Date submitted: 03/06/2009

A Trust P2P network for the Access to Open Archive resources

Emanuele Bellini
Paolo Nesi
Department of Systems and Informatics
University of Florence,
Florence, Italy
belliniema@gmail.com, nesi@dsi.unifi.it
http://www.disit.dsi.unifi.it/

Meeting: **193. Information Technology**

WORLD LIBRARY AND INFORMATION CONGRESS: 75TH IFLA GENERAL CONFERENCE AND COUNCIL
23-27 August 2009, Milan, Italy
http://www.ifla.org/annual-conference/ifla75/index.htm

Abstract

The number of works made freely accessible on the Web by cultural institutions joined to Open Archive Initiatives such as libraries, archives and museums, along with research institutions and foundations, is constantly growing and it is becoming more complex. Unfortunately, these institutions are slow to offer innovative services and do not pay sufficient attention to the access functionalities required by users, whilst current technology solutions for wide dissemination and manipulation, like OAI Service provider implementation, are not enough. In this scenario, the benefits of the free access to resources is lost.

The proposed approach aims to increase resource circulation among users exploiting the P2P network and it provides a single application to search and access physical resources. Furthermore, the solution aims to transform the OAI resource into some intelligent content exploiting the structural richness and semantic awareness of the new OA complex digital objects, while preserving their authority and reliability. These new interaction possibilities with contents can stimulate any developing of new manipulation features according to each specific domain user requirements.

1. Introduction

The Web has drastically changed the information environment where users of the humanities work and study. This is particularly true in research and cultural heritage fields where the information needs authority, reliability and wide dissemination to become collective knowledge.
The usefulness of many on-line journals and scientific digital libraries existing today is limited by the difficulty of a) **accessing directly** and b) **manipulating** (like content enrichment) these resources through a c) **unique interface**.

The evolution of technologies also improved the Open Archive (OA) capabilities of storing more complex and structured contents, encapsulating resources and metadata as single entities. This has

caused the introduction of formats that are more complex, expressive, and accurate in their description of digital resources with respect to DC [1], MARC [2] and similar bibliographic formats as shown in [28]. Such formats are generally referred to as complex digital objects or reach media according to [4], [5], [26]; examples include MPEG21 [6], METS [7], SCORM [8], etc.

Unfortunately, in [9] Access to Digital Library project has shown that cultural institutions restrict themselves, only to permit passive access to digital collections, thus limiting the usage to the possibility of performing metadata searches and hyper-textual navigations inside the work or collections. From the user point of view, more functionalities are required such as manipulation of texts, possibility of concordances, and making notes and references to other works, etc. In this case, the organisational obstacles to accessing digital contents are linked together with the essentially legal and economical problems related to the access tools that still have to be considered quite unsuitable. In fact, a data provider often has records pointing to both freely-available and restricted-access digital resources and a web is required for each different provider.
To face these issues the proposed solution:

1) collects the OAI institutional repository resources in a trustable way, while preserving their authority and reliability
2) generates more intelligent content from OA resource for direct content manipulation
3) implements a 'scalable and distributed access point' going beyond the current non-homogeneous research and access tools.

The paper is organized as follows. Section 2 presents an overview of the state of the art of Open Archive implementations and the main rationales and motivations for using P2P solutions for content sharing. Section 3, the AXMEDIS architecture overview for content distribution and monitoring, is presented in terms of high level functionalities allowing any control on the P2P architecture and facilities. Section 4 presents the architecture proposed and the interactions among components. Section 5 describes some experimental results. Conclusions are drawn in section 6.

2. Open Archive overview

The OAI architecture [10] identifies two logical roles: "Data Providers" and "Service Providers". Data Providers deal with both the deposit and publication of resources in a repository and they "expose" for collecting the metadata about resources in the repository. They are the creators and keepers of the metadata and repositories of resources. At present many institutions have implemented the OAI Data Provider also to manage complex resources [27] [3], thus choosing the following repository software: Dspace [11], Fedora [12], Eprints [13], etc.

Service Providers use the OAI-PMH interfaces of the Data Providers to collect and store their metadata. They use the collected metadata for the purpose of providing one or more services across all the data. The types of services, which may be offered, include a search interface, peer-review system, etc. The key architectural shift was to move away from only supporting human end-user interfaces for each repository, in favour of supporting both human end-user interfaces and machine interfaces for collecting.

