

## Open Archives Initiative – Protocol For Metadata Harvesting

### Practices of cultural heritage actors

September 2003

Editor : Muriel Foulonneau (Relais Culture Europe, France)

Contributor : David Dawson (Resource, United Kingdom)

*A special thanks to Martin Sévigny (AJLSM), Sara Aubry and Thierry Cloarec (French National Library)*

<b>Project Number:</b>	IST-2001-32015
<b>Project Title:</b>	Open Archives Forum
<b>Deliverable Type:</b>	Public
<b>Deliverable Number:</b>	D4.8
<b>Date of Delivery:</b>	September 2003
<b>Title of Deliverable:</b>	Expert Report 3 – Open Archives Initiative – Protocol for Metadata Harvesting - Practices of cultural heritage actors
<b>Workpackage contributing to the Deliverable:</b>	WP4
<b>Nature of the Deliverable:</b>	Report
<b>URL:</b>	<a href="http://www.oaforum.org/otherfiles/oaf_d48_cser3_foullonneau.pdf">http://www.oaforum.org/otherfiles/oaf_d48_cser3_foullonneau.pdf</a>
<b>Author:</b>	Muriel Foulonneau (Relais Culture Europe, France) with David Dawson (Resource, United Kingdom)
<b>Contact Details:</b>	Philip Hunter, UKOLN The University of Bath, Bath BA2 7AY, UK. URL: <a href="http://www.ukoln.ac.uk">http://www.ukoln.ac.uk</a>
<b>Abstract</b>	This report is about the issues raised in the cultural heritage sector by the OAI-PMH, how several people have handled them or set hypothesis to handle them. It gathers a series of lessons learned, issues to take into account and practices to assess about the use of OAI-PMH for cultural resources.
<b>Keywords</b>	cultural heritage, open archives, Open Archives Forum, Open Archives Initiative, metadata sharing.
<b>Distribution List:</b>	Public, via OA-Forum website; OA-Forum partners & Project Officer
<b>Issue:</b>	1
<b>Reference:</b>	32015-DEL4.8-20031001
<b>Total Number of Pages:</b>	56

## Table of contents

1	Introduction.....	4
1.1	The OAI-PMH in the cultural heritage sector .....	4
	▪ Metadata harvesting and aggregation .....	4
	▪ Cultural heritage metadata on the Web .....	4
	▪ Underlining the value of memory organisations' intellectual asset .....	5
1.2	Methodology and intended audience of this report.....	6
2	Metadata harvesting in the cultural heritage sector .....	7
2.1	Main principles .....	7
	▪ Data providers, service providers, aggregators .....	7
	▪ Harvesting .....	7
	▪ Metadata representation .....	8
	▪ Interoperability is not solved.....	8
2.2	Use of OAI-PMH for the cultural heritage sector .....	8
	▪ Make resources accessible to the Web .....	9
	▪ Cross-institution and cross-domain services.....	9
	▪ Data-reusability .....	9
	▪ Standard for data interchange.....	10
	▪ Downstream services.....	10
2.3	The OAI model to build heritage repositories and services .....	10
	▪ Shareable and disclosed models for data provision.....	10
	▪ An asynchronic model .....	11
	▪ An economic and organisational model for resource discovery .....	11
3	Building a repository .....	13
3.1	Setting records .....	13
	▪ Many heritage standards are not proper XML schemas .....	13
	▪ The OAI-DC schema .....	13
	▪ More complex schema .....	14
	▪ Managing relationships.....	14
	▪ Mapping tasks .....	15
	▪ Identifiers.....	16
	▪ Defining sets for selective harvesting .....	17
	▪ Data update .....	17
3.2	Technical issues.....	18
	▪ Defining the system and architecture.....	18
	▪ Security and access rights .....	19
	▪ Data delivery .....	19
3.3	Communication and management .....	21
	▪ Budget issues .....	21
	▪ Available competences .....	22
	▪ Register the repository.....	23
	▪ Publish information on the repository.....	23
	▪ User and evaluation.....	24
4	Building a service .....	25
4.1	A service for cultural heritage resources.....	25
	▪ Services for accessing aggregated content.....	25
	▪ Setting a community.....	25
4.2	Cross-collection content is always heterogeneous.....	27
	▪ Data analysis .....	27
	▪ Access points .....	28
	▪ Display issues.....	30
	▪ Common metadata schema .....	31
	▪ The granularity of descriptions .....	32
	▪ Metadata crosswalks.....	33
	▪ Terminology issues .....	34
	▪ A multilingual environment.....	36
4.3	Resource aggregation.....	36

▪	Technical features .....	36
▪	Harvesting frequency to improve synchronization .....	36
▪	Harvesting profiles .....	37
▪	Reprocessing content .....	38
▪	Managing aggregated collections.....	39
4.4	Displaying information on the service .....	39
▪	Information reliability .....	39
▪	Integrated access to resources.....	40
▪	Aggregation procedure for newcomers.....	41
▪	User approach and service evaluation .....	42
5	Conclusion .....	43
	Annex 1 – Glossary.....	45
	Annex 2 – Bibliography .....	46
	Annex 3 – Tricks, traps and issues.....	50

## 1 Introduction

For several years, the cultural heritage sector is facing a challenge to represent the wealth of its resources on the Internet and benefit from the potentialities of information technologies. As all sectors had to consider similar issues, digitisation has led to new opportunities for cooperation between institutions and between the various heritage sectors.

