Alma-DL, the Digital Library of the University of Bologna: a case study

Marialaura Vignocchi, Simone Sacchi, Fabrizio Morroia; C.I.B Bologna University; Bologna, Italy

Abstract

Alma-DL, Alma Mater Studiorum Digital Library, is a complex project started by CIB (Inter-library Centre) of Bologna University late in 2001, with the aim to develop an organizational and a technological infrastructure to collect, organize, archive, integrate and offer an access to the digital contents that our University acquires or produces and makes available primarily to institutional users and when possible worldwide. Alma-DL has completed its beta-test phase and can now evaluate its first outputs. This contribution intends to focus on the two subprojects that represent more original outputs of the digital library project that is the OAI compliant platforms developed to collect, archive and distribute the digital contents produced by the University of Bologna. The DigLib: digitization of the cultural heritage of Bologna University and the Open Access Institutional Repositories and E-Journals projects offer paradigmatic solutions in the field of digital image and file archiving, metadata storage and exchange, e-content management, integration and fruition. Finally, the contribution points to immediate future action plans that have been devised as further steps of the development of the Digital Library services.

Alma-DL, *Alma Mater Studiorum Digital Library*, is a complex project started by CIB (Inter-library Centre) of Bologna University late in 2001, with the aim to develop an organizational and a technological infrastructure to collect, organize, archive, integrate and offer an access to the digital contents that the University acquires or produces and makes available primarily to institutional users and, when possible, worldwide.

To facilitate management and organization, the *Alma-DL* project was split into five sub-projects that have been pursuing specific objectives whose coordination was ensured and monitored by a central project leader. The sub-projects have covered the following areas: institutional repositories and e-publishing, digitization, archiving and web publication of valuable collections owned by the University libraries, digital collections development and preservation, resources integration and search, information literacy and support services for users, statistical monitoring, users authentication systems and fault tolerance mechanisms.

Alma-DL has completed its beta-test phase and can now evaluate its first outputs. The University provision of digital contents has been enriched by a number of new commercial collections of e-journals and books. A linking system between digital resources and services has been implemented using open-url technology and personalizing a commercial knowledge base while a meta-search interface to search across all resources included local and national OPACs has been developed with the contribution of a local private software company. Our portal website has been revised in order to provide an aggregated and friendly gateway to OPACs, METAOPACs, digital resources, news, information, services and external links. Users instruction courses have been organized at all levels, on-line tutorials have been made available through the portal website and a collaborative reference service has just been started. However, more original outputs of the digital library projects are represented by the OAI compliant platforms that have been developed to collect, archive and distribute the digital contents produced by the University of Bologna.

The scientific and cultural patrimony as well as the current outputs of research and teaching activities of the University are of great importance and value. The University of Bologna is recognised as the oldest and perhaps the first university in the western world. Its foundation year dates back to 1088. Famous students included Thomas Beckett, Dante, Petrarca and Copernico while important scholars and scientists like the jurist Irnerio, the naturalist Ulisse Aldrovandi, the medicine scholar Malpighi brought prestige to the "Studium bolognese". The collections of rare and ancient books, codices and manuscripts of the University Library and of many faculty and department libraries document the historical importance of the University scientific and cultural heritage. Nowadays the University of Bologna offers a multi-campus structure which is divided into 23 faculties, 68 departments and 5 university campus branches within the cities of Bologna, Cesena, Forlì, Ravenna and Rimini (www.unibo.it). Structured academic staff amounts to almost 3000 professors and researchers and students are about 100.000. Research production has obtained high scores in almost every discipline in the recent Italian RAE (Research Assessment Erxercise) and annually accounts for about 40.000 entries, mainly journal articles, in the University Research Registry. It is not possible to quantify the great amount of teaching materials created for courses that so far have been sparsely distributed in faculty offices, photocopying centres, the professors' personal web sites and university libraries.

To collect, preserve and make available this conspicuous patrimony CIB, the university centre for library automation has developed the *DigLib: digitization of the cultural heritage of Bologna University* and the *Open Access Institutional Repositories and E-Journals* projects which offer paradigmatic solutions in the field of digital image and file archiving, metadata storage and exchange, e-content management, integration and fruition.

The objective of the *DigLib* sub-project has been to provide the University of Bologna with a competent technical support and with an efficient technological infrastructure to carry out the digitization of culturally and scientifically significant, often unique and rare, materials owned by the University libraries. *DigLib* has already completed several digitization projects and many others are still in progress.

