The standard Celestial installation re-exposes the same metadata it harvested; with this feature Celestial exposes the converted metadata, in order to facilitate their possible reuse by different service providers.

All metadata harvested from different repositories require harmonization, so they are converted in an internal schema through specific crosswalks. Several fields have the same semantics of Dublin Core unqualified or a similar one. The element of this internal schema that requires more work is dcsubject: it contains Italian MIUR classes.[20] For every archive there is a specific conversion tool from their original data in dc:subject.

Indexes are built from metadata in an XML format, using Cheshire2 as indexer software.[21] Cheshire2 allows to install a server that works with two protocols, http and Z39.50. This has been the reason why Cheshire2 was selected as indexer. At the moment Cheshire is used only with the Z39.50 protocol.

The end-user search functions are implemented through a specific portal interface, developed using PHP and based on YAZ libraries [for the translation of http queries into Z39.50 standard. The interface functions allow to define several search criteria, and a classification filter for the search results. This is obtained directly through the Z39.50 query, using the dcsubject datum within the XML payload. Data within the Z39.50 payload package, formatted in XML, are passed to a PHP module that has been developed to allow different sorting criteria, namely Author/Title/Year. This module is based on DOM libraries [22].

Next developments will concern new user services linked to the search functions: "search history", "saved search strategies" and "my articles" areas, to allow registered users to navigate through previous searches, to access their preferred articles in a reserved area, to receive search results from automatic search processes based on customized search strategies. User areas will be stored in a database, and the module will be built with MySQL back-end for compliancy with XOOPS environment, to be possibly reused in a general-purpose context.

5   Bibliographic resources integration: OAI-PMH and Z39.50

The release of the OAI-PMH protocol in 2001 opened the way to the development of a variety of tools for metadata sharing across different platforms, applications, installations. Wherever a document is deposited, its potential readers are able to locate it via centralized search facilities known as service providers.

PLEIADI constitutes an attempt to exploit the potential of the new OAI-PMH together with the well established Z39.50, a model to build centralized search facilities for several categories of information resources. Academic e-publishing initiatives, particularly open-access journals and conference proceedings, would benefit from an integration with other information sources, such as the library catalogue and bibliographic references from databases and commercial publishers, that constitute the traditional core of information

provision in universities.

E-prints, dissertations, data sets and all the material deposited in open archives can contribute to extend the amount of information available to academic users, especially if they can be retrieved via the same resource discovery systems that are already in use to disseminate traditional works. Also cross-institutional collaboration may be encouraged through metadata exchange, resulting in a wider dissemination of relevant scholarly content.

Metadata crosswalks, combined with OAI-PMH harvesting and Z39.50 search, offer a possible solution for bibliographic resources integration.

References

[1] Budapest Open Access Initiative (BOAI), http://www.soros.org/openaccess, April 29th 2005

[2] Open Archive Initiative (OAI), http://www.openarchives.org/, April 29th 2005

[3] OAI- PMH, http://www.openarchives.org/OAI/openarchivesprotocol, April 29th 2005

[4] Berlin Declaration on Open Access to Knowledge in Sciences and Humanities, http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html, April 29th 2005

[5] Scottish Declaration on Open Access, http://scurl.ac.uk/WG/OATS/declaration.htm, April 29th 2005

[6] The complete list of Italian open archives, http://www.openarchives.it/, April 29th 2005

[7] Gli Atenei italiani per l'Open Access: verso l'accesso aperto alla letteratura di ricerca, Workshop, 4-5 November 2004, Università di Messina.http://www.aepic.it/conf/index.php?cf=1, April 29th 2005

[8] PLEIADI, http://www.openarchives.it/pleiadi/index.php?sel_lang=english, April 29th 2005

[9] AEPIC, http://www.aepic.it, April 29th 2005

[10] XOOPS, http://www.xoops.org/, April 29th 2005

[11]] PHP, http://www.php.net/, April 29th 2005

[12] MYSQL, http://www.mysql.com, April 29th 2005

[13] APACHE, http://www.apache.org/ April 29th 2005

[14] SMARTY, http://smarty.php.net/, April 29th 2005

[15] The whole system architecture is similar to the one used by the ePrints UK Project. We are grateful to UKOLN and RDN for support in implementing this solution, http://www.rdn.ac.uk/projects/eprints-uk/, April 29th 2005

[16] YAZ, http://zoom.z3950.org/], April 29th 2005

[17] CELESTIAL, http://celestial.eprints.org, April 29th 2005

[18] OAI Dublin Core: http://www.openarchives.org/OAI/2.0/oai_dc.xsd., April 29th 2005

[19] DCMI (2004), *Dublin Core Metadata Element Set, Version 1.1: Reference Description*: http://dublincore.org/documents/dces/, April 29th 2005,

[20] MIUR- The Italian Ministry of University and Research has issued a discipline based classification scheme to define syllabi and careers This scheme is adopted by most of the Italian institutional repositories http://www.murst.it/atti/2000/alladm001004_01.htm,, April 29th 2005

[21] Cheshire II Project, http://cheshire.berkeley.edu/, April 29th 2005

[22] DOM http://it2.php.net/manual/en/ref.domxml.php, April 29th 2005