On the Internet, the user could access numerous heritage resources from different sectors in various formats. A user-centred approach to Internet services should take into account the necessity to cope with these multiple pieces of cultural knowledge and set suitable services to successfully access digital cultural assets.

“The DigiCULT-study in particular highlights the importance of co-operation in creating value added services and rich environments for broader user groups as well as fostering more cross-domain co-operation of cultural heritage institutions.”<sup>1</sup>

The cross-collection approach exists for a long time notably in the library sector to build resource discovery services (union catalogues). Many institutions are orienting their strategy towards a cross-domain approach. The OAI protocol for metadata harvesting (OAI-PMH) represents an opportunity to support those services in the cultural heritage sector.

### 1.1 *The OAI-PMH in the cultural heritage sector*

The OAForum project aims at exploring the applications of the OAI-PMH in the heritage sector and providing tools and means for fostering its development. For several years, the protocol appears of uppermost interest for memory organisations, although many of them still adopt a “wait and see” approach<sup>2</sup>. This report shall therefore help defining a series of practical issues raised to implement an OAI-PMH architecture in the cultural heritage sector.

- **Metadata harvesting and aggregation**

The Budapest Initiative for Open Archives, associated with the development of the protocol, is only of concern to the ePrints community which is facing specific issues and has proposed an economic and organisational model to produce and disseminate grey literature.

The memory organisations rather consider the OAI-PMH as a protocol which facilitates resource discovery on the Web and underlies a specific organisational model for accessing cultural heritage resources and exchanging metadata. Several ePrints actors (University libraries) are also developing similar architectures for their heritage resources.

The OAI-PMH offers a technology to support the organisational evolution of digital cultural content creation and collective services to access cross-domain collections. That technology is based on metadata which are core wealth of memory organisations. It allows to “aggregate” the metadata of heritage resources in a central location while leaving the resources under the responsibility of the creator (cultural institution in charge) or of a specialised service provider which can maintain the digital surrogates or their metadata.

- **Cultural heritage metadata on the Web**

The cultural heritage sector is composed of archives, libraries, museums, monuments, archaeological sites, galleries ... To manage, preserve and provide access to their heritage asset, the memory organisations create rich metadata, very important to retrieve resources, particularly non textual resources.

<sup>1</sup> Mulrenin Andrea, “The DigiCULT Report Technological Landscapes for Tomorrow’s Cultural Economy Unlocking the Value of Cultural Heritage”, European Communities, 2002, [ftp://ftp.cordis.lu/pub/ist/docs/digicult/executive\\_summary\\_en.pdf](ftp://ftp.cordis.lu/pub/ist/docs/digicult/executive_summary_en.pdf), p 38

<sup>2</sup> see Seamus Ross and Ian Anderson, “Discovering Good Practice: Metadata and the NINCH Guide”, talk in OAForum Berlin workshop March 2003, [http://www.oaforum.org/workshops/berl\\_abstracts.php#ninch1u](http://www.oaforum.org/workshops/berl_abstracts.php#ninch1u)

The memory organisations have sometimes connected their catalogues and finding aids to the Internet, but the main challenge and incentive have appeared while they were digitising their material and as new cultural content, born digital, was created. Then online access to cultural heritage resources could provide a much improved service to the public and help enlarging audience.

However, the content of online databases is not indexed by search engines, although they are estimated 400 to 500 times the content reached by search engines<sup>3</sup> so that it has become a challenge for memory organisations to improve access to those resources, thanks to formats adapted to the Web.

The presence of memory organisations on the Web has also led to present information in new contexts, notably to relate information from various cultural heritage sectors, thus providing new information and adding value to those assets through new modes of access. This cross-domain approach has increased over time and new formats and standards are emerging to improve access to cross-domain heritage resources, especially based on metadata.

This evolution is driven either by a philosophy of enlarged access to resources, or the opportunity to build value-added services for communities which work similarly (union catalogues for example for libraries), or by the opportunity to set cross-institution (and possibly cross-domain) platforms with user-driven approach for access to resources, wherever it comes from.

In this context, the actors and stakeholders of the new services to discover heritage resources may be different. Many gateways to cultural heritage material are not set by memory organisations, although they used to be the only ones to conceive and set up access services to their collections. This trend may appear more and more significant with the emerging technologies for resource discovery and the multiplication of cross-domain access services on the Web.

This allows small memory organisations to benefit from other people's competencies to publish their content. A protocol such as the OAI-PMH can allow small institutions' collections to be largely visible and accessible in value-added services, in new sectors, gain new public with a single relatively low investment which is to set up an OAI-PMH repository. They may then concentrate on their core know-how to furnish high quality metadata and possibly on an advanced and specialised way of disseminating their asset, such as specialised databases, including extensive descriptions of material and thesaurus-based search functionalities. Other types of services could be built (if the memory organisation does not have the human and/or financial resources to do it) by other institutions, whether commercial, academic, agencies....