The projects carried out so far cover a wide range of typologies and disciplines. The project team has digitized and published the digital edition of printed books, manuscripts and journals dating back to 16th and 17th centuries, modern scientific journal, a collection of 19th century students' periodicals, architectural tables for as much as approximately 150.000 pages. Among these, some deserve special mention for their unique relevance and value such as the digitization of the University

Library owned watercolored copies of the books by the 16th century naturalist Ulisse Aldrovandi; a collection of 17th and 18th centuries books and documents on the longest meridian line in the world build by astronomer Gio Domenico Cassini in S. Petronio Basilica in Bologna; the 18th century complete collection of *Commentarii* of Bologna Science Academy; *Scientia*, an important "Bolognese" scientific journal that published one of the first articles written by Einstein on the Relativity theory in 1914; a collections of architectural tables by famous architects like Otto Wagner.

As in the case of the digitization of the Ulisse Aldrovandi's printed works and manuscripts owned by the University Library, which will be used to create the first complete critical edition of the famous naturalist with the support of the Italian Ministry for Cultural Heritage and Activities, all *DigLib* initiatives reflects not only local needs but national and international recommendations to maximize access to collections while preserving originals and contributing to a distributed network for the integration of digital resources thanks to system interoperability and metadata exchange. *DigLib* may be considered part of the BDI (Italian Digital Library) Project that has been developed by the Italian Ministry for Cultural Heritage and Activities within the European Commission Minerva project.

The Minerva project inscribes itself in the general objectives of *eEurope Action Plan* that in 2000 recognized the strategic and public value of cultural and scientific heritage for European Information Society. The representatives and experts of all European Member States met in Lund in 2001 where they agreed on a series of objectives and principles as to establish a European coordination of digitization policies and programs within specific technical guidelines and recommendations that guarantee quality and interoperability of digitization outcomes in order to provide an integrated access to all digital resources. In Italy the BDI project has first of all carried out a survey of all national digitization initiatives that should become accessible through a national web portal. BDI has also carried out a study on metadata sets providing an original solution with the MAG dataset that combines DC elements with technical data on imaging.

The *DigLib* project team consists of librarians and computer science professionals and engineers who have been working together to establish policies, technical requirements and procedures that together with the implementation of an on-site developed technological platform permit to control the workflow of each digitization initiative.

The digitization project includes all the necessary steps from selection and analysis of originals to digital acquisition, from descriptive metadata attribution to final publication of the digital editions through the web.These steps include:

- high-definition digital photography of the original
- creation of the image back-up copies
- high-definition original compression
- metadata collection through management software
- metadata storage in the repository

At an early stage, the scanning of the originals was carried out in outsourcing by specialized professional companies. Yet, the control and finishing activities, as well as the qualitative differences among the results due to diverse equipments and working methods were so time-demanding that the project team has considered the purchase the scanning equipment and has employed dedicated staff. This choice has allowed the speeding up of the acquisition process and control phases nad finally the improvement of the image quality.

The equipment includes a Jenoptik Eyelike DCS camera with AF-MICRO NIKKOR Nikon lenses, the apropriate cold lighting and the required Machintosh G4 to support the camera. This is all stored in an air-conditioned room with monitored lighting which allows the preservation of the originals, the correct functioning of the digital optical sensors as well as tone uniformity of the scanning light. This latest aspect is vital to ensure colour correspondence which may be compromised by light variations. The working schedule of the scanning staff is limited to 4 hours maximum not to compromise the scanning quality and limit mistakes in the file nomenclature. Further necessary care when treating antique books may include page scanning with a 90°-opening not affect or strain the bindings, and the use of gloves to protect the originals from hand skin secretions.

The following step consists in the creation of the back-up copies made on high-capacity magnetic tapes, exploiting SDLT technology using the Tandberg SDLT 600 magnetic tape recorder. At least, two copies of each work are usually recommended. One is kept in staff offices while the other is stored in a building specifically equipped with fireproof and air-conditioned cabinets for the stocking of magnetic storage supports.

Compression can be carried out both in JPG and DJVU [1] format. The first is a standard format for web diffusion that can be easily displayed by any browser without plug-ins. The second format allows a much more fine and powerful image view management. The main features of this format consist in the compression degree, allowing the web publishing of images in their original capturing size without affecting quality, and the possibility of carrying out optical characters recognition on digital images. This provides good results only if the text features modern printing standards and is in good conditions; this is why it was not applied to antique books and manuscripts. For the latter, compression parameters were aimed at obtaining a better graphic correspondence between the originals and the compressed image. Moreover, successive OCR, whose settings affect displaying quality, was not required. The free viewer allows to activate view management tools like zoom, image rotation and full-text search in the whole file.

The metadata used for repository structuring follow the MAG [2] standard developed by BDI. This standard consists of three main groups of metadata: bibliographic, structural and administrative. Bibliographic metadata describe the originals using the main 15 elements of the Dublin Core. Structural metadata track the relations among the various parts of the works like the division of the journals into volumes and issues and allow the definition of description labels associated to each part. Administrative metadata contain technical data about the digital objects and the system used for digitization: image size, format, model and manufacturer of the equipment employed.

The management software used for metadata collection has been developed by CIB staff. It allows the manual data entry of the data that cannot be collected automatically by means of a graphic user interface. Administrative image information and structural information are automatically collected. Structural data are modeled following the hierarchy according to which the image groups referring to one part are stored in the file system. Once

collected, the metadata are saved in XML text format and, only after a further control, saved inside the repository. During the information rewriting inside the repository, the records harvested by OAI-PMH [3] queries are automatically filled-in. The use of phpoai2 [4] instrument, and working as a Data Provider for static OAI repositories, allowed *DigLib* validation and registration as OAI standard compliant repository (protocol 2.0).

Web browsing is made possible by a simple PHP interface that permits to visualize the various parts of the originals or to choose a specific part of it by means of a scrolling list menu that contains all the labels associated to the various parts of the original work; making thus possible a direct access to the selected parts. The web interface permits also to choose the visualization format of the images. Partial bibliographyc information are showed on the interface creating thus a link to the local OPAC where the complete record of the browsed work can be obtained.

Originally the whole system was supported by Open Source software and developed on Linux platform, later on the project team has decided to replace the DjvuLibre Open Source project with the commercial Document Express DJVU creator developed by LizardTech. This choice has greatly improved the quality of our Djvu images and has allowed to perform the OCR process on the images. The introduction of a commercial software solution has implied the introduction of a Microsoft Windows platform in our system since the promised Linux version of Document Express has not yet been realesed. Most of the system, however, remains supported by Open Source software.

The JPG compression script is developed in Perl and uses the ImageMagick library to perform the conversion. The whole management software is also written in Perl and the repository is based on MySQL database. We keep two copies of the repository: one interfaced with the management software for metadata entry and manipulation, and the mirror copy interfaced directly with the web interfaces and OAI Data Provider.

All the production environment is hosted on a SUN Cluster and provide the remote replication of the whole system allowing a optimum degree of fault tolerance and services stability.

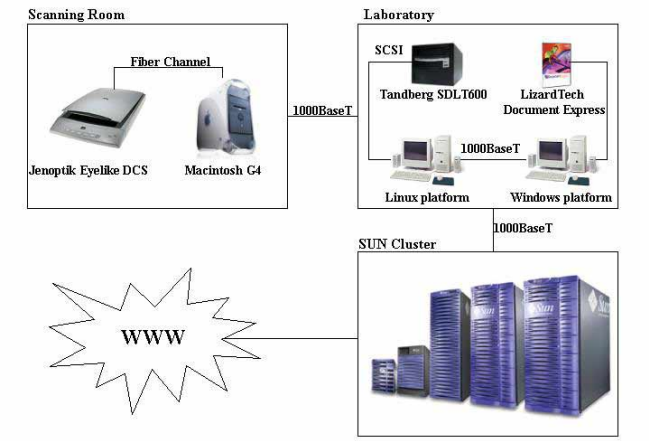

Figure1. *DigLib System Description*

The *Open Access Repositories and E-Journals* sub-project has provided the researchers of Bologna University with new options for the dissemination and communication of research and teaching materials. The project group has customized GNU EPrints software to implement three institutional repositories - AMS Acta , AMS Miscellanea and AMS Campus – and, up to now, a born-digital journal *RPD Ricerche di Pedagogia e Didattica* edited by the Education Department and a digital version of a current journal in print, *Statistica*, edited by the Statistical Sciences Department.

The institutional repositories are dedicated to and can manage different types of documents. GNU EPrints customization has implied the add-on of new features, a better integration with other components of the University information system and the graphic harmonization of the user interface with the University brand and style.

AMS Acta [5] is the Institutional Repository for scholarly communication and makes freely available the results of research carried out at Bologna University in all fields. It has been configured and customized to contain pre-prints, post-prints and e-prints. It provides our researchers with a simple, yet powerful, instrument to submit, preserve and grant access to their scientific papers as it is full OAI-PMH compliant and part of the Open Access community.

AMS Acta is rather paradigmatic of the status of deployment of IRs in Italy. In Italy, as in other European countries, Open Access principles are considered to have been at the basis of institutional repositories setting up, according to the results of the international survey on IRs carried out by JISC, SURF and CNI in the Spring of 2005 [6].

In general Open Access principles have obtained a good institutional acceptance. In Italy, thanks to the advocacy carried out by a voluntary task force of librarians, 74 out of 77 universities have signed the Berlin Declaration on Open Access to Knowledge in the Sciences and the Humanities. Recently a sub-group of experts appointed by the Library Commission of the Council of Italian University Rectors has officially presented a document on Open Access issues that recommends practical actions according to the roadmap presented during the Berlin3 follow up Meeting held in Southampton in February 2005.

Yet, despite institutional awareness and advocacy initiatives, no Italian university has so far elaborated an official policy in favor of Open Access and at a national level it seems very difficult to achieve a coordination given the lack of a common table of discussion for all the stakeholders. As a consequence IRs grow rather slowly even though steadily. The IRs so far registered in the Italian service provider PLEIADI (Italian Portal to electronic scientific literature in IRs) [7] managed by CILEA, one of the consortia that provide Italian universities with innovative technological services and solutions, are 10, two of them belonging to Bologna University. The average number of documents in the repositories is about 400.

AMS Acta was launched halfway through 2004 and the amount of submitted papers has grown steadily since then. Statistics of March 2006 provided by the Registry of Open Access Repositories (ROAR) [8] show that it gives free access to 1900 documents and it is by far the most active IR in Italy.

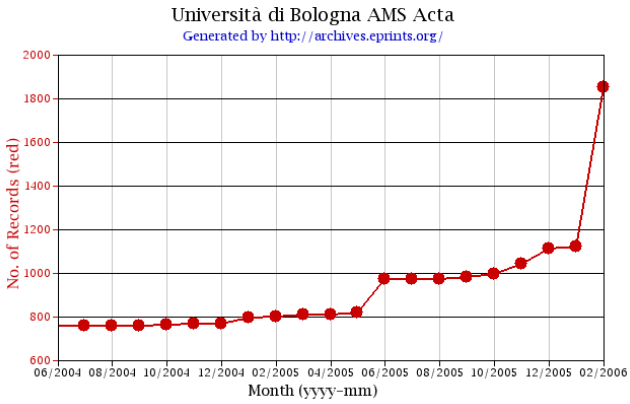

Figure 2. *AMS Acta growth graph*

Yet, the amount of documents is not significant if compared to the total production of research outputs of the institution and the peak in the submission trend that can be seen in figure 1 is to be attributed to a massive data entry of older materials edited by a University-based Scientific Society.

In fact, the faculty feed-back is poor. While adhering to Open Access on a principle basis, few academic authors use IRs. Looking for reasons, a number of surveys have investigated the authors' attitude towards scholarly communication system and Open Access. The answers to the questionnaires show that inertia, ignorance of existing options, quality and copyright concerns are still the obstacles to the academic authors' full adoption of Open Access alternatives [9] [10].In particular, intellectual property right issues and management have appeared to be major concerns with academic staff of Bologna University.

In response to the authors' fears and doubts, a legal advice has been asked and the self-archiving procedures have been completely revised in accordance with regulations in force. After the revision that has just been completed, authors have to sign and return a written contract – once for all - in order to be admitted to the self-archiving service. It is a non-exclusive contract that identifies the copyright owner as the copyright holder and states his responsibility for the contents that he wants to make available through the repository web site. It provides the conditions and duration of the service, and states the role of the IR manager as mere service provider with no responsibility or rights on the contents or their forms. It includes a list of "use conditions" which are presented as click wrap licenses in the course of the self-archiving procedure and are attached to each e-prints document, clearly stating the admitted and authorized uses of the contents. The contract includes a list and a description of other services offered to academic authors such as book-on-demand and the deposit in the Central National Library in Florence which are both presented as click wrap options during the self-archiving on-line procedure.

Thanks to these additional services, original and unpublished works have the possibility to reach a sort of official published status thanks to the IR infrastructure. The deposit in the National Library is required for legal attribution of date of publication and intellectual property. Unfortunately the deposit of digital documents distributed on the Web has been included in a recent law Act that is still lacking its set of rule to come into force. So far the deposit in the National Libraries is carried out on a voluntary and experimental basis by some universities that have signed an agreement with the National Library which harvests their sites on request. The only legal deposit requires a print version of the original work. This is one of the reasons why a print-on-demand service has been activated.

Proper technical measures of digital rights management have not been implemented in the IR since GNU Eprints has been designed to contain Open access full-texts. Consequently, authors can activate protections on their own .pdf. only. Nevertheless this new concern for legal correctness that it has becomes visible on the site thanks to a number of disclaimers and licenses statements may hopefully change the academic authors' perception of IRs as unregulated and potentially subject to abuse and infringements. The new Acta is still at an early stage to make possible an evaluation of its outcome but hopefully the new registration procedure may help authors consider it a trusted archive not only on account of its technological reliability.

GNU EPrints customization has also achieved a better integration with other services offered by our university. Different user registrations have been added to and have been integrated via LDAP with our University Active Directory Service that contains a built-in anagrafe of university researchers. In this way academic authors may access the repositories with their institutional usernames and passwords, making log-in procedure friendlier.

AMS Campus [11] is the institutional repository that contains and give access to teaching packages prepared by lecturers for courses. It is based on a net of distributed autonomous repositories (one per Faculty) that can be searched through a unique OAI interface. Each AMS campus IR is totally indipendent and teachers and students can register and access to the service according to their academic status. In the case of AMS Campus, integration with the University Active Directory Service allows not only a single sign on to the system through institutional username and password, but also a selective access to full-texts according to faculty affiliation. As a consequence the faculties that do not want to make teaching materials freely accessible in the Internet can choose to limit access to the full-text to registered students only, while still making metadata freely searchable through the common OAI interface.

AMS Campus presents a working-area interface for documents submission and a search and retrieve interface for students totally customized to deal with teaching and learning materials. The organization of navigation and the search options reflect the students' need to retrieve the files according to course titles, teachers and academic year. AMS Campus was born late in 2002 and as the chart shows, its submission trend is growing fast, considered that self-archiving is voluntary. As far as March 2006, about 300 teachers have self-archived more than 700 teaching packages.

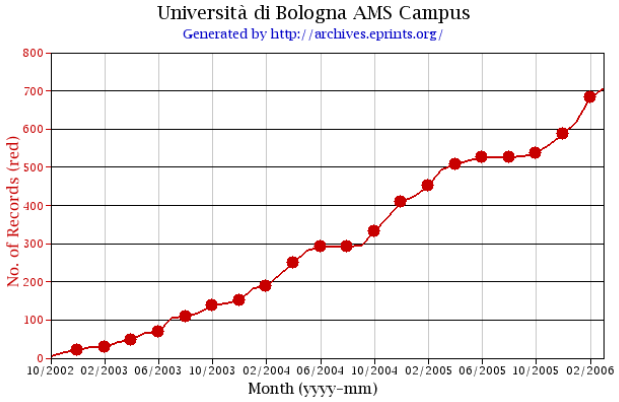

Figure 3. *AMS Campus growth graph*

Finally, AMS Miscellanea [12] is Bologna University institutional repository for cultural contribution not strictly related to research and teaching activities. It shares the customization with AMS acta but in a more free and open fashion. Submission is not subjected to any validation, either formal or substantial, and outputs are totally under authors' responsibility. Due to his very generic nature independent from any institutional activities, Miscellanea is practically empty.

From the analysis of submission trends of the different IRs of Bologna University, it is easy to come to the conclusion that the success of an IR depends on its positioning in the general workflows of the institution and on its immediate practical utility. The distributed solution adopted for AMS Campus combined with the possibility of graphic customization of web interfaces is being appreciated by Heads of Faculty who can well control and integrate their teaching materials management platforms into the faculty information systems and portals. AMS Acta seems to suffer from its ideological affiliation. As the results of the first international survey on IRs show [13], most IRs may never move forth from mere "political" icons to concrete tools for the research community. In such a respect it becomes crucial for those in charge of IRs to gain the support of institutional stakeholders by demonstrating how IRs can integrate with, support and add value to the activities of the institution itself.

It is necessary to re-think IRs in a more pragmatic way to achieve a repositioning of their services within the more general workflow of the institution. Library automation centers in charge of the IRs should seek an interoperation with other components of the general information system of the institution that could easily and profitably re-use data and contents contained in the repositories. In a context characterized by a growing demand for accountability, scale economies in data management and use may save resources that can be more efficiently and effectively employed otherwise.

CIB has been seeking the collaboration of the University Research Management Unit in order to integrate AMS Acta with the research management and information system. The latter is currently under revision to respond to an increasing need for national coordination of the research evaluation system which has been lately modeled on the British RAE and on which have come to depend the selective distribution of public funds. This integration will provide researchers with a single workflow for submission of publications data and full-texts to the repository and registration for evaluation exercise by local academic committees and national panels that grant research and institution funding. Through a cooperation with library services the research evaluation management system can improve the quality of data necessary to document research outputs. By becoming an essential part of the organizational workflows library services like IRs may be no longer perceived as a supplementary burden by faculty.

Immediate future planning includes the development of a common search and retrieval interface for all our repositories via OAI-PMH to aggregate them with all other electronic resources, both free and acquired that make up the rich patrimony of the University Digital Library. In general, integration and inter-operability seem to be the keywords for our way forward. Lorcan Dempsey, reporting on IRRA project, has affirmed: "As more of our working, learning and playing lives moves onto the network we need better workflow support" [14].

Networkflows is the key concept that Lorcan Dempsey suggests in order to start thinking about IRs and, in general, about library services anew. To consider networkflows means to come out from a library-centered universe and enter an open space with no centers - a space dominated by flows that meet or cross forcing us to reinterpret the role and the significance of services, even traditional ones, accepting their plurality. To regard our digital collections as part of apparatus that support research and teaching activities implies thinking pragmatically upon possible solutions of integration, which is perhaps the way forward to institutional legitimacy and recognition.

References

[1] Léon Bottou, Patrick Haffner, Yann Le Cun, Paul Howard, Pascal Vincent. DjVu: Un Système de Compression d'Images pour la Distribution Réticulaire de Documents Numérisés. (DjVu: An image compression system for distributing scanned document on the Internet). Actes de la Conférence Internationale Francophone sur l'Ecrit et le Document, Lyon, France, July 2000.

[2] Progetto MAG - Biblioteca Nazionale Centrale di Firenze. <http://www.bncf.firenze.sbn.it/progetti/mag>

[3] The Open Archives Initiative Protocol for Metadata Harvesting. <http://www.openarchives.org/OAI/openarchivesprotocol.html>

[4] Heinrich Stamerjohanns, University of Oldenburg. <http://physnet.uni-oldenburg.de/oai/ >

[5] Almae Matris Studiorum Acta <http://amsacta.cib.unibo.it>

[6] Clifford A Lynch, Joan K Lippincott Institutional Repository Deployment in the United States as of Early 2005, D-Lib Magazine 11;9 Sept. 2005.

[7] PLEIADI <http://www.openarchives.it/pleiadi>

[8] Registry of Open Access Repositories <http://archives.eprints.org/>

[9] JISC, OSI, Journal authors survey: report. [Web document]. 2004 Feb. <http://www.jisc.ac.uk/uploaded_documents/JISCOAreport1.pdf>.

[10] A.Swan, S. Brown, Open access self-archiving: an author study. [Web document]. May 2005. < http://cogprints.org/4385/01/jisc2.pdf>

[11] Almae Matris Studiorum Campus <http://amscampus.cib.unibo.it>

[12] Almae Matris Studiorum Miscellanea http://amsmisc.cib.unibo.it

[13] Gerhard van Westrienen, Clifford A. Lynch Academic Institutional Repositories: deployment status in 13 Nations as of Mid 2005, D-Lib Magazine 11;9 Sept. 2005.

[14] Lorcan Dempsey, Networkflows. Digital asset management, Research, learning and scholarly communication, User experience,

Lorcan Dempsey's weblog. Jan. 28, 2006,
<http://orweblog.oclc.org/archives/000505.html>.

Authors Biographies

Marialaura Vignocchi is Head Librarian in charge of the Digital Library services of Bologna University. She spent 12 years as senior librarian responsible for users' services in an academic central library. She recently took part to an IFLA meeting in Oslo with a contribution on new trends in scholarly communication.

Morroia Fabrizio is Technical Project Magager of the DigLib digitalization project of the Digital Library services of Bologna University.

Simone Sacchi iswebmaster of Alma-DL and is in charge of all Open Access technical aspects of Alma-DL Digital Library services of Bologna University. He is involved since early 2000 in Open Access applications and projects.