There are several service providers as CiteSeer [14] (information science), RePec [15] (research paper in economics), Pleiadi [16] (OAI resources),etc. One of the most important Service Providers is OAIster [17]. OAIster is a union catalogue of digital resources. It provides access to digital resources by "collecting" their descriptive metadata (records) using OAI-PMH on thousands of contributors. The proposal has tried to eliminate the so called 'dead ends' (collected records which do not link to an accessible digital resource) of the query results provided by OAI service providers.

In fact users retrieve not only descriptions about resources, but they have access to real digital resources thorough the URL of the access pages of CMS (i.e., http://aei.pitt.edu/7400/). The first problem is that the physical access to the resource is not provided in the same request action, and the user may see different user interface, from portal to CMS for each resource. Please note there are no guarantee of successful access because the local server could be out of order or the record is no longer updated on service provider.

Besides, Service provider solutions force the user to "jump" among different portals to get the resources not found on a first portal. As a matter of fact, in the OAI architecture, each portal or service provider collects metadata only from the data provider chosen and available on line. In fact, the www.openarchives.eu (http://www.openarchives.eu/home/home_do.aspx) service implements a unique interface for searching a digital object by selecting the service provider from a list.

The analysis has been integrated also with the results of the Access to Digital Libraries project [9] funded by Fondazione Rinascimento Digitale. The Final Report sketches out the state of the art of digital libraries created by cultural institutions in Italy, in particular the focus was on standards and related good practises.

In further details the report has outlined what follows:

a) the lack of a unique user interface prevents the retrieval of resource, thus forcing the user to "jump" among different web portals and making difficult for users to identify what is available on line.
b) It is difficult to sustain a unique global meta-catalogue for all OAI resources in the world.
c) Getting access to the OAI physical resources is not guaranteed and it is often discouraged or even prevented by web page design of CMS.
d) Features for exploiting the structural and semantics richness of the content are missing.

At present, the user still uses general purpose search engines such as Google which gives no guarantee of reliability for the retrieved information. In fact, robots in use by such search services do not delve into the CMS-CGI that sends this resource information to the web because of the presence of conditional access mechanism [18]. As a consequence, such resources are generally made accessible only to those who know how to look into particular repositories or service providers.

3. AXMEDIS overview

The proposed solution is based on the AXMEDIS framework (http://www.axmedis.org) [19]. The AXMEDIS is the ultimate solution for cross-media content business. The novel AXMEDIS approach provides tools for an efficient and valuable content management: from production, elaboration, management and protection, to a real multi-channel distribution. AXMEDIS Framework supports and automates the digital content management along the whole value chain, while offering an enlarged set of tools able to satisfy even the most complex business requests. With AXMEDIS multi-channel distribution the same content can be distributed for streaming or downloading for different platforms and channels: Satellite, PC Windows or Mac, PDA, Mobile, Set-Top-Box, IP-TV, and many other systems.

3.1 Trusted P2P: AXP2P network

The AXMEDIS P2P network is an open and scalable solution for setting up P2P networks for multimedia content distribution and sharing. In particular, AXP2P solution allows content owners

and distributors to exploit the capabilities of P2P protocol to create efficient, controllable, legal and secure P2P networks for content distribution and sharing. By using AXP2P solution a distributor may publish content in the P2P network and the content may navigate freely among peers, with the supervision and control of the AXMEDIS protection and monitoring tools. The AXP2P is based on the BitTorrent technology with several innovations to enable actions such as: monitoring the network, managing DRMed content, making queries, distributing MPEG-21 objects.
The AXMEDIS P2P solution is based on the following active elements.

AXTracker is a modified BitTorrent Tracker managing the AXMEDIS P2P network and community. The tracker is used to publish new content by posting on it BitTorrent information and obtaining IDs for nodes and files. Main AXMEDIS Tracker's features are: tracking of BitTorrent files; catalogue of tracked files; tracker's download statistics. AXTracker serves as an authoritative source for listing the available files on the network and for tracking statistics. In the global geographic system, many AXMEDIS BitTorrent Trackers may survive to provide services and share mutually information regarding nodes and content files, so as to distribute the workload and make it fault tolerant. Thanks to the AXTracker, the AXEPTool /AXMEDIA P2P client tools are able to know where they can retrieve the segments building a file.