This involves that the exposition of memory organisations' metadata, which are their invaluable intellectual asset, lead to more or less formal "contracts" between the memory organisations as content providers and service builders. Those may be other memory organisations or other types of institutions. These "contracts" include the commitment by data providers (memory organisations) to provide high quality data and by service providers to add value to this data and use them in the conditions allowed by data providers.

- **Underlining the value of memory organisations' intellectual asset**

Usage of descriptive metadata created by the memory organisations can lead to set up high quality services such as making heritage resources accessible by search engines, capitalising on union catalogues, reusing cultural heritage resources in other environments and building catalogues to retrieve and/or store heterogeneous resources.

Nevertheless, the potential applications of the OAI-PMH in the cultural heritage sector may not be limited to aggregation of metadata from catalogues and finding aids. It can help discovering editorial products and value-added services built by memory organisations to broaden their audience. The issue of OAI-PMH is not restricted to databases-style applications but also Websites, including virtual exhibitions and galleries, or other types of resources, on users, terminologies... with the only condition that this information is structured.

The OAI-PMH architecture appears of major interest to improve user access to digital cultural content. It could usefully be integrated in the digitisation strategies of cultural institutions and several funding

---

<sup>3</sup> Evaluation mentioned in Bergman K. Michael, "The Deep Web: Surfacing Hidden Value", The Journal of electronic publishing, University of Michigan Press, 2001, <http://www.press.umich.edu/jep/07-01/bergman.html>

programmes are considering the possibility to include the “open access” to online material created so as to re-build large distributed digital libraries from the investment they make<sup>4</sup>.

## **1.2 Methodology and intended audience of this report**

The present report is not a guide to interoperability.... though it is impossible to avoid dealing with interoperability issues. This is not a report on how to make cultural heritage resources available on the Web (protection issues, how to publish images or sound, defining IPR..), although applications of the OAI-PMH in the cultural heritage sector may be related to the publication of resources.

This report is about the issues raised in the cultural heritage sector by the OAI-PMH, how several people have handled them or set hypothesis to handle them. It gathers a series of lessons learned, issues to take into account and practices to assess about the use of OAI-PMH for cultural resources.

The use of the Open Archives initiative protocol for metadata harvesting raises on the one hand functional issues of cross-collection access and management of aggregated resources and on the other hand technical issues related to metadata harvesting rather than another type of cross-collection application. Consequently, the examples provided do not all use OAI-PMH but may be experiments of cross-collection services.

This report has been done with :

- experiences gathered from several institutions (interviews, meetings, mailing lists) of OAI-PMH and cross-collection access;
- work on organisational issues raised at European level, with partners involved in designing national strategies for digital cultural content creation, notably the discussions of the National Representatives Group on digitisation of cultural and scientific heritage (NRG)<sup>5</sup> and the MINERVA project which coordinates technical working groups for preparing the political decisions adopted by the NRG<sup>6</sup>;
- articles and presentations, exposing experiences and lessons learned in various contexts.

This material is not enough to make a proper guide to good practices but it pretends to help discovering key issues for using the OAI-PMH in the cultural heritage sector. It is desirable that this document gets enriched by larger community's experience.

This report is intended to institutions of the heritage sector which consider getting involved in an OAI-PMH architecture, project managers, and decision-makers which think at launching such a project. It pays a particular attention to defining how to get small or medium cultural institutions benefit from such an architecture. Indeed, many large institutions with extended competencies and international networks are already aware of the potentialities of the OAI-PMH for them and should help disseminating their know how. This report shall focus on the benefits to smaller institutions and the opportunity to support collaborations between institutions for providing enhanced access to digital cultural content.

---

<sup>4</sup> Dawson David, “Open Archives Initiative, Metadata Harvesting and the NOF portal – an information paper from the NOF technical advisory service”, <http://www.ukoln.ac.uk/nof/support/help/papers/oai-pmh>

<sup>5</sup> <http://www.cordis.lu/ist/ka3/digicult/eeurope-overview.htm>

<sup>6</sup> <http://www.minervaeurope.org/>

## 2 Metadata harvesting in the cultural heritage sector

The OAI-PMH is first of all a protocol which underlies an architecture based on metadata harvesting. This process is a simple way of collecting data. An operator, the *data provider*, prepares its data for another operator, the *service provider* who can use those data in a specific representation, the metadata schema.

### 2.1 Main principles

Functionally, the protocol allows to centralise metadata describing various resources but leaves the resource in its original location. It is a “third-party metadata model” which provides access to distributed resources by gathering metadata and exploit them according to the specific needs of a service.

- **Data providers, service providers, aggregators**

The OAI-PMH defines various roles in an architecture built on metadata. A *data provider* makes its metadata available for use in one or more description formats. From the common finding aids, records are made available so as to match a formal XML schema (ie Dublin Core, METS... or possibly made for that specific purpose). A *service provider* launches a programme called *harvester* to visit a data provider and collect metadata in the format it requires, if this is available, at least in unqualified Dublin Core. The service provider processes the metadata gathered and offers a service based on those metadata.

