OPEN ARCHIVES FORUM: PROJECT DELIVERABLE

Project Number:	IST-2001-32015
Project Title:	Open Archives Forum
Deliverable Type:	Public

Deliverable Number:	Project Deliverable D4.1
Contractual Date of Delivery:	March 2002
Actual Date of Delivery:	August 2002
Title of Deliverable:	"Creating a European Forum on Open Archives Activities" Workshop Report of the 1st Open Archives Forum Workshop, Pisa, 13-14 Maggio, 2002.
Workpackage contributing to the Deliverable:	WP4
Nature of the Deliverable:	Report
URL:	http://www.oaforum.org/documents/d41workshop1a.php
Author:	Donatella Castelli with Leona Carpenter, Susanne Dobratz, Philip Hunter
Contact Details:	Donatella Castelli, +39 050 3152902

Abstract	This report documents the First Open Archives Workshop held in Pisa (Italy) on 13th -14th of May 2002.
Keywords	open archives, Open Archives Forum, Open Archives Initiative, e-print archives, data archives.

TABLE OF CONTENTS

1 GENERAL OVERVIEW

The aim of the first OA-Forum Workshop was bring together researchers, technical implementers and project managers who are experimenting, or are willing to experiment, with the open archives approaches. The Workshop wanted to set up the basis of a European forum for sharing experiences and solutions, and encouraging networking among projects.

The interest about the open archives approach has been growing out since 1999 when the Open Archives Initiative (OAI) started discussing this approach within the e-print community. In order to support interoperability among open archives OAI has proposed a technical solution based on the use of a harvesting protocol that allows gathering metadata records from heterogeneous archives in a uniform way. At the present this is the most widely known and discussed solution for supporting the interoperability among repositories of electronic documentation. Many e-print archives have implemented/are implementing this the OAI-PMH protocol and many people are ready to discuss and report about their experiences. It is was obvious for the OA-Forum to initiate the series of its workshops with presentations of the OAI solution and of the experiences related to the implementation of OAI-compliant archives and services.

In parallel with the monitoring of the European OAI related activities, the OA-Forum also intends to pay attention to the solutions for interoperability implemented by different communities for other types of archives. In this first workshop we started this exploration process by investigating what is done within the astrophysical community. This is a small, but strong, community that uses widely both data and e-print archives. There are several research and economic motivations that push the members of this community to share the content of their archives. The interoperability between these archives is based on de-facto standards that are successfully used since several years. Very recently, however, some work has been done to take into account the current development of general standards and to create a link with other disciplines. It is certainly interesting for the OA-Forum to follow the evolution of this new trend.

The Workshop was organised in presentations and break-out sessions in order to better achieve the workshop objective, i.e. facilitate the creation of a discussion forum. Four break-out sessions were set up to discuss relevant issues and the invited specialists were asked to keep into account these issues in their presentations. In addition to the formal opportunities, coffee breaks, lunches, and the workshop dinner, were organised in such a way to provide the best conditions for facilitating networking and dialogue between participants.

The workshop was slightly oversubscribed, with more than fifty registered participants attending, along with six invited speakers and nine OA-Forum project workers. Eleven European countries were represented: Belgium, France, Germany, Italy, The Netherlands, Norway, Portugal, Spain, Sweden, Switzerland and the United Kingdom. The subject specialties included were agriculture, law, space, medicine and astrophysics. The types of organisation sending representatives included archives, universities, research bodies, national and research libraries, commercial organisations, and computing centres.

Our general impression is that we have achieved our goals. This first opportunity for interested parties and OAI implementers in Europe to meet under the banner of the Open Archives Forum was highly successful in the degree of participation, and smoothly run thanks to the support of CNR staff on the ground in Pisa. We were confirmed in our impression by comments received through the evaluation questionnaire, as well as by several e-mails we received after the workshop (e.g. "... just a quick note to say how much I enjoyed

the workshop in Pisa... thank you and all your colleagues for arranging such an enjoyable, informative, and well-organized event"; and "Congratulations for the workshop of Pisa. It was a real success. And for French librarians it was a good opportunity to discover open archives' issues", "In Paris VIII, we are going to create an eprint-server with Paris IV, and use the OAI Protocol, in order to be able to participate to your initiative").

2 THE CONTENT OF THE WORKSHOP

Invited talks

This section contains the abstracts of the presentations provided us by the invited speakers and brief comments written down by the project partners. A session is dedicated to each presentation. The slides of these presentations can be found on the OA-Forum project Website (http://www.oaforum.org/workshops/pisa_programme.php).

"New development in OAI (OAI-PMH2), by Michael Nelson

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is an evolving protocol and philosophy regarding interoperability for digital libraries (DLs). Previously, "distributed searching" models were popular for DL interoperability. However, experience has shown distributed searching systems across large numbers of DLs to be difficult to maintain in an Internet environment. The OAI is a move away from distributed searching, focusing on the arguably simpler model of "metadata harvesting". Perhaps the strongest and distinguishing feature of OAI is its simplicity: by being "smaller" than previous interoperability projects, it actually allows for more powerful and adaptable configurations and deployments. Key concepts in OAI include the separation of responsibilities of "service provider" and "data provider" and the use of community-specific metadata sets (with Dublin Core as a lingua franca).

Key to understanding the philosophy of the OAI is understanding the separation of responsibilities of "service provider" (SP) and "data provider" (DP). In practice, a SP and a DP can reside in the same entity, but it is important to understand the distinction. A DP is a repository (or "archive") - simply a collection of metadata records (which may or may not point to corresponding full-text documents). A SP provides value added services (e.g., searching, browsing) on the metadata extracted from one or more DPs. SPs are free to define their own services, presentation and interfaces tailored to the user base. These services could be complimentary or competitive.

The OAI-PMH consists of six verbs, 3 of which reveal the characteristics of the repository (ListMetadataFormats, ListSets, and Identify) and 3 verbs for extracting metadata from the repository (GetRecord, ListRecords, ListIdentifiers). OAI-PMH 2.0 is going to be realised in June of 2002, after 18 months of testing version 1.1. Although OAI-PMH 2.0 is not backwards compatible with 1.1, 2.0 represents only an evolutionary progression of the OAI-PMH. Some optional features for richer harvester – repository negotiation have been added, ambiguities removed, and extensibility hooks added for desirable features that lie outside of the scope of the protocol (e.g., machine readable rights management information).

OAI-PMH projects and services are being announced frequently; check the Open Archives Initiative home page (www.openarchives.org) for the latest news. The OAI is entering the

exciting phase where the focus is no longer just on the protocol, but more rightfully on the various services that use the OAI-PMH in novel and compelling ways as well as the community building that the OAI-PMH facilitates.

Some remarks

Some new functionalities are to be found in version 2.0, e.g. the identify-requests now contains more information about the data provider and the administration; the header of a metadata set contains information of the membership to several "subjects" sets; the introduction of a friends-container that will help to discover new repositories that are similar to the one that is being searched at the time.

After going into some details of the new defined set definitions, error codes, the possibility to encode provenance information within the metadata set, Michael Nelson talked about the alpha and beta time schedule and phase. Some European activists took part in the alpha test in addition to US American testers, including The British Library, Cornell U. -- NSDL project & e-print arXiv, Ex Libris, FS Consulting Inc -- harvester for my.OAI, Humboldt-Universität zu Berlin, InQuirion Pty Ltd, RMIT University, Library of Congress, NASA and OCLC.

All in all the OAI has about 77 registered repositories in May 2002. They expect the number to grow immensely in future and hope that the OAI-PMH use will become very wide-spread, and embedded in practice as the HTTP-protocol. The OAI Steering Committee looks on the development of the OAI protocol as something that is similar to the development of the HTTP protocol itself. So they barely see the necessity to set up large Data and Service Provider registries. But if those registries would provide more information about usage and experiences, as is planned in the Open Archives Forum, they seem to be a good way to spread the usage of the OAI protocol.

"CERN Document Server Software " by Martin Vesely, CERN, Switzerland

The presentation gives an overview of implementation approaches of the OAI-PMH protocol and its application potential within communities that are involved in projects where metadata exchange and document handling are essential. The experience gained within the e-prints community is presented, particularly the implementation in the CERN Document Server Software (CDSware), featuring as it does both the service and the data provider aspects foreseen in the protocol.

In general, the OAI-PMH protocol offers a framework for metadata exchange between repositories in a distributed, structurally and syntactically heterogeneous environment, and allows the repositories to practise metadata exchange in a uniform way, taking advantage of the widely deployed web infrastructure that makes it applicable at a significantly lower cost. Services provided within the OAI framework can be characterized as information products created either by a suitable assembly or correlation of metadata gathered from various data providers or by other eventual value-added activities. The core service of the CERN Document Server (CDS) within the OAI framework is the *search engine* for documents in *particle physics* and related disciplines using *value-added* metadata. Today the CERN repository contains more than 550000 metadata records containing mainly original CERN records that may be of interest to potential service providers as a source of scientific and engineering documents in the field of particle physics and related research at CERN. There are also records harvested from other repositories, enriched, maintained and updated periodically by the library staff. In the centralized model that we use today every involved

institute submits its contribution of metadata records to the master repository and then harvests the entire contents back. We discuss possible potential of the protocol for an application in decentralized or distributed systems with different topologies based on hierarchical and reciprocal harvesting.

We conclude by discussing protocol features that require customization on the application level rather than on the level of the protocol itself. In order to practise the metadata exchange some policy concerning optional issues has to be agreed within the community. Several issues that we consider essential for broader interoperability within the e-prints community (and most likely also within other communities) are pointed out, such as (i) the http error-prone data flow control, (ii) dealing with the semantic heterogeneity of OAI sets, (iii) the relation of OAI identifiers to the value-added records, and (iv) information loss caused by an exclusive support of the default metadata format.

Some remarks

The CERN documents server seems to be a good example of a community-specific workflow management tool for digital scientific documents. The CERN Document server software is freely available under the GNU public license and uses Mark 21 as internal metadata format.

"Revealing a new Dynamic: Interaction in an Open Access Archive" by Steve Hitchcock

The talk will show that open access works for authors and users, and reveals some new aspects of the social life of an eprint archive. Illustrating software and services developed as part of the Open Citation Project, and using data from our associated studies of arXiv user behaviour, it will be shown that a new 'dynamic', the speed of interaction between users, becomes evident when access to full resources is free, open and unrestricted. This is important for all those who are building open archives, and for those who are tentatively moving towards building open archives (e.g. the biomedical community).

Some remarks

Steve Hitchcock's presentation, 'Revealing a New Dynamic: Interaction in an Open Access Archive', was an introduction to the ideas of the e-prints community. E-prints have been around since long before the OAI-PMH, and improving access to e-prints was the original focus of the OAI (now widened to include 'access to a range of digital materials'). Hitchcock's presentation drew on the Open Citation project (http://opcit.eprints.org/), which is developing software for reference linking in large-scale open archives other software, and investigating citation impact and usage patterns of these archives, among other activities to support the take-up of self-archiving and the use of open archives of scientific literature. Using data provided by the Open Citation project, he illustrated how 'open access' and the increased speed of interaction between users creates a new dynamic when access to full resources is 'free, open and unrestricted'. Hitchcock maintained that when the idea of interoperability is not enough to convince decision-makers (especially the authors who will decide whether or not to self-archive) of the value of open archives, we talk to them instead about the critical value in access for users of research literature, impact for authors of research literature, and quality for research.

"Online Information in Astronomy - From networking to a virtual observatory", by

Francoise Genova – CDS, Strasbourg

Astronomy relies on long-term observations of variable phenomena, and conserving and reusing data is the key for major scientific objectives, such as the definition of objects and of their properties, or the study of variability and evolution, all this requiring statistical studies on large number of objects. Observations at different wavelengths, with different techniques, allow one to understand the physical phenomena at work in objects. In addition, astronomical observations rely more and more on large ground-based and space observatories, and on large surveys of the sky, and reusing their data for new scientific objectives is necessary to optimise the scientific return of these large projects.

This is an old, but increasingly complex endeavour: information volume and complexity increase, and information is heterogeneous and distributed. Moreover, data must be properly documented to remain usable. A technical revolution has of course occurred in the last years in that domain, with the increased technical capacity to store and manage information, and the new possibilities offered by the WWW in terms of information distribution, of integration of data with documentation, and of navigation between on-line services. These useful and appealing tools are widely used, but one has to keep in mind that careful work on the service contents and functionalities remains mandatory, and that information validation remains critical. Moreover, services and links have to be maintained on the long term.

The development of services for the usage of the scientific community puts new constraints on the Agencies, with competition between data conservation/diffusion and the implementation of new instruments or operational costs. The scientific community also has to put a sufficient priority on this activity, in projects, evaluation, and strategic plans, and to encourage motivated scientists and engineers to work on data conservation and diffusion. Projects have to make their data available, in a usable form, i.e. data has to be properly selected, organized and documented, and the "project memory" has to be kept.

Astronomy has very rapidly taken advantage of the new technical possibilities, by developing on-line information services and information networking. It is a small discipline with few commercial constraints, which has helped to build long term partnership to define community standards, thus allowing the development of links and of generic tools to access data. The network of astronomical information goes from observations, distributed in observatory archives, to results published in electronic journals, with also disciplinary centres distributing data and information in a given domain, and Data Centers building value-added services and generic tools. For instance, the Centre de Données astronomiques de Strasbourg (CDS), created in 1972 with the mission of taking care of electronic data, of building expertise about the data, of implementing tools for science, and of playing an international role, now summarizes its charter as follows: "Collect, homogenize, preserve, distribute astronomical information, for the usage of the whole astronomy community".

Astronomy has developed disciplinary standards and tools for interoperability: for instance, FITS is a widely used data format (images, spectra, tables), and data from any telescopes is kept in FITS format, thus allowing the usage of common tools to deal with it. Another example is the "bibcode" (a 19-character description of published references), first defined by database managers who had to exchange bibliographic information, then adopted and extended by the reference astronomical bibliographic database, ADS, then by the journals when they developed electronic versions, then by observatory archives when they wished to implement links with published papers using their observations. The successful networking of bibliographic information in astronomy demonstrates how a high level of interoperability can

be built using de facto standards, with a bottom-up approach defined by a small group of practitioners (long before the advent of the Internet!) and "snowball effect" in the community. In this model, links are easy to build but the quality of contents and validation are fundamental and in the hands of specialists. To take into account the current development of general standards for bibliography, a gateway with bibcode will be developed, to preserve on one hand the human readability of the bibcode and the functionalities and networking of astronomy on-line resources, and on the other hand to permit links with other disciplines (e.g. a correspondence table between the bibcode and DOI).

Another example of disciplinary standard is the description of tabular data, the "ReadMe", an ASCII file, which contains information about the physical organization of the data and about its scientific meaning. It is common to catalogues, tables published in journals, surveys, and catalogues of observations in archives. It allows also a check of the homogeneity of tabular data in journals before publication, which improves the quality of published information, in addition to the peer review. This also means that published tables are usable data, and not only figures printed on paper. In the recent years, an XML standard for astronomical tabular data has been developed (astrores), and the usage of XML is currently widely discussed and implemented in the context of the Virtual Observatory projects.

The Virtual Observatory (VO) can be defined as "an enabling and coordinating entity to foster the development of tools, protocols, and collaborations necessary to realize the full scientific potential of astronomical databases in the coming decade" (NVO White Paper, June 2000). The VO has many components, going from network infrastructure or computer and data GRID, to tools and standards for data mining, or to statistical tools able to access very large, distributed data sets. Several RTD/Phase A projects have been accepted in 2001, the Astrophysical Virtual Observatory in Europe, the National Virtual Observatory in USA, AstroGrid in UK, ... The European project (PI: European Southern Observatory, partners: ESA-ECF, AstroGrid, CDS, Terapix, Jodrell Bank) has three work areas: Science use case and requirements, Interoperability deployment and demonstration, and Technology needs. CDS is responsible for the Interoperability Work Area, and a prototype using the CDS data federation and data integration tools is being developed, to give access to ground- and space-based, multi-wavelength, multi-technique archives. The prototype is made available to the community for scientific usage, in order to obtain science results and user feedback at an early stage of the project. Another objective is to establish a set of usable recommendations for helping archive managers to implement interoperability. CDS also leads the Interoperability Working Group set up by the OPTICON European Network, which aims at studying cost effective tools and standards for improving access and data exchange to/from data archives and information services. The VO projects are coordinating their activities at the international level, and the first common milestone has been the definition of an XML standard for tabular data, VOTable (V1.0 was released on April 15, 2002).

CDS	http://cdsweb.u-strasbg.fr/
AVO	http://www.eso.org/projects/avo/
NVO	http://www.us-vo.org/
AstroGrid	http://www.astrogrid.org/
OPTICON	http://www.astro-opticon.org/

VOTable http://cdsweb.u-strasbg.fr/doc/VOTable/

Some remarks

The astrophysical community is a good example of a well-organised community. The need of specific services for supporting the works of the astrophysical researchers motivated the creation of de-facto interoperability standards (e.g. the description of tabular data or bibcode) long ago. Recently, this community has started to discuss the possibility of mapping some of these de-facto standards into more widely accepted standards in order to enlarge the access to, and the exploitation of their documents. The discussions on the choices to be made for such mappings are still in progress, and no consolidated result exists as yet. The astrophysical community is also one of the first communities that is experimenting with the GRID technology for achieving different levels of interoperability. It is certainly very interesting to follow the developments made by this community, to monitor their progresses and to see to what extent and how they will open the access to their archives to other parties.

"TORII: Access to Digital Research Community" by Fabio Asnicar-SISSA, Italy

The communication of the results of scientific research, and in many ways, research itself have changed in recent years as digital means of information production, distribution and access have become widespread. Paper preprints have been replaced by electronic archives, mail and phone calls by e-mail, typewriters and hand drawing by text and graphics software programs, cabinet files by saved directories on hard disks. These new tools, together with multimedia presentations and conference websites, constitute the growing digital network of information that is taking over many aspects of the working place of research. It is a system in which the information flow is regulated, integrated and made available by the software and the network.

This digital network of research is currently organized in three layers:

- Repositories of information: open archives and databases. This first level is the analogous of library and publishers stacks.
- Services over and for information: e.g. review journal, cross-citation. They are the analogous of, for instance, library desks and paper journals.
- Digital communities: synergic union of services and information. Ideally, they replace your desktop environment by giving access to the tools you use in your everyday work.

In this talk we will present Torii, a system that gives direct access to the digital research community. All tools and documents the user need are collected under an unified access point, organized according to his needs and ready for him everywhere he is are and at any time he may need them. An intuitive user interface helps the user to navigate. All the tools the user need are at his fingertips. Choice of archives and subjects are easily costumized to fit his interests. This platform grows as the digital community grows. New features will be added as they become available in the future.

The personal folder is the hub of the system. The user can store his documents here for future reference or to be printed or sent to others. The personal folder is easy to use by means of its drag-and-drop interface. It ideally replaces the cabinet filer where paper documents used to be stored. Stored documents can be ranked according to his profiles, impact factors or

evaluation tools. The user will find in his personal folder new documents suggested by the social filtering engine and he can attach to any documents comments for himself or to be shared by the community.

Documents to be manipulated can be organised in a multi-layered stack. It could be an entry in a database, and as a new layer is added so is the entry column modified in the database, or it could be a collection of documents managed by a web server that keeps track of their relationships and modifications. The access to a multi-layered document is dynamical. According to who the user is at a given moment---reader, author, referee, editor---he has access to different layers. Dynamical access requires an appropriate interface between the multi-layered documents and the users. It also requires intelligent agents to sift through the increasingly large amount of information to shape it into some hierarchy and thus making it usable.

Key features for the integration of dynamic access to the information into the portal are the XML language and the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH). The XML language is used to encapsulate in a common structure the exchanged information, originally stored in a variety of formats. This XML metadata structure represents the semantic aspects related to the data.

The OAI-PMH defines a HTTP-based mechanism for harvesting XML files containing metadata from repositories. This is the basic communication protocol between Torii and the underlying services, operating in a location-transparent way. Through the use of the OAI-PMH, Torii will be easily extensible to any archive implementing the protocol.

In a user-friendly information society, the information overload is limited and the information delivery is personalized: the broadcasting of information is replaced by a more effective narrow-casting and mass-media are replaced by personal media tailored to each user's needs. These aims can be reached by more effective systems for information access. Torii provides a filtering component to skim too large a set of retrieved information and thus providing the user only with the information nearest to his interests.

In Torii the user defines his research interest profiles by filling in a form; from this a user profile is derived, based on a semantic network. The profile is automatically updated every time the user provides explicit relevance feedback on some new documents. Documents to be evaluated by the cognitive filtering module are processed through information extraction techniques aimed at capturing the meaning of the document content. These techniques exploit linguistic processing and statistical analysis. Every day the filtering module filters the submissions to the archives accordingly to the user profiles and the graphical user interface provides tools to rank displayed documents accordingly to the user's profiles. Out of the 30-50 daily submissions, the user is able to see the 3-4 most relevant at the top of the list.

Social filtering circulates interesting documents among users who share interests. It automatically feeds in the personal folder those documents that are potentially relevant for the user. The relevance of the documents is evaluated for similarity with the selections done by other users with similar interests. The process is the digital analogous of sharing paper among colleagues. It fosters the growth of the digital community.

Quality control tools memorize and exploit human evaluation of documents. They provide users with the possibility to express their evaluation of a document by filling in a predefined form and writing free textual comments. The form results are used to statistically evaluate

numerical scores about the scientific quality of the document, the comments are general, each user can choose whether his comment will be public or for himself. Users can read all public comments on a document. These tools embody a first instance of open peer review in which the community as a whole participate in the review process.

A search engine, Okapi, is accessed directly from Torii. It offers a sophisticated search environment where you can look and search among the more than 150,000 documents currently stored in the archives. Okapi offers advanced retrieval mechanisms based on the probabilistic model of retrieval and relevance feedback. It runs on both the document metadata and their full text. It is fast and accurate.

An assistant monitors the user search and helps him with helpful hints and terminological and contextual suggestions. It alerts the user for dead-end searches leading to hundreds of documents or no document at all. The user is made aware of strategic aspects of searching that allow him to fully exploit all information resources and services. The assistant comes fully integrated into the Okapi search engine of Torii.

Torii integrates also iCite. This tool extracts all citations from all the documents submitted to the archives. These are used to rank documents in Torii so that you can order them according to their impact factors. It is a completely automatic system that creates a net of cross-references inside the archives. It is an instance of service, the second level of the three-layer structure, that can also be accessed independently at icite.sissa.it to search for citations patterns and ranking.

Torii is ready to move on into the future of digital networking. As the next generation of wireless systems comes into production, Torii will be accessible from the user mobile phone. The user can connect already and use it via WAP at \texttt{torii.sissa.it/wml/ia.wml} but the full potentiality of the system must wait for the 3G broad bandwidth to come into being. At that point, the user will be able to browse documents use his personal folder and any other of the features of Torii as he travel.

More information about Torii can be found at: http://tips.sissa.it/docs/booklet.pdf

Some remarks

Most of the OAI-MPH compliant services that have been developed until now are cross-archive search services. Torii is one of few examples of services that exploit the records harvested thought the OAI-MPH protocol for providing the end user with another kind of functionality. It is also an example of a system that uses the metadata records harvested to retrieve the documents themselves. Currently, the experimentation made by Torii is limited to the use of selected archives whose properties (metadata formats, quality of the records, etc.) are well known. This restriction in some way reduces the problems encountered in building new kinds of services. However, the experience made by Torii is certainly valuable. The OA-Forum intends to monitor the progresses made by SISSA on this subject and establish a permanent link with them.

"Cyclades: an Open Collaborative Virtual Archive Environment", by Umberto Straccia – IEI-CNR, Italy

The main goal of CYCLADES is to develop an open collaborative virtual archive service environment supporting both single scholars as well as scholarly communities in carrying out

their work. In particular, it will provide functionality to access large, heterogeneous, multidisciplinary archives distributed over the Web and to support remote collaboration among the members of communities of interest.

CYCLADES will run on the data environment composed by the archives that adhere to the Open Archives Initiatives harvesting protocol specifications (http://www.openarchives.org).

From the technical point of view, CYCLADES will consist of the following federation of interoperable services:

- *Access Service*

It harvests information from the set of OAI compliant archives, and indexes and stores the gathered information in a local database.

- *Collection Service*

It provides mechanisms for dynamically structuring the overall information space into meaningful (from some community's perspective) collections.

- *Search and Browse Service*

It supports the users in formulating queries and develops plans for their evaluation. In particular, it provides an advanced multilevel browse facility, completely integrated with the search facility, that allows one to browse at schema, attributes, and document levels.

- *Filtering Service*

It supports information filtering on the basis of individual user profiles, and profiles of the working communities the user belongs to. User and community profiles are automatically inferred by monitoring the user behavior.

- *Recommendation Service*

It provides recommendations about new published articles within a working community. The choice about what recommendations to send and to whom is based on both the user and the working community profiles.

- *Collaborative Work Service*

It supports collaboration between members of communities and project groups by providing functionality for creating shared working spaces referencing users' own documents, collections, recommendations, related links, textual annotations, ratings, etc.

In this presentation, we will introduce each of the above services and we will discuss how each of them is influenced by the content harvested by the OAI-MPH.

Some remarks

Cyclades is an example of a complex system that uses the information harvested from the OAI compliant archives not only for information discovery but also for the provision of other features. In particular, it uses the metadata harvested for building "folder profiles" which express the user/community information needs. The folder profiles are used for filtering information and for building recommendations.

The project proposal does not contain any pre-defined selection of the archives to be harvested. It will be interesting to see to what extent the quality of the services is influenced by the quality of the metadata records harvested.

Another interesting experimentation will result from the implementation of the Collection service. This service provides a virtual and dynamic organisation of the information space and support source selection. The implementation of this functionality requires some knowledge about the underlying archives. This knowledge could be acquired by exploiting the content of the OAI-PHM Identify container. At the moment, however, no community has established an agreed standard for this. In order to implement this functionality by relying only on the established protocol, Cyclades is experimenting query sample algorithms. It is interesting to see to what extent this solution turns out to be effective within this context.

Break-out sessions

Some preliminary work was done before the workshop in order to prepare a ground for the workshop discussion. In particular, a technical questionnaire was circulated among the registered participants, and many of them where interviewed to ask them for information about OAI activities in their countries.

Terminology breakout session

The terminology breakout session was facilitated by Philip Hunter of UKOLN (University of Bath), and reported by Joe Yates of Space S.p.A. (Italy). The session covered a large amount of territory in the limited time available. Michael Nelson of the OAI was present throughout, and contributed both extensively and constructively to the discussion. Many of the terms present in the Forum's Metadata Glossary were discussed, as well as two of the terms in the name of the OAI - 'Open' and 'Archive', which were found to be problematic ('archive' in fact is not used in OAI documents).

The breakout session also looked at terms that it would be useful to add to the glossary. These included definitions of 'resource' and 'rights', despite the fact that the OAI-MPH explicitly is not concerned with the latter. 'Repository' was another term whose meaning might have difficulties for potential implementers: a draft glossary redefinition resulted from the discussion. Other additional terms that might be added were also discussed: these included 'item' and 'set': these terms were defined in the course of the discussion, and will be added.

Other issues discussed included the potential usefulness of translating the OA-Forum glossary into different European languages. It was decided that this would be useful for those producing introductory articles and papers, especially those aimed at managers and other decision-makers. This was unlike the situation for the OAI-PMH documents themselves, where maintaining versions in several languages was an unnecessary overhead, since implementers generally have a good knowledge of technical English, and would go directly to the original publications. A number of those at the session expressed interest in providing translations for the OA-Forum. The workshop also looked at the target audience for the glossary, and at how difficult it was to locate it on the OA-Forum site. It was decided that a new menu item should be placed on the home page to facilitate access.

Other documentation that it might be useful to make available was discussed. It was suggested that a 'non-normative' document might be created which indicated 'useful things to do with OAI' to stimulate the development of the technology.

Overall, rights management, and the fact that the OAI does not address this important issue, ran through much of the discussion.

Organisational Issues

The breakout session on organisational issues was facilitated by Denis Nicholson of the University of Glasgow, and reported by David Casal of the University of East Anglia. It provided an opportunity to share experience and identify issues that were arising across many initiatives. A full report of the session is available on the OA-Forum website from the 1st Workshop programme page at http://www.oaforum.org/workshops/pisa_programme.php.

Participants represented a substantial body of implementation experience, which showed that academic attitudes are important. This applied to convincing a user base to actually use the service either as contributors or users of data, though the scientific community are more eager to contribute electronic copies. Institutional structures and polices constitute a framework within which an open archive must find its proper place. Unless repositories are firmly embedded in institutional life, they won't survive the end of their project funding, or the departure of their original implementers.

IPR (Intellectual Property Rights), metadata, quality of metadata and content, and business models emerged as important issues. Intellectual Property Right (IPR) was acknowledged as an important issue, with a need to protect institutional IPR on one hand, and the dangers of institutions inadvertently violating IPR on the other. Better understanding and awareness of IPR is needed. There was considerable discussion of subject metadata requirements, the role of authors in the creation of metadata, the need for guidelines for the application of unqualified DC (Dublin Core), and the adequacy of DC in meeting the needs of service providers.

Participants in the organisational issues breakout session agreed to establish a European working group that will carry out discussions and produce best practice guidelines in key areas. It will communicate via the OA-Forum discussion list in the first instance, and meet at upcoming OA-Forum workshops. The first task of working group is to scope its work and decide which guidelines to produce as a priority.

Technical validation

The breakout session on technical issues was facilitated by Susanne Dobratz, and reported by Steve Hitchcock of the University of Southampton. The session focused on the experiences made with existing OAI implementations and the problems, that particular demands caused. The session was prepared by an web-based questionnaire on the Open Archives Forum website, that participants were asked to fill out before the workshop.

18 People contributed to the questionnaire, 6 from Germany, 5 from Italy, 2 from Belgium, 2 from the Netherlands, 1 each from France and Sweden and 1 from the UK.

Half of them said, that they were still in a planning or test phase and had no fully established ideas for which services they would use OAI. If people were using OAI as data or

service provider in most cases they implemented the software used. If not, one of the following software tools was reused from the OAI software archive: ETD-db software with OAI extension or the Eprints software.

The main point about using OAI as an approach to open archives was that people want to provide additional services to existing services, such as a supplement to a paper catalogue, a delivery of electronic holdings info to an OpenUrl resolver system, or they want to use other, different and additional possibilities of dissemination of information about scientific results of their research staff.

Some participants mentioned the replacement of existing services through an OAI Interface, e.g. to replace the single, non-searchable lists of eprints available on the websites or to develop the practice of alternative means of dissemination of the scholarly communication.

A number of answers referred to research projects, e.g. one person answered: We used to run a search-engine that tried to combine different HTML-outputs from different sources. This has been replaced by a compatible OAI search-engine.The importance of the OAI technical framework was also measured by the better retrieval, people expect by using OAI. In order to make metadata exchange available single projects started harvesting their own archive via OAI (replacement of search engine) or to build an interface for metadata exchange within several projects (new service). Another goal of using the OAI technical framework was that to establish an exchange of the library catalogue with other universities and the integration into a virtual union catalogue for the whole country. Many people started using OAI because they want to be able to develop services based on OAI compliant archives maintained by others.

As the advantages of OAI the following points were seen: OAI provides a chance to share scientific knowledge and to harvest other knowledge databases, it provides an opportunity to import metadata in library software and to implement a major dissemination of researchers' results. It is simple in implementation for data providers and provides a simple to implement facility of exchanging metadata in comparison to more complex protocols like Z3950, or others because in many cases the archives are too small for a Z39.50 service. By harvesting several relatively small archives at once it is possible to maintain a Z39.50 service.

There were about 25 persons attending the session on " Technical Validation ". After the partners from Humboldt-University gave an overview of the answers to the questionnaire, the following topics were particularly mentioned during the discussion:

- How can sets and set hierarchies within the OAI protocol be defined? Is there any standardisation either on a subject based level or on a cultural and country based level?

- How does the implementation of an OAI data or service provider fitted into the local existing IT infrastructure?

- Which are the points that need to be solved by special communities in order to get clear data? Are there standards?

Thomas Place from Tilburg University spoke shortly about the Dutch ARNO project (http://www.uba.uva.nl/en/projects/arno), where they integrated an OAI based search into an Z39.50 environment.

Jens Vindvad, Riksbibliotektjenesten Oslo, talked about the situation in Norway, where there is no OAI implementation, but the new library system contains a full text search and uses OAI.

Heinrich Stamerjohanns, Institute for Science Networking at the University of Oldenburg, Germany, gave a short overview of the implementation costs and strategies used for the Physnet/PhysDoc (http://www.physics-network.org/PhysNet/physdoc.html) OAI based Search engine.

Many of the participants in this breakout session had not yet any experiences with OAI, so contributions to the session were made by a few specialists that talked about very special aspects of their OAI implementation.

Finally the discussion got back to the metadata topic and to the use of classification schemas and controlled vocabulary and the quality of metadata and the XML output used by data providers.

Communities and Services

The breakout session on communities and services was facilitated by Donatella Castelli, and reported by Michael Popham of the Oxford University Computing Services, and Paul Child of UEA at Norwich, United Kingdom.

Three questions were raised to start the discussion: i) Which kind of services can be built on OAI-PHM compliant archives? Under which hypotheses? ii) Is OAI-PHM a basic tool for supporting interoperability among archives or it is just a component of a wider infrastructure? iii) Do we need open services?

It emerged clearly at the beginning of the session that at the moments there is not yet much experience to answers these questions. The services that have been developed until now are mainly cross-archive search services. These services works on the Dublin Core metadata records. The quality of these metadata is the crucial point for these services. Dublin Core only defines a set of optional metadata fields, not restricting their content any further. Thus the use of Dublin Core differs significantly among the individual archives. For example, several archives list all the creators in one single tag of the metadata record – sometimes not separated by delimiters - whilst other repeat the creator tag; the format of the date is often different; etc. It was noticed that, Dublin Core as it is now only allow for free text search.

Projects like Cyclades will provide better insights on the questions formulated above. Certainly, everyone agreed that the quality of metadata records will play a significant role. This quality is measured in terms of richness (the information provided) and consistency (subject scheme, authority files). There was a general agreement on the statement that the service providers need to identify and to select the archives that offer metadata with the required level of quality. This implies the possibility of retrieving these archives and getting information about them.

During the Workshop the OAI-PHM protocol was presented not only as a protocol for cross-domain discovery but also as a protocol that can be used to achieve intra-community interoperability. Some members of the astrophysical community commented this point by point out that the protocol is based on certain assumptions, like the use of the http protocol or the use of XML, that are not necessarily acceptable for all kind of applications. They also pointed out that the Dublin Core is not sufficient to cover the descriptive needs of data archives.

A lot of discussion was raised about this last point. We noticed that people tend to identify OAI-PHM with Dublin Core and often to confuse harvesting issues with search issues. The criticism to OAI-MPH often are really criticisms directed towards Dublin Core.

The break-out session ended by discussing about the idea of extending openness also to services, i.e. make services available to other service providers. It was felt that a lot has still to be done before being able to reach this goal.

3 Outcomes and Actions

The outcomes of the workshop and the actions agreed are fully outlined in the summing up presentation slides on the OA-Forum site at:
http://www.oaforum.org/otherfiles/Pisa_summingup.pps

The major outcomes of the workshop include communication between participants, an initiative to set up a European working group on organisational issues, and an identified need for a mailing list to support continuing discussion. In response to requests from participants, the OA-Forum project will also provide an OAI training session and an expert review of IPR issues for the next workshop, more pointers to external information from the OA-Forum web site, and a revised and translated glossary.

Participants will continue to share information by discussing technical issues on the OAI-general mailing list, discussing organisational issues on the OA-Forum mailing list, and registering project and service information on the OA-Forum website. Some participants will form a working group on organisational issues. The remit of the working group will be to produce best-practice guidelines, and to discuss organisational issues. The scope is to be determined, and could include (for example) metadata use and the adequacy of unqualified DC; costing, planning, and managing repositories.

4 Lesson Learned

- Some people are not yet well informed about OAI-PMH: many of them need a tutorial.
- Many people see the OAI development as a chance to open their resources, but a large number of people are still waiting until this protocol has been established on a broader basis. They cannot afford to put money and staff into developments, until they have some confidence that use of OAI will be widespread and successful in future. This slows down the impact that initiatives like OAI have in Europe.
- The open issue is an important one; many people want to give open access to their resources – not only in the e-print community.
- People are digitising, but they often need ways of making digitised material openly available.

- Some communities have found their solution for internal interoperability. Each solution is known only within the community implementing it. There is potential for exchange of info/solution among different communities.
- The socialization aspect is very important in a workshop like this one: people talk and establish agreements.