AXEPTool is a specific P2P BitTorrent Client Node, suitable to play the role of a P2P Node for B2B activities such as producers, distributors, integrators, etc., for B2B content distribution. It is able to accept automated requests of publishing, downloading, monitoring, etc., from other tools; for example, from an AXCP GRID node in the AXMEDIS Content Production/Distribution Factory. These facilities have been included to allow the P2P network controllability. In addition, the AXEPTool accepts also manual commands of publishing, downloading and monitoring via a graphical user interface.

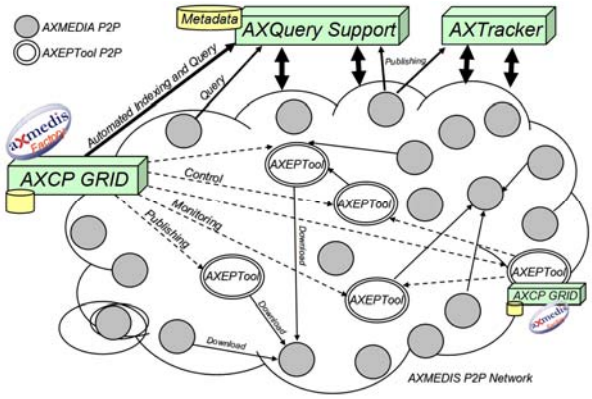

Figure 1 – AXMEDIS P2P Architecture

AXMEDIA is a specific P2P BitTorrent Client Node for final user content sharing and B2C (Business to Consumer) content distribution. Internet distributors may use the P2P Network to reduce costs for scaling the infrastructure for content distribution. AXMEDIA does not include some of the AXEPTool's features dealing with publication, downloading and activity monitoring. Therefore, it is easy to use, lighter to be executed and more easily acceptable by final users.

4. Proposed architecture

Our approach comes from the success of P2P networks for searching, accessing and distributing information. The P2P approach provides a unique view and a simple access point of heterogeneous

resources, it solves the repositories cross-search issue and increases the digital preservation capability thanks to the duplication of the resources. Until now, the use of the P2P network for distributing trusted Digital Object has been prevented due to a certain lack of capabilities to:

a) certify content providers that publish their resources on the network
b) certify digital copies against the original digital resources
c) Intellectual Property Right (IPR) management via DRM or other means.

The proposed solution is based on the AXMEDIS framework and defines a production flow (see figure 2) from the content managed by OAI Data provider to the OAI resource distributed on a trusted P2P network. This solution redefines the Data provider roles and introduces a new actor in the OAI architecture: the AXMEDIS **Content Production node** (AXCP GRID).

The role of the Data Provider consists in:

a) selecting and preserving the resource to be managed,
b) certifying the resources either exchanged (or downloaded/published) through P2P network,
c) providing the resource to the content production, for packaging in a new digital object for multi-channel distribution.

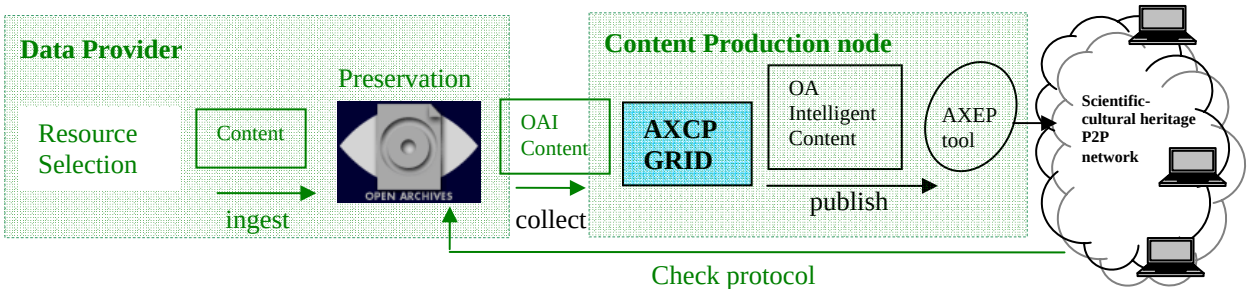

Figure 2 Content Production flow

In particular, with the proposed solution, content could be provided only once to the AXCP, then the dissemination/publication and access activities are delegated to the AXMEDIS P2P infrastructure.

Data providers manage different kinds of resources in terms of types (digitized books and articles, born-digital texts, audio files, images, movies), file formats (mp4, quicktime, wav, mp3tiff, gif PDF, PDF/A, HTML, etc.), and complexity (compound object or multi resources record).
The solution collects by OAI-PMH the simple (Dublin Core) or complex (MPEG-21, METS) [20], metadata and related physical resources from OA. Then, the system bundles the resource and its metadata together in an MPEG-21/AXMEDIS object, mapping the DC or other schema into a descriptor item in the MPEG-21 and it may encapsulate any other metadata sets via additional descriptors (see figure 3).

This new OA content may include any kind of digital resources, behaviours and also some presentation interactive layers based on HTML, SMIL or other formats. The model also enables to wrap into the object some degree of intelligence in terms of semantics and java script processing methods to be run at certain conditions and events [21]. In fact, to increase the accessibility to resources it is necessary to increase the number of copies available on the network, while preserving digital object reliability and authority given by Digital Library responsible. To this end a check protocol with the Digital Library responsible for the master copy of Digital Object has to be

arranged according to the Persistent Identifiers (PI) technologies [22] (as the National Bibliography Number [23][24]). It is reasonable to assume that each Digital Object managed by an institutional Open Archive could have a credible Persistent Identifier assigned, because it is a basic requirement for the reliability and long time preservation of that resource.

The intelligent object connecting to the appropriate resolution service resolves the PI to the master copy of that digital object and with a mathematics (hash) and/or semantics comparison (metadata) it verifies its integrity. In this approach, the Digital Library is covering the role of reliability "supervisor" (see figure 3).

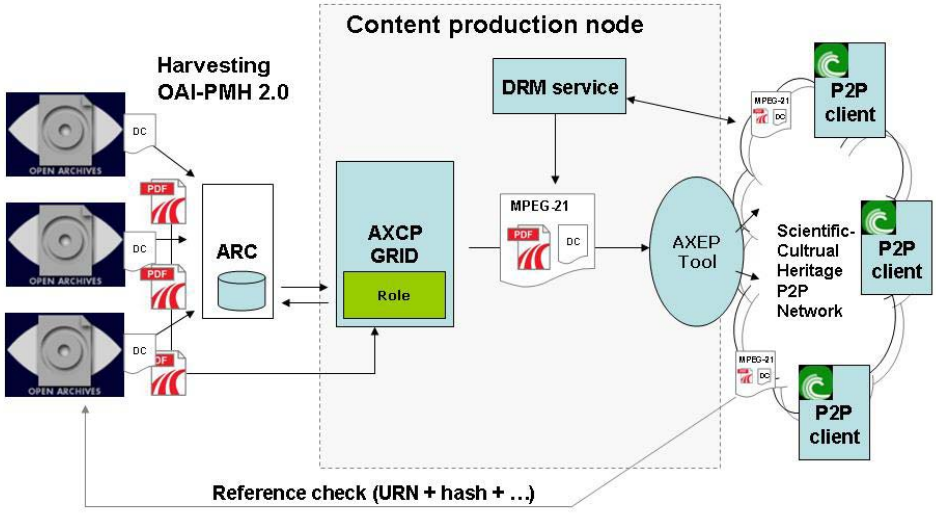

Figure 3 Proposed Architecture trusted P2P network for the access to Open Archive resouces.

The system is able to distribute the AXMEDIS/MPEG 21 objects with DRM or not, according to specific resource access policies of each OA. In particular, it is possible to protect only a part or the entire set of MPEG21/AXMEDIS objects. This feature is useful to manage the IPR according to the data provider's distribution policies and user profile. Furthermore, it can be used to manage separately the original content of a digital library, with respect to the user generated content added by end user such as: annotation, tags, translation, etc.

Besides, the content production node AXMEDIS can be replicated in order to distribute the load of the content production (see figure 4). In particular, a service provider that has already linked a number of Data providers can easily implement a content production node by connecting the data provider to the node.

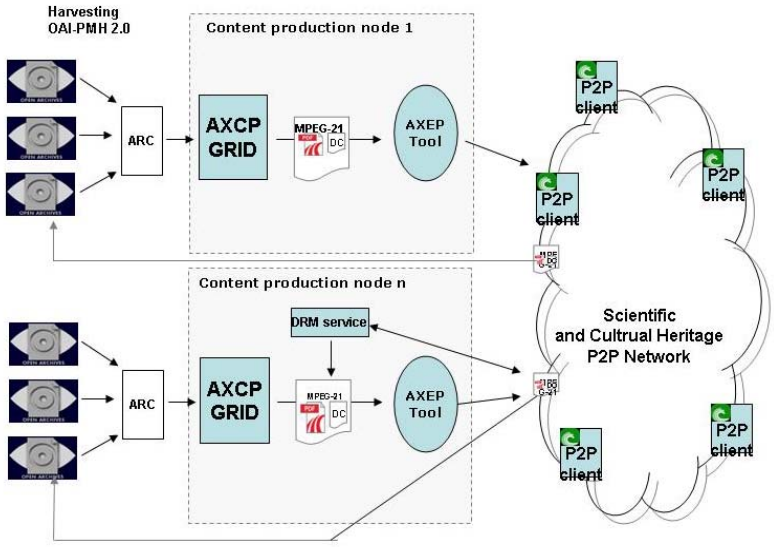

Figure 4 Scalability of the AXMEDIS infrastructure

As presented in [21], once the user has obtained the OA intelligent content, this content allows and helps the user to:

- make queries internally to the content on the basis of the content.
- provide information in a proactive manner, such as user generated content. Once provided, the intelligent content may elaborate to produce some integrated intelligent content for the user: a collection, a sliding show, a guide.
- provide cross media or multimedia annotations on the basis of some early content or from scratch, thus creating collections in either simple or complex structures, with multiple files of different types and their annotations and links/relationships.
- build cross media or multimedia scrapbooks, with annotations and user generated content, etc.
- exchange annotations and OA complex content among end-users. They can be annotated guides, scrapbooks, annotated school book, collections of mixed files of different natures.
- interact with these collections of intelligent content in order to (i) migrate them to other devices, (ii) organize them into local database, (iii) share them with other users, etc.

6. Experimental results

The experiment has exploited the Arc module for implementing the collecting engine. Arc (http://arc.cs.odu.edu) is the first federated search service based on the OAI protocol. It originates from the Universal Preprint Service (UPS) prototype (Van de Sompel, Krichel, Nelson, et al., 2000), which was developed as a proof-of-concept and discussion piece for various digital library technologies, including the feasibility of constructing a cross-archive searching service [25].
Similar to a web crawler, the Arc harvester traverses the data providers automatically and extracts metadata. The significant differences include: normalizing the metadata, thus producing more complete and accurate results; and exploiting the incremental, selective collecting defined by the OAI protocol.

The experiment is based on the resource and metadata collecting from the EPrints of University of Florence (http://eprints.unifi.it/index.php) and the DSpace of University of Parma (http://dspace-unipr.cilea.it/).

The URL of the resource is specified in dc.format but the lack of a general mechanism is well known and it is crucial for applications requiring the gathering of content, not only metadata [28].
The experimentation has started with the metadata collecting in DC format from the Data provider using the Arc module. The Arc module saves a record in its database for each OAI record collected with the date-stamp. The Arc module stores the metadata collected in a single field of the database record in XML format. Then, the javascript production role starts on the AXCP node following these steps:

a) it reads from the Arc database only the new record generated by the collecting repository and it parses the field which contains the OAI-PMH record in XML format.
b) it creates an AXMEDIS Object and maps the OAI DC in the MPEG21 item.
c) For each <dc:format> field found, the role extracts the URL of the resource and downloads it (<dc:format>pdf http://www.unifi.it/abc/doc.pdf</dc:format>). If the resource is an HTML file, the role crawls all external resources as GIF file, through the URLs found in the page.
d) it packages the physical resources in the MPEG21 AXMEDIS object By-Value: embedding a base64-encoding [29] of the datastream inside the wrapper XML document.
e) it requires a licence for the new intelligent object.
f) Finally it publishes the object on AXEP Tool.

Users can perform many types of search on their P2P client. With the AXMEDIS P2P solution, it is also possible to have facilities to make queries into the P2P network. This search facility is provided for MPEG-21 objects and it is based on Dublin Core metadata and classification model plus additional business information such as licensing information, distributor information, etc. The user can browse and query the content obtained from the P2P and he can enrich it with tags and annotations.

7. Conclusion

AXP2P solution allows the creation of efficient, controllable, legal and secure P2P networks for content distribution and sharing. By using the AXP2P solution an OA may publish content in the P2P network and the content may navigate freely among peers with the supervision and control of the AXMEDIS protection and monitoring tools. With the AXP2P solution it is also possible to make queries based on metadata, classification model or additional business information into the P2P network for MPEG-21 objects.

In this approach, the Digital Library plays the role of “supervisor” on the reliability of the copy required, thanks to a check service based on hash and persistent identifiers technologies.
Final users have a direct access to certified resources and they may be stimulated to create and share their collection, setting the conditions for developing a certified Personal Digital Library, in the respect of IPR.

8. References

[1] Dublin Core Metadata Initiative. Dublin Core Metadata Element Set, Version 1.1. http://dublincore.org/documents/dces/.
[2] MARC ISO 2709 http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=41319
[3] Jeroen Bekaert, Patrick Hochstenbach, Herbert Van de Sompel, "Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library", D-Lib

Magazine, Nov 2003 http://mirrored.ukoln.ac.uk/lis-journals/dlib/dlib/dlib/november03/bekaert/11bekaert.html
[4]Kahn, R. and Wilensky R. “A Framework for Distributed Digital Object Services. Corporation for National Research Initiatives,” http://www.cnri.reston.va.us/k-w.html, 1995.
[5] Carl Lagoze The Warwick Framework A Container Architecture for Diverse Sets of Metadata - Digital Library Research Group, Cornell University, D-Lib Magazine, July/August 1996,ISSN 1082-9873
[6] MPEG-21, Information Technology, Multimedia Framework, "Part 2: Digital Item Declaration," ISO/IEC 21000-2:2003, March 2003.
[7] METS, <http://www.loc.gov/standards/mets/>
[8] Advanced Distributed Learning, "The Sharable Content Object Reference Model (SCORM) - Version 1.3 - WD," March 2003
[9] Access to Digital Library http://www.rinascimento-digitale.it/indexEN.php?SEZ=494
[10] C. Lagoze and H. V. de Sompel. The Open Archives Initiative Protocol for Metadata Harvesting, version 2.0. Technical report, Open Archives Initiative, 2002. http://www.openarchives.org/OAI/openarchivesprotocol.html.
[11]. DSpace <http://www.dspace.org/>.
[12] Fedora <http://www.fedora.info/>.
[13] EPrints for Digital Repositories <http://www.eprints.org/>.
[14] Citeseer http://citeseer.ist.psu.edu/
[15] RePec http://repec.org/
[16] Pleiadi www.openarchives.it/pleiadi/
[17] OAIster, 2005. OAIster Home. <http://oaister.umdl.umich.edu/o/oaister/>.
[18] Conditional access system CAS - http://en.wikipedia.org/wiki/Conditional_access_system
[19] AXMEDIS: Automating content f Cross Media Content for Multichannel Distribution http://www.axmedis.org
[20] Herbert Van de Sompel et al. Resource Harvesting within the OAI-PMH Framework. D-Lib Magazine December 2004 Volume 10 Number 12 ISSN 1082-9873
[21] P. Bellini, P. Nesi, I. Bruno, M. Spighi - Intelligent Content Model based on MPEG21 AXMEDIS 2008 Conference
[22] H.-W. Hilse, J. Kothe Implementing Persistent Identifiers: overview of concepts, guidelines and recommendations, 2006, ix+57, pp. 90-6984-508-3 http://www.knaw.nl/ecpa/publ/pdf/2732.pdf
[23] J. Hakala. Using national bibliography numbers as uniform resource names, 2001. (RFC3188) http://www.ietf.org/rfc/rfc3188.txt.
[24] E. Bellini , C. Cirinna, M. Lunghi, E. Damiani, C. Fugazza, 2008 Persistent Identifiers distributed system for cultural heritage digital objects, IPRES2008 conference
[25] Xiaoming Liu et al. « Arc - An OAI Service Provider for Digital Library Federation” D-Lib Magazine April 2001 Volume 7 Number 4 ISSN 1082-9873
[26] Wernher Behrendt, Guntram Geser, and Andrea Mulrenin. Ep2010 - the future of electronic publishing towards 2010. Information Society DG - Unit E2, EUFO 1-275
[27] Jeroen Bekaert, Emiel De Kooning, and H. Van De Sompel. “Representing digital assets using mpeg21 digital item declaration” International Journal on Digital Libraries 6(2)
[28] Herbert Van de Sompel et al. Resource Harvesting within the OAI-PMH Framework D-Lib Magazine December 2004 Volume 10 Number 12 ISSN 1082-9873
[29] Freed, N. and N. Borenstein. 1996. "RFC 2045: Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies," November 1996. <http://www.ietf.org/rfc/rfc2045.txt?number=2045>.