From the *OAI repository* (repository of metadata on the data provider machine), metadata can be available in various formats to match various requirements. For example a general service provider may only need Dublin Core format, while a service for library information will harvest full MARC descriptions.

A third type of operator may, in certain cases, intervenes in this configuration. An *aggregator* gathers metadata from various data providers and make them available in an OAI repository, possibly after processing them to improve their quality for building added-value services. The aggregators can guarantee a better quality of metadata, ensure a common storage, perform a first set of normalisation tasks.

- **Harvesting**

OAI-PMH only manages data transfer, it is not a cross-searching protocol since it does not support querying functionalities. However, querying can be processed when making the OAI repository or within the OAI service by rebuilding a finding aid on the server of the service provider. Cross-searching functionalities may be added and used as complementary functions to an OAI architecture.

Data collection includes all records added, modified, deleted from last harvest, thanks to timestamps of modification and creation and keeps track of deleted items. The harvesting process does not collect all data but information on repository modification since last harvest. The timestamps recorded in the repositories are the key elements on which the harvesting process relies. Consequently, the repository updating process should adequately modify records so that harvesters can understand the actual modifications since last harvest.

A scheduler may be implemented to harvest metadata which have been modified on a regular basis. The schedule shall depend on the type of repository and its update frequency.

The protocol allows to define the quantity of data being transferred, to divide element sets to be transferred in order to perform the harvest in several steps if the amount of data to transfer is too large. The OAI-PMH is based upon HTTP protocol so that additional functionalities can be implemented for transferring data, for example data encryption or data compression to make the transfer more fluent<sup>7</sup>.

---

<sup>7</sup> not defined in the protocol but as an example, the Physnet service from Oldenburg university in Germany has implemented it

- **Metadata representation**

The data provider defines the representation mode of data for a specific harvester, which is XML encoding according to a formal XSD schema. Indeed, the protocol defines that the unqualified Dublin Core (DC) is the minimum schema to implement : every repository must be harvestable in DC, but it is possible to represent metadata according to various schemas. A single item can be harvested with 1 unqualified DC record, one MARC-XML record and one MODS record, all three applied to resources held in libraries. The OAI protocol allows to manage multiple representations of a single metadata set to be used in different environments.

The OAI-PMH also allows to define *sets* of items in a repository. Data can be extracted from catalogues on a regular basis, to be integrally copied in another location. Alternatively, the protocol controls the possibility to define data which can be harvested by specific services. The harvester then queries the repository for the sets it uses and only collects the necessary records.

- **Interoperability is not solved**

Still, the protocol does not solve all issues raised to process material from various origins in a single application.

The OAI protocol defines a standard way for memory organisations of making its resources available to other institutions. The Openarchives Website<sup>8</sup> provides various software components to implement for making its system OAI compliant but this does not solve all situations. It is a standards way of reaching technical interoperability for catalogues and finding aids but the necessary tools are not yet available for all types of applications used in the cultural heritage sector.

Moreover, it does not provide any solution to organisational interoperability which should allow to synchronize content creation and build a similar way of working which facilitates mutualised services. The OAI-PMH implies the existence of a community with coherent organisational practices.

Finally, the protocol does not take into account semantic issues (the fact that metadata sets in use are not similar and not used in the same way), syntactic issues (metadata values not encoded in the same way) and granularity issues (the metadata do not describe the same "resources", whether a collection, an item, part of an item....).

Still, the OAI-PMH, by providing a standard and simple way to cope with technical interoperability, can be a facilitator. Semantic and organisational interoperability may be much more complex issues to face for working on cross-collection services, and moreover cross-domain services, although it is possible to benefit from several ongoing experiences of services based upon aggregated resources.

## **2.2 Use of OAI-PMH for the cultural heritage sector**

In the OAI architecture, the role of memory organisations are at least content providers so that they must act as data providers, optionally as service providers or as aggregators.

All information types created within the cultural sector may appear interesting to exchange and expose, whether descriptions of locations to visit and relate to other touristic information, events published on cultural Websites, articles for professionals, descriptions of Websites from cultural gateways, terminological tools and indexes (thesaurus, name authorities) .... However, memory organisations have a major interest in making available finding aids and catalogues, preferably when digital surrogates or born digital resources are available online.

---

<sup>8</sup> <http://www.openarchives.org>

*The US Digital Libraries Federation (DLF) has defined hypothesis for interesting use of OAI-PMH in the cultural sector<sup>9</sup> :*

- searching information on a specific subject;
- access information in a specific format (eg. EAD, TEI);
- services developed by an institution or a particular network to be presented in an environment specifically conceived for specific needs;
- services with search functionalities like search engines to search any type of information.

▪ **Make resources accessible to the Web**

“DP9 is an open source gateway service that allows general search engines, like Google, to index OAI-compliant archives.”<sup>10</sup>. The OAI-PMH may be a standard solution in the future to provide access to the “hidden Web”, which is all the information contained in databases connected to the Internet. They are accessible through the Internet but not through classical search engines.

▪ **Cross-institution and cross-domain services**

Data are always heterogeneous when they are produced in two different institutions, even when the institutions are using similar metadata schemas. When implementing services to provide access to various catalogues, one always faces interoperability issues such as the definition of common access points, the way in which search results should be displayed, the way to provide coherent results to queries, the necessity to display content formatted similarly (image size for example)...

All types of resources may be accessed through a single interface, whether they cover a particular area (regional portal) or a specific subject. The traditional portals which offer references of Web resources may become “deep portals” which also allow to search and retrieve information in the given Websites<sup>11</sup>. This can lead to broaden the audience and define a structure to the digital representation of datasets held in various applications, possibly in remote institutions.

▪ **Data-reusability**

Cross-domain services may not take advantage of full bibliographic, museum or archival information. Only a part of those information may be of interest for resource retrieval for example. On the other hand, the service may define additional information needed to fully exploit the cultural resources, administrative metadata, preservation metadata, specific metadata for e-commerce, geo-referencing to connect the heritage information to a geographic information system....

A specific representation of the data for being used in the education sector or by an e-commerce application could be implemented to connect the local database to a specialised e-learning system or e-commerce platform.

<sup>9</sup> Digital Libraries Federation, “DLF evaluation of the Open Archives Initiative”, January 2003, <http://www.diglib.org/architectures/testbed.htm>

<sup>10</sup> <http://arc.cs.odu.edu:8080/dp9/about.jsp#link>

<sup>11</sup> see Johnston Pete, Dawson David “Collections et services : construire un environnement informationnel pour l’Europe” in Culture & Recherche, Paris, 2002, <http://www.culture.gouv.fr/culture/doc/index.html>

## Using OAI-PMH for preservation<sup>12</sup>

In order to ensure the long-term preservation of arXiv (which is the largest open access e-prints archive in Physics, set up by Paul Ginsparg in the early 90s), the French National Library manually collects the e-prints (flat files) from the CCSD<sup>13</sup>, which is responsible for the French mirror. The descriptive information (e.g. title, abstract, keywords, ... ) filled in by the authors, while submitting their articles, is also harvested in order to simplify the collects and control the updates. This information will then be reformatted and reused in long-term preservation metadata.

- **Standard for data interchange**

Generally, XML has become a standard way to exchange data from distinct databases, the OAI-PMH is becoming a standard way to exchange metadata, based on XML formatting and HTTP protocol. Given that the protocol may also collect resources, the OAI protocol can lead to exchange both metadata and resources.

- **Downstream services**

This principle can be used with part of a catalogue : libraries and archives have implemented a common name authorities services within the LEAF<sup>14</sup> project and the TEL library service<sup>15</sup> is using it through OAI harvesting.

The GSAFD<sup>16</sup> thesaurus is another example of an intermediary service which offers the thesaurus records in various formats and their mapping to another thesaurus (LCSH). "As a result, by storing the GSAFD Thesaurus as an OAI-PMH repository, its content becomes an integral part of the Web infrastructure where it can be seamlessly used by both human and machine using standard Web tools."<sup>17</sup>

Other types of application can be built from similar information collected in all digital libraries. This is the opportunity to build knowledge bases on several subjects. The Usage logs project launched by the OCLC and the Los Alamos Laboratory aims at collecting and taking advantage of experiences from various digital libraries and allow to exchange usage logs information to improve user experience.

All those applications of the OAI-PMH may however not be equally valuable. The OAI-PMH is not only a technology, it underlies an organisational model which may differently apply to the various types of information managed in memory organisations.

### **2.3 The OAI model to build heritage repositories and services**

The OAI architecture for heritage repositories and services indeed raises a number of organisational issues to clearly identify in order to define who shall make available what type of information for which benefit.

- **Shareable and disclosed models for data provision**

The organisation of service provision relies on the availability of digital cultural content. Memory organisations may provide their content according to a specific service or as a way to "share" their metadata, possibly for any (declared) service.

<sup>12</sup> on the archiving strategy, see presentation in Chanay Daniel, « Centre pour la Communication Scientifique Directe », dec. 2001 web.ccr.jussieu.fr/urfirst/presse/ccsd\_charnay.ppt

<sup>13</sup> Centre pour la Communication Scientifique Directe of the French National centre for Scientific Research (CNRS)

<sup>14</sup> Linking and Exploring Authority Files <http://www.crxnet.com/leaf/>

<sup>15</sup> <http://www.europeanlibrary.org/>

<sup>16</sup> Guidelines on Subject Access to Individual Works of Fiction, [http://www.ala.org/Content/ContentGroups/ALCTS1/Cataloging\\_and\\_Classification\\_Section/Committees3/Subject\\_Analysis/MARC\\_21\\_Authority\\_Records\\_for\\_GSAFD\\_Genre\\_Terms.htm](http://www.ala.org/Content/ContentGroups/ALCTS1/Cataloging_and_Classification_Section/Committees3/Subject_Analysis/MARC_21_Authority_Records_for_GSAFD_Genre_Terms.htm)

<sup>17</sup> Van de Sompel Herbert, Young Jeffrey A., Hickey Thomas B., "Using the OAI-PMH... differently", in D-Lib Magazine July/August 2003, vol 9, number 7/8, <http://www.dlib.org/dlib/july03/young07young.html>

A *shareable* model is an open model for repositories built for making metadata available, as widely as possible, for any harvester to harvest them, with a genuine intention of encouraging as many harvesters as possible to collect their data. This does not exclude to control access to the repository and to record harvesters which collect data but complies with a generic policy of knowledge sharing.

A *disclosed* model refers to repositories built for the specific use of a service, within a global community-based project.

The disclosed model is still the main reason why setting a repository and this is based on a clearly defined community. It constitutes a service-led approach to setting up repositories. It must be considered as the major way for memory organisations to feed an OAI-based service. Then, they can possibly define a generic policy of “making their metadata available”, but this does not motivate the initial decision. As an example, the French National Library intends to build a disclosed repository for making available resources of the Aquitaine region in the regional portal on Aquitaine heritage<sup>18</sup>. Then, as a second step, the repository of the National library will become shareable, including more data, formatted for a larger sample of library-specific services.

- **An asynchronous model**

The OAI architecture is asynchronous, the content of the original system (memory organisation's database for example, where original metadata are stored) may not be synchronised with the repository and the repository with the service. This can entail a delay between original system update and service's data update. A major challenge is to ensure that repository update is as much synchronised as possible with original system update and the service as much synchronised as possible with the repository.

This architecture may hardly be applied to data which are frequently modified. Typically, cultural heritage catalogues and finding aids do not have a high modification frequency and if data are not fully up-to-date, this does not make the full set of data irrelevant. However, records of cultural events for example, may not be adequately handled by the OAI-PMH.

- **An economic and organisational model for resource discovery**

The OAI architecture underlies an economic and organisational model where memory organisations may take part, as stated above, as data providers, possibly as service providers or aggregators as well.

In this model, the share of efforts, investments, and responsibilities for building a service is important and should be compared to the one involved for cross-searching for example.

The technical investment for the data provider may be lighter than with Z39.50. However, the main cost (documentary mapping and quality insurance) are identical for a cross-searching gateway and for disclosed OAI repositories. For shareable repositories, investment can be made once for various services, according to different needs.

Translations and mapping shall be conceived with the data providers but better performed on service-side since the repository may be re-used in another context and therefore the metadata content may be as close to the original as possible, given that re-processing and mapping tasks may always entail information loss. The data provider prepares data to be collected and the service provider adapts them to its specific use.

Aggregators can perform intermediary processing of data, possibly to build OAI-repositories from cross-searching or indexing of remote resources<sup>19</sup> and compensate the lack of local competencies and/or investment capacity to build proper OAI repositories. It is possible to use both cross-searching and central indexing to build a single service. An OAI architecture can then include an aggregator gathering data by performing cross-searching on various repositories and exposing those records in an OAI repository.

---

<sup>18</sup> [http://bnsa.aquitaine.fr/article.php3?id\\_article=12](http://bnsa.aquitaine.fr/article.php3?id_article=12)

<sup>19</sup> see the Resource Discovery Network

*Harvesting and/or Cross-searching<sup>20</sup>*

<u>Central indexing</u>	<u>Distributed search</u>
Uniform search options	Search options may differ per target
Predictable performance	Performance determined by slowest target
Easy integration of search results	More difficult to integrate of search results
Centrally controlled index	No control on distributed indexes
Less overall system load: one search is one request	System load is multiplied by number of requested systems

The OAI architecture can support an Application Service Provider - style model of common service provision for small institutions. It offers a framework to mutualise competencies and investments (an editorial department for example)<sup>21</sup>. Even the basic investment of an OAI repository can be avoided since an aggregator can gather the content by import / export procedures. That intermediary institution then processes the content so as to expose it for a specific service or as a shareable repository.

Economic models for service maintenance are identical to the ones implemented for any type of mutual service, whether through institutions' contributions or through users' contribution or through public funding or a mix of all this<sup>22</sup>.

This model must be ensured for the sustainability of both the service and the data providers and this issue is particularly important if data providers "delegate" part of their functions to a service provider : data preservation, unique access to their catalogues .... The metadata preservation and the possibility to bankrupt an OAI service may have important consequences, especially in the case of small institutions' service provision.

But in this case, an important feature of the model is that the data provider keeps responsibility of its data and exposes them under formal or informal contractual terms. For memory organisations, rights issues on resources shall be taken into account and they are very much concerned with controlling the data they expose and the use of their intellectual asset.

The economic and organisational model may not only involve traditional data publishers but rather evolve, together with digital libraries actors within the next years. By allowing new roles to appear in the field of digital libraries, the type of actors and therefore of economic model may be modified<sup>23</sup>.

<sup>20</sup> Van Veen Theo, "The European Library:opportunities for new services", Oaforum workshop "Open Access to Hidden Resources", 2002, [http://www.oaforum.org/otherfiles/lib\\_tel.ppt](http://www.oaforum.org/otherfiles/lib_tel.ppt)

<sup>21</sup> see the Territorial Service Centres of the OpenHeritage project (not OAI-based) in Lusso Salvatore, "OpenHeritage: Developing cultural tourism in lesser-known regions", in Cultivate Interactive issue 9, Feb. 2003, <http://www.cultivate-int.org/issue9/openheritage/>

<sup>22</sup> As an example, Picture Australia requires data providers' financial contributions

<sup>23</sup> see the DELOS/NSF working group on "Reference Models for Digital Libraries: Actors and Roles", <http://www.delos-nsf.actorswg.cdlib.org/work.html>

### 3 Building a repository

In an OAI framework, the memory organisations can make their content available to harvesters through OAI repositories. The implementation of a repository within an institution is a project which may have important impacts and a disclosed repository may lead to an open, shareable configuration. The cooperation policy of the institution is here materialised. It is important to define how a repository shall be built and the main issues to take into account, whether related to data, technical system or the "OAI repository project" as a whole.

#### 3.1 Setting records

The content of the repository is a set of records which may be organised by sets. The data provider defines the data exposed, the schema(s) according to which those data are made available, the datasets built to allow harvesters to collect data, finally the possibilities for harvesters to collect those data.

The original system, which is the database or system in which the metadata are usually managed in the institution will be used to build the repository. The repository content then relies on the quality and the organisation of the original system.

- **Many heritage standards are not proper XML schemas**

According to the wealth of the metadata available (schema used, possibility to map to other schemas), it will be possible to define which metadata formats can be provided to harvesters.

The content of the repository shall be exported or transformed to comply with a proper XML schema<sup>24</sup>, not simple database records. The conversion of descriptive metadata into proper XML formatted records is part of the standardisation process launched in most European countries<sup>25</sup>. Still, that crossroad is not always easy and it must be done with much attention since it is the basis of the quality of the metadata provided.

Moreover, this is a major constraint for cultural heritage resources. The standardisation process has started with SGML DTDs and all standards are not yet translated into XML Schemas. This is notably the case of EAD, the AMICO-DTD or BiblioML<sup>26</sup>. BiblioML is not a proper XML schema, but rather a DTD so that BiblioML files would not be OAI compliant, unless a proper schema would be developed out of the BiblioML DTD. The development of standard crosswalks is then very important.

The effort is on the way, the SPECTRUM-XML, MARC-XML, METS, MODS and of course DC schemas are proper standard schemas. But the effort needed is still large and standardisation processes can be long for all DTDs to be translated into proper XML schemas, accepted all over the communities. This can constitute an important barrier to the use of large cultural heritage specific schemas for metadata interchange and harvesting.

- **The OAI-DC schema**

The OAI protocol (OAI 2.0 specifications) requires the OAI Dublin Core schema as the minimal schema available in any OAI repository. If the institution is willing to expose larger metadata sets, it is still possible, but it may first allow OAI-DC records.

The OAI-DC schema<sup>27</sup> is unqualified Dublin Core. A specific OAI schema has been built in order to ensure the DC version even if the DC committee validates a new version of the DC schema.

But, the Dublin Core format is not always meaningful for the data concerned. Especially, in disclosed repositories, it may not be used. It is only required for full OAI compliance. The DC requirement is

---

<sup>24</sup> <http://www.w3.org/XML/Schema>

<sup>25</sup> see "Progress report of the National Representatives Group: coordination mechanisms for digitisation policies and programmes 2002.", European Commission : The Information Society Directorate-General, 2003, <http://www.minervaeurope.org/publications/globalreport.htm>

<sup>26</sup> DTD for UNIMARC

<sup>27</sup> [http://www.openarchives.org/OAI/2.0/oai\\_dc.xsd](http://www.openarchives.org/OAI/2.0/oai_dc.xsd)

even considered as an inhibitor for the development of Open Digital Libraries, especially when considering disclosed systems<sup>28</sup>.

However, as no element is mandatory in the DC schema, the single « title » element is enough to correctly match OAI-DC. Most existing implementations only use OAI-DC, at least as a first step to exchange data but it is important to take into account that simple DC may only be useful in the context of heterogeneous resources and the wealth of cultural heritage metadata is not fully exploited when using this format.

- **More complex schema**

Many implementers consider implementing or have implemented richer schemas. As memory organisations usually use richer formats, it is possible to expose richer formats. Formats such as MODS, METS, MARC XML, IMS have been implemented in OAI repositories.

But for accessing heterogeneous resources, whether from a single sector (museum community) or from various, the process consists of finding the highest common denominator, thus simplifying the metadata representations. This shall be done according to the need of the access interface (service) by identifying the necessary access points and descriptive information to display.

The trend is certainly to implement lighter formats than the ones which are created for the institutions' catalogues, possibly qualified DC. Several communities have created qualified DC for Web applications (DCMI Libraries and the DC Libraries Application Profile), it may be useful to explore that issue for other communities (such as the museum community for example<sup>29</sup>). Although controversial, those simplified versions of sector-specific schemas may be developed within the next years if they help preserve the wealth of cultural heritage metadata information in Web-based applications<sup>30</sup>.

In any case, service-specific schemas are commonly used to improve display or records management and it is predictable that most OAI applications will need to build specific schemas, based on standard ones for data interchange. The "24 Hour Museum" project distinguishes data to retrieve information from a schema to display results. The "Aquitaine Patrimoine" project has included specific information to build geographic representations of regional heritage and set specific categories for the portal that the repositories and/or aggregator shall include.

- **Managing relationships**

As the protocol provides means to transfer XML files, the relations between records and the hierarchical structure of records, as well as external references (such as terminologies) must be considered on service-side.

The relationship between a record on an article and a record on the periodical must be correctly interpreted on service-side. The relation must therefore be adequately encoded, for example through the DC:relation element and possibly its qualifiers HasPart, IsPartOf. This solution is proposed by Christopher J. Prom and Thomas G. Habing<sup>31</sup>, together with XPointers encoding model, to refer to nodes of original EAD documents from multiple OAI-DC records.

It is important to identify what the record represents and its relation to other records. If the schema is common to a group of disclosed repositories, this information (what the records actually represent) shall be part of the community agreement to build the service and the service shall have all the necessary material (terminologies, codes...) to correctly interpret the harvested records. However, for shareable repositories, this may be more complicated and a standard way to encode relationships shall be adopted.

---

<sup>28</sup> Suleman Hussein, Fow Edward A., "Beyond harvesting : digital library components as OAI extensions", 2002, [http://oai.dlib.vt.edu/odl/pubs/cstr\\_2002\\_odl\\_1.pdf](http://oai.dlib.vt.edu/odl/pubs/cstr_2002_odl_1.pdf)

<sup>29</sup> Perkins John, "Disclosing Digital Cultural Wealth: Museums and the Open Archives Initiative", in *Cultivate Interactive* issue 6, Feb 2002, <http://www.cultivate-int.org/issue6/cimi/>

<sup>30</sup> see Cole Timothy W., "Using OAI-PMH to aggregate metadata describing cultural heritage resources", presentation ALA/CLA annual meeting, 22 June 2003, Toronto, <<http://dli.grainger.uiuc.edu/Publications/TWCole/ALA2003OAI/>>

<sup>31</sup> Prom Christopher J., Habing Thomas G., "Using the Open Archives Initiative Protocols with EAD", Joint Conference on Digital Libraries 2002, <http://dli.grainger.uiuc.edu/publications/jcdl2002/p14-prom.pdf>

- **Mapping tasks**

The encoding task consists of setting equivalence between metadata elements of the schema and the original system's records, then to format the original system's records with an XML syntax. Traditional database-style records must be mapped to XML-formatted standard records according to the schema(s) selected to expose data, at least OAI-DC.

To create the equivalent XML records from original system's ones, it is important to ensure that the element represents the same concept and that the content of it (syntax and terminologies) is correctly interpreted. When mapping the data, it is better to create poor metadata records than wrong ones. However, several values may be hard coded, such as the publisher of a resource if it is common to a full dataset. Existing guidelines help defining correctly which concept must be represented in an element (guidelines to a specific schema, such as the CIMI guide to DC good practices), other help setting the equivalence between a standard metadata set and another one (crosswalks).

*CIMI Guidelines to DC*<sup>32</sup>

**'Reality Checking**

Creation of DC records is easy provided that three criteria are kept in mind:

- Is the record itself, and each element within that record, *useful* for resource discovery? If not, leave it out.
- Is the value of the element known with certainty? Is it readily available from existing databases or information sources? If not, leave it out.
- Have you selected values from enumerated lists recommended to assist in cross-domain searching? If not, please recognize that interoperability will be degraded and records will be harder to maintain."

In case of a service-specific schema, it is necessary to set specific guidelines. This may also be true for standard schemas, although they should be inspired by standard guidelines since for example a DC repository may be re-used by another service later.

*Mapping guidelines : some crosswalks*

- MARC21 to DC <http://www.loc.gov/marc/marc2dc.html>
- FINMARC to DC <http://www.lib.helsinki.fi/meta/dcficross.html>
- GILS to DC <http://db1-www.sub.uni-goettingen.de/servlets/metaformList1?Table=GILSCoreElementsDC&Head=GILS>
- EAD to multiple OAI-DC files <http://dli.grainger.uiuc.edu/publications/jcdl2002/p14-prom.pdf><sup>33</sup>

List of mapping guidelines : <http://www.ukoln.ac.uk/metadata/interoperability/>

Those guidelines specify which concept must fit in the element. They may also define the syntactic rules and possibly terminology used. The equivalence of the latter should preferably be done on service-side, to avoid information loss if necessary to map this data again for another service. For a shareable repository, the mapping task to a standard terminology is time-consuming and shall be avoided, unless it really provides an added value to the metadata exposed.

<sup>32</sup> CIMI – “Guide to Best practices Dublin Core DC 1.0 RFC 2413, version 1.1”, 21 April 2000, [http://www.cimi.org/public\\_docs/meta\\_bestprac\\_v1\\_1\\_210400.pdf](http://www.cimi.org/public_docs/meta_bestprac_v1_1_210400.pdf)

<sup>33</sup> in Prom Christopher J., