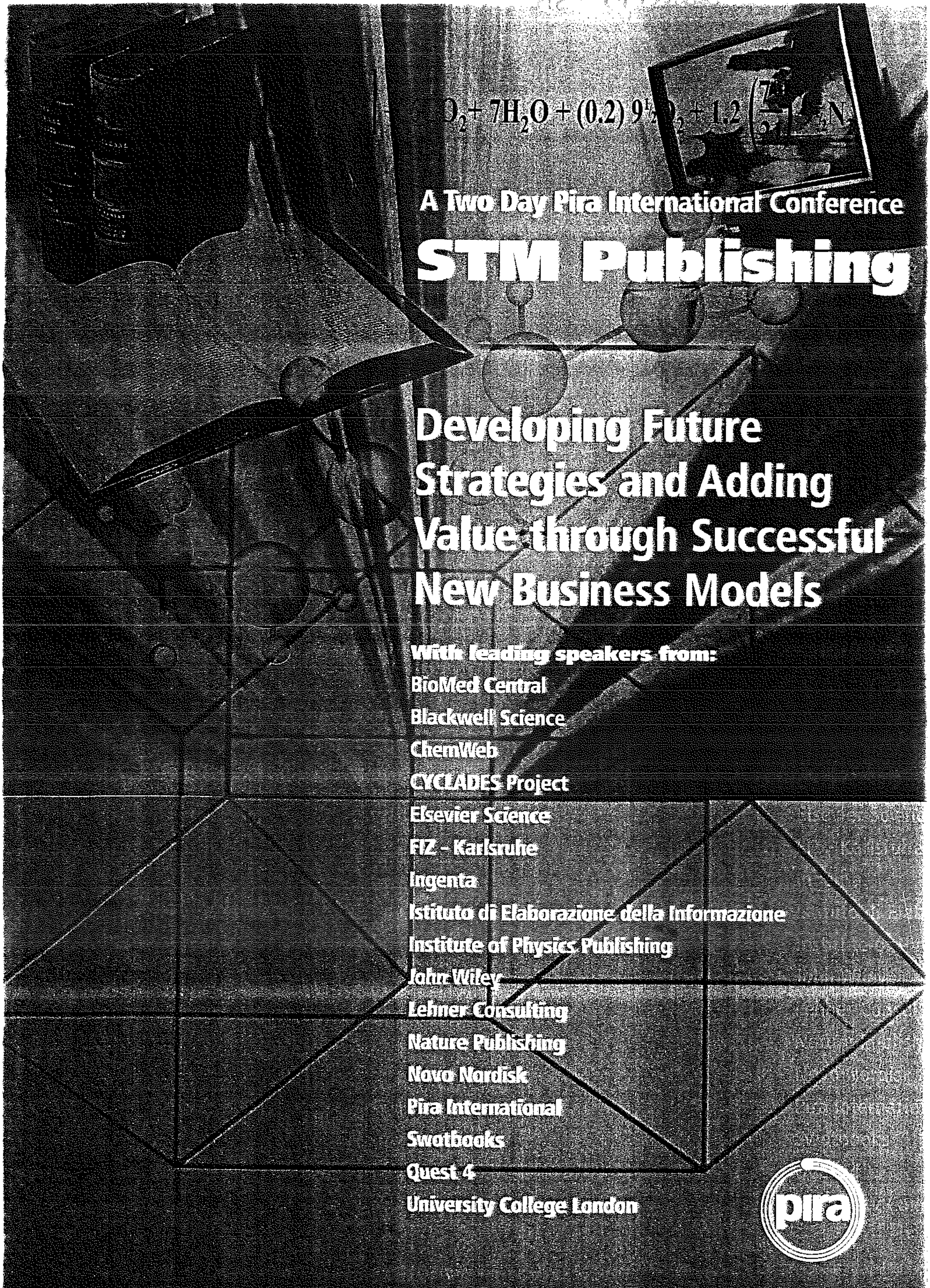
92-29 [illegible]


$O_2 + 7H_2O + (0.2)\,9\tfrac{1}{2}\,O_2 + 1.2 \left(\tfrac{7}{3}\right) N_2$

A Two Day Pira International Conference

STM Publishing

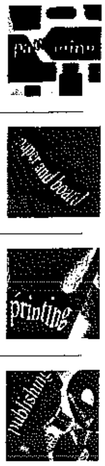
Developing Future Strategies and Adding Value through Successful New Business Models

With leading speakers from:

BioMed Central
Blackwell Science
ChemWeb
CYCLADES Project
Elsevier Science
FIZ - Karlsruhe
Ingenta
Istituto di Elaborazione della Informazione
Institute of Physics Publishing
John Wiley
Lehner Consulting
Nature Publishing
Novo Nordisk
Pira International
Swotbooks
Quest 4
University College London

pira

Tuesday 21st & Wednesday 22nd November 2000 Gladstone Library, One Whitehall Place, London, UK



STM Publishing

Pira International conference proceedings

Tuesday 21st & Wednesday 22nd November 2000
One Whitehall Place, London

Pira International
Registered office: Randalls Road, Leatherhead, Surrey, KT22 7RU, United Kingdom
Telephone: Leatherhead (code UK +44) 01372 802000 e-mail training-services@pira.co.uk Fax: (code UK +44) 01372 802243
Registered number: 3858209 England (Limited Liability)

Paper 5

STM Publishing

An open collaborative virtual archive environment

Carol Peters
Istituto Di Elaborazione Della Informazione - CNR
Italy

One Whitehall Place, London
Tuesday 21st & Wednesday 22nd November 2000

Pira International
Registered office: Randalls Road, Leatherhead, Surrey, KT22 7RU, United Kingdom
Telephone: Leatherhead code UK +44 (0)1372 802000 e-mail training-services@pira.co.uk Fax: code UK +44 (0)1372 802243
Registered number: 3858209 England (Limited Liability)

An Open Collaborative Virtual Archive Environment

Donatella Castelli and Pasquale Pagano

Istituto di Elaborazione della Informazione-CNR

Pisa (Italy)

e-mail:castelli/pagano@iei.pi.cnr.it

1. From e-print archives to collaborative environments

The history of automated archives for electronic communication of research information (e-print archives) begins in 1991 with the went on-line, at the Los Alamos Research Laboratories, of an archive for research information in formal areas of high energy particle theory [Ginsparg]. This archive was set up for circumvent inadequacies of research journals, especially the speed and the cost of dissemination. Unexpectedly, within a very short period, this e-print archive became the primary means of communicating research information in that particular area of physics. Since then, following this successful experience, and motivated by the exponential increase in the usage of the electronic networking, several other e-prints archives serving other research disciplines have been set up.

Initially, these e-print archives were intended as a means for supporting the submission of abstracts and research papers, their permanent storage, indexing and retrieval. During time, however, the new electronic medium had given to the researchers the opportunity to reconsider many aspects of their communication and has raised new expectations about e-print archives that go far beyond the usual functionality of an electronic clone of a printed journal. New initiatives have been thus set up to move these archives from a simple networked distributor of information objects to a medium for supporting more advanced forms of scientific collaborations. In order to meet this objective these initiatives work mainly in two directions: (i) to define mechanisms for achieving interoperability among existing archives, so to permit a wider access and use of the e-print archives and (ii) to construct new tools for transcending the limits of conventional journal in structure, content and functionality.

The purpose of this paper is to exemplify this tendency by presenting two of these initiatives.

The first, called Open Archives initiative (OAi) [OAi] (see Section 2), has been set up by a consortium which group the currently most widely used e-print archives. The initial aim of this initiative has been to create "a forum to discuss and solve matters of interoperability between author self-archiving solutions, as a way to promote their global acceptance". The first act of this initiative has been the establishment of a set of quite powerful interoperability specifications to be implemented by e-print archives in order to facilitate the development of cross-archive end-user services implemented by third parties.

The second initiative, called Cyclades [Cyclades] (see Section 3), is an EU V Framework funded project (IST - 2000 - 25456) which aims a constructing a set of services for supporting a community of scientists in accessing information from a set of multidisciplinary archives and in collaborating with members of their own communities. These services will be experimented on top of the OAi established interoperability layer.

These two initiatives have been proposed within for a scholarly framework but their principles can equally be applied to other domains.

Few other initiatives with similar objectives are briefly described in the concluding section.

2. The Open Archives initiative

2.1 Motivations

There are currently several e-print archives that serves large communities of scientists. The most widely know is the already cited Los Alamos E-Print Archives, called ArXiv.org, which at the present maintains non peer-reviewed research papers, not only in the physic research areas but also in the mathematics, non-linear systems and computers science. This archive stores more than one hundred and fifty thousands documents, it has a monthly submission rate of two thousand documents and typically process more than one hundred thousand connections per-day.

Other significant archives are NCSTRL, an international collection of computer science research reports that serve more than one hundred of institution around the world; and NDLTD a digital library that maintains electronic theses and dissertations published in more than one hundred different countries.

These archives have been set up by specific communities that use them as an instrument for supporting collaboration among its members. Their services address the peculiar needs of their own specific community and are accessible through ad-hoc interfaces. This specificity prevent the exploitation of the full power of these archives since it precludes the possibility of serving cross-domain communities and, when there are more archives, also limit their full utilisation even within the same community. The Open Archives initiative (OAi) (http://www.openarchives.org) has been set up to discuss and solve this problem of interoperability between author self-archiving solutions.

OAi originated by a meeting held on July 1999. The meeting was sponsored by the Council of Library and Information Resources(CLIR), the Digital Library Federation(DLF), the Scholarly Publishing & Academic Resources Coalition (SPARC), the Association of Research Libraries (ARL) and the Los Alamos National Laboratory (LANL). The participants were representatives of different scholarly e-print archives. The aim of the meeting was to explore the co-operation among these archives as a way to contribute in a concrete manner to the transformation of the scholarly communication.

A first result of this initiative has been the establishment of a set of technical specifications and organisational principles known as "Open Archives Harvesting Framework Specifications" (HFS) (previously called "Santa Fe Convention"), that states a set of agreements for achieving interoperability among archives. The implementation of this agreement is a sufficient condition for an archive to become interoperable with the others that implement the same agreements.

2.2 The Open Archives Harvesting Framework Specification

The approach to interoperability proposed by HFS[1] relies on the distinction between the proper archive functions, which include data collection and maintenance, and end-user functions, like cross-system search, recommendation and linking. Although proprietary archives can implement their own end-user services, it is essential that the archive remains "open" in order to allow others to equally create such services. This concept has suggested the distinction between *providers of data* and *implementers of services.* HFS makes easy to implement technical recommendations for data providers that allow data from e-print archives to become widely available through end-user services developed by third-parties.

These recommendations allow to achieve interoperability among archives at the level of metadata harvesting. This level of interoperability permits the extraction of metadata descriptions for the resources that are maintained in the archive.

Within HFS for "archive" it is meant an abstract object with a state made by a set of "containers" and a set of operations which implement the following functionality:

- a submission mechanism
- a long tem storage system
- a management policy with regard to the submission of documents and their preservation
- an open machine interface, that enable third parties to collect data from the archive.

For "container" it is meant a "black box" that represents the resource. It encloses the content of the resource (or a link to this content) and the description of the resource in one or more metadata formats.

The HFS technical recommendations for data providers impose constraints on the archive open machine interface. In particular:

[1] The Open Archive Harvesting Framework Specification has not already been officially released. The description we present in this paper had been derived partially from the description of the Santa Fe Convention [VanDeSompel] and, partially, from a draft version of this specification not yet available for distribution.

- *It must comply with the Open Archive Harvesting Protocol.* This protocol has an HTTP-based implementation. It provides a communication procedure, as well as the syntax for the corresponding messages and responses, that allow service providers to harvest metadata selectively from open archives that comply with HFS. The protocol requests to be implemented provide the following functionality:

  1. List the full identifier of records stored in the archive
  2. Return the metadata for a specific record in a requested format
  3. Return the list of metadata formats supported by an archive
  4. Return the list of metadata formats available in a specific container
  5. Return the list of partitions by which an archive is organised

- *It must provide a Dublin Core metadata description* [DCMES] of the resources in the archive. This shared metadata format permits a coarse granularity resource discovery among the records in distributed and dissimilar archives. Beside that, each archive can decide to made available through its interface also the specific metadata formats that suit the need of its community and the types of data they handle.

The HFS organisational guidelines for data providers establishes the following requirements:

1. *Choose an unique identifier for your e-print archive*
2. *Use unique persistent identifiers for containers in the archive*

   These identifiers are built as a concatenation of the unique archive identifier and the unique persistent identifier of a container in an archive.

3. *Implement and document other metadata formats supported by your e-print archive*

   Service providers will be able to provide more powerful services for users if a metadata format that is richer than DC can be harvested from an archive. The convention encourages data providers to provide access to the full richness of metadata available to support discovery and retrieval of records in their archive, preferably by adopting an exchange format already used by other e-print archives. A registry of existing formats is maintained so that each data provider can determine if there already exists a format that serve its needs. In case there is no such format, a data provider can compile its own format. In this case a compilation of an XML representation for this format is required.

4. *Register your print archive as open*

   A facility to register an e-print archive as being compliant with the Open Archive Infrastructure is provided by means of a filled-out version of a data provider template that describes crucial characteristics of the archive.

Nine e-print archives have currently implemented the HFS. They are briefly listed in Appendix A.

2.3 Remarks

OAi is currently modifying its status: it is moving from a spontaneous aggregation of e-print representatives, committed to render their archives interoperable, to a more stable institution. In an OAi meeting, held on September 2000 at Cornell University, a Steering Committee was appointed with the task of overseeing the pursuit of the mission and a Technical Committee was formed to focus on the generalisation and stabilisation of the technical framework. This Technical Committee is now in charge of evaluating possible evolutions of the HFS. In particular, it is now working at a revision of the HFS in order to meet the pressing requirement that emerged from other communities, especially the research library community, who are interested in applying the framework for a wide variety of scholarly material beyond e-prints. This revision is expected for the beginning of 2001. In the meantime, other e-print archives, located in different countries, have expressed their intention to become OAi complaint.

3. The Cyclades Project

3.1 Motivations

Mechanisms for interoperability offer the ground for research supporting services that may revolutionise the research communication model. These new services can span from simple tools for cross-archive searching or for keeping track of the relevant information, to more automatic tools that can create virtual working spaces for scholarly communities or even to recognise and recommend new emerging research communities. This new collaboration form may significantly impact the way in which research is carried out. For example, it may allow the setting up of virtual research institutions whose members span across geographic and domain boundaries.

The aim of the Cyclades project is to create a set of core services for experimenting this change. In particular, Cyclades will develop an open service environment that will support scholars as members of networked peer communities in collaborating and interacting with multi-disciplinary archives. The effectiveness of this environment will be tested on top of the OAi compliant data archive layer.

3.2 The Cyclades environment

The Cyclades environment is composed by a set of services, which build on a uniform archive harvesting interface, that makes it possible to collect metadata descriptions of existing resources.

The logical architecture of the Cyclades environment is depicted in Figure 1. It consists of a set of independent services, each implementing a well defined functionality, and *a Cyclades Mediator Service* that implements the specific Cyclades functionality on top of the other services. Some of these services (Access Service, Query&Browse Service) implement standard functionality for searching and accessing information, others implements advanced functions for supporting the scholarly collaboration (Recommender Service, Personalisation Service, Collection Service, Collaborative Service).

Each service is accessible through a well established protocol and it is able to serve "describe itself" requests. The last two are key features for openness since they permit an easy extension of the environment with the addition of new services and the exploitation of each of these services by other parties for the construction of different environments.

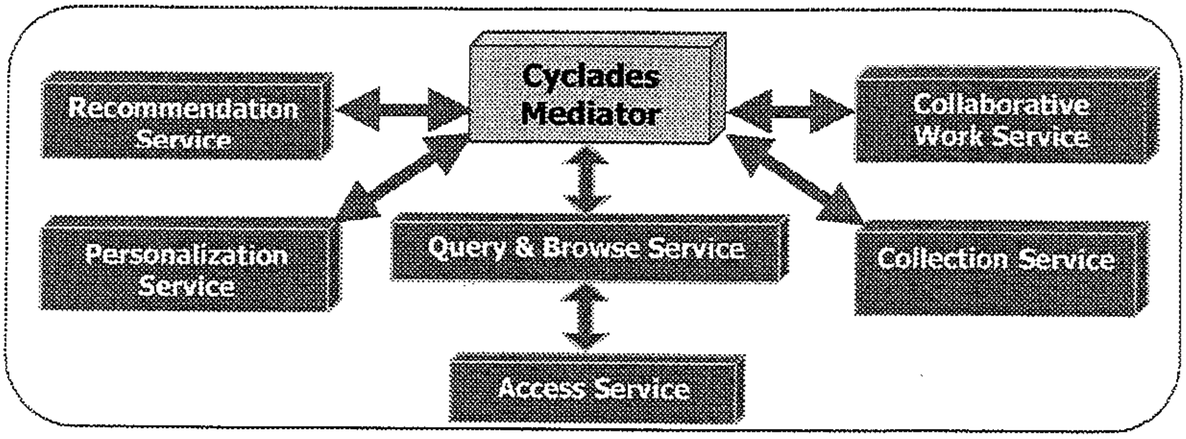


Figure 1 - The Cyclades logical architecture

For brevity, we cannot describe here the specific functionality of each service. We will concentrate our attention on the functionality provided by the Cyclades Mediator Service. This will be described in terms of the different kinds of functionality provided to the Cyclades end-user.

Virtual collections

The Cyclades user perceives the information space as organised into a set of (virtual) *collections*. The correspondence between these collections and those from where resources are harvested is transparent to the user. A collection groups the resources that are meaningful from the perspective of some scholar community and the operations available on them. As such a collection is a means for focussing on a particular resource space. The user can list the existing collections and can access the description of each of them. A description can simply report the subject of the resources in the collection (e.g. resources in "computer science" domain) or it can contain more complete information (e.g. resources in "physics" domain stored in ArXiv and PhysNet published after 1990).

The set of collections is dynamic. New collections can be created by an automatic or human administrator, for example, to respond to the needs of emerging communities. This permits to a new community to quickly establish their own dedicated communication means (i.e. something analogous to a specialised journal or to an e-print archive) and dismiss it in a flexible way if things do not work out.

Each collection makes available a set of specific operations. These are meaningful operations for that collection. For example, a cross-language search may be provided for a virtual collection of multilingual documents, the same operation is not certainly appropriate for a collection of mono-lingual documents.

Search and Browse

Search and Browse are two operations that can be associated with a collections. These may differ from collection to collection. The difference may regard, for example, the search/browse fields or the set of acceptable values for the fields.

Given a query condition expressed in terms of a set of descriptive fields, the search operation returns the resource descriptions of the given collection, in the requested metadata format, that match the query. The browse operation provides different levels of browsing: schema, attribute and document level. In order to help the users in the formulation of the query conditions (and its refinements) these levels of browsing can be intermixed among them and with search operation. This provides to the users a powerful interactive means for focussing and expressing their information needs.

Automatic user profiling

The Cyclades service provides a facility for storing the metadata descriptions retrieved as a search operations in a personal, dynamic hierarchy of folders. This facility is similar to that provided by an e-mail system where the users can define their mail-boxes and move, either manually or automatically, the incoming messages into the appropriate boxes. By observing the content and the organisation of the folders the system learns the user's information needs, i.e. it automatically creates a profile of the user. This profile is continuously updated following the changes that occur in the personal folders. This automatically acquired profile is a very important resource of information for the different services. For example, based on such an information about the user, the search service can support queries user-profile based such as for example: "Find all documents in the collection "Computer Science" that are *similar* to those that I have stored into the "Metadata" folder" or "Find all documents in the collection "Computer Science" on "interoperability" that are *similar* to those stored in the "Metadata" folder".

Recommender

Users of the Cyclades service may express their opinion on documents: typically, on a scale with high values representing a strong interest in a document and low values representing a strong lack of interest, and free text comments intended for other users to read.

By exploiting the similarity between the rating patterns of a user with those of the other users, the user profile and other information about the user activity (e.g. URL references and download requests), the system is able to identify a user community and keep track of both its behavior and its information needs (community profile).

Recommendations based on both the user and community profiles can then be delivered on user demand or in an asynchronous way (i.e periodically by e-mail, fax, etc.). A user can, for example, specify a document identifier, or an author name, and then ask to the system to receive recommendations on relevant documents/authors that "similar" to that specified.

Collaborative Work

Users that often access similar resources in a collection are automatically made aware of each other so that they may enter into a long term relationship and eventually evolve into a common scholarly interest group. A set of ad-hoc facilities is given to support collaboration within these groups. A member of the group can create a workspace and invite other members to share this space, open communication channel and exchange information. Usually, a user is member of several workspaces (e.g. one workspace corresponding to each of the research projects a user is involved with). A group shared workspace is a small repository accessible to the members of a group using a simple user name and a password scheme. In a workspace users can upload documents, hold threaded discussions, and obtain information on the previous activities of the other users to co-ordinate their own work. A share workspace can contain different kinds of information such as user documents, descriptions and recommendations of resources belonging to specific collections, related links, annotations, ratings, etc. Members of the group can set access rights to control the visibility of the information they have transferred from their machines or the operations which can performed on these documents by the others.

A particular form of collaboration is peer-reviewing. This can simply obtained by: (i) creating a community consisting of the reviewers, (ii) uploading the to be reviewed documents; and (iii) uploading the reviews.

3.2 Remarks

One of the main goal of Cyclades is to contribute to the definition of more powerful communication and collaboration scholarly networked environments. To this aim Cyclades will transfer to the archive environment a set of services, deemed important for the support to the scholarly communication and collaboration, that have been experimented and found useful by other communities in other contexts (mainly the Web). The Collaborative Work Service, and partially the Recommender Service, for example, will be built by adapting and extending the Social Web Cookpit [Cookpit], a tool to support Knowledge sharing Communities that has been developed at GMD. Similarly, the Personalisation Service will be built by relying on the experience acquired in implementing a similar tool, developed for Web applications, within the EUROgatherer project [EUROgatherer]. Finally, the Search& Browse service [DesIRe].

Another goal of Cyclades is to show how, by relaying and a clear separation between data handling and services, and on well established interoperability rules, more and more services can be developed and integrated in existing archive service environments by independent third parties which are not required to know the details of the underlying archives. The achievement of this form of extensibility is a key feature for any service environment that serve scholarly communities since it creates a framework for coping with the diverse needs, traditions and opportunities of the different communities.

The first experimentation of the Cyclades environment will be done on top of the OAi interoperability layer. This will permit a strong evaluation since the environment will be used by large and different communities. This experimentation will permit to measure the impact of the proposed services on both the communities that already use extensively e-print archives and on those interdisciplinary communities that currently have no e-print support at all.

4. Concluding Remarks

Indeed scholarly communication is changing. For a long time scholars have published their research in printed journals and conference proceedings. Access to current research required subscription to journals, usually through an academic library. More recently e-print archives have been created that permit a more widely and faster access to research information. At the present, the role of e-print archives is evolving: from simple distributors of information they are becoming instruments for supporting the scholarly communication and collaboration. This paper have presented two initiatives that aims at promoting this change in the scholarly communication and collaboration model.

Others initiatives with similar objectives exist. Among these, let us cite the German Digital Library Project [Rush-Feja] and the DEF-Denmark's Electronic Research Library Project [DEF]. The first initiative is based on furthering the co-operation with university, scientific publishing houses, book dealers and academic and research libraries to achieve 1) efficient access to worldwide information; 2) directly from the scientist's desktop; 3) while providing the organization for and stimulation fundamental structural changes in the information and communication process of the scientific community. The DEF aims at establishing the basis for a Danish electronic research library. The core of this library is the DEF Directory, a metadata based index of resources at the Web sites of Danish research libraries, and five subject gateways. The last are developed to cater for subject specialists requiring in-depth access with narrow fields. The overriding idea behind the DEF Directory is a system of highly compatible services which at the same time maintain the individuality needed to deliver access to specialised information resources within distinct subject areas.

5. References

[Ginsparg] First Step toward Electronic Research Communication,

http://xxx.lanl.gov/ftp/hep-th/papers/macros/blurb.tex

[DCMES] *Dublin Core Metadata Initiative*, http://purl.oclc.org/dc/

[VanDeSompel] The Santa Fe Convention of the Open Archives Initiatives, in *D-Lib Magazine*,

February 2000,

http://www.dlib.org/dlib/february00/vandesompeloai/02vandesompel-oai.html)

[OAi] *The Open Archive Initiative* , (http://www.openarchives.org/ups.htm)

[Cyclades] *Cyclades – An Open Collaborative Virtual Archive Environment*, http://www.iei.pi.cnr.it/cyclades

[Cookpit] *Social Web Cookpit*, http://orgwis.gmd.de/cookpit

[EUROgatherer] *EUROgatherer - Personalised Information Gathering System*, http://pc-erato2.iei.pi.cnr.it/eurogatherer/

[DesIRe] *DesIRe -Dortmund extensible structured Information Retrieval engine*, http://ls6-www.informatik.uni-dortmund.de/ir/projects/DesIRe/

[Rush-Feja] Global Info (The German Digital Libraries Project), in *D-Lib Magazine*, April 1999, http://www.dlib.org7dlib7april99704rush-feja.html

[DEF] *DEF Project Home Page*, http://www.deff.dk/admin.html

6. Appendix A

List of OAi compliant archives:

arXiv e-print archive (arXiv)

(http://arXiv.org/help/oa/sfc_data_provider)

Hosted by Los Alamos National Laboratory, is considered the premier example of e-print archives. The archive was started in 1991 by Paul Ginsparg, who is internationally recognised as one of the leaders in the area of scholarly publishing alternatives. Over the past decade, the arXiv archive has evolved towards a global repository for non peer-reviewed research papers in a variety of physics research areas. arXiv has also incorporated mathematics, non-linear sciences and computer science.

Clinical Medicine and Health Research Netprints (clinmed)

(http://clinmed.netprints.org/open_archives/oai_provider.shtml)

Launched in December 1999, it provides a place for authors to archive their completed studies - before, during, or after peer review by other agencies. Its scope is original research into clinical medicine and health.

Cognitive Sciences Preprint Archive (CogPrints)

(http://cogprints.soton.ac.uk/sfc-info.html)

Hosted by the University of Southampton in the U.K., it is modelled on arXiv and focuses mainly on papers in Psychology, Linguistics and Neuroscience.

Networked Computer Science Technical Reference Library (NCSTRL)

(http://www.cs.cornell.edu/cdlrg/ncstrl/cornell-oams.htm)

It is an international collection of computer science research reports. NCSTRL is based on a distributed model. Documents are stored in distributed archives and are made available through distributed services that communicate via the Dienst protocol.

The European sub-collection of NCSTRL is implemented by the ERCIM Technical Reference Digital Library(ETRDL).

Networked Digital Library of Theses and Dissertations (NDLTD)

(http://www.dlib.vt.edu/projects/OpenArchives/NDLTD_SFC.html)

It aims at building a digital library of electronic theses and dissertations (ETD) authored by students of member institutions. In ongoing research, NDLTD addresses issues such as the creation of a workflow to submit ETDs, the development of an XML DTD for ETDs and the support of a digital library for ETDs.

RePEc (RePEc)

ftp://netec.mcc.ac.uk/pub/RePEc/all/docu/RePEc_oai_provider_template.html

It is an initiative in economics that operates on a distributed model. It provides authors with the option to submit working papers to a departmental archive or -- if one does not exist -- to the EconWPA archive at Washington University. These archives support the so-called Guildford protocol that guarantees interoperability between the RePEc archives and has enabled the creation of a variety of end-user services.

Web Characterization Repository (WCR)

(http://repository.cs.vt.edu/oai.htm)

It is a database of meta-information relating to trace files, tools and publications that are relevant to characterisation of the World Wide Web. The project is managed by the W3C Web Characterisation Activity Working Group.

Computer Science Teaching Center (CSTC)

(http://www.cstc.org/OpenArchives/CSTC_sfc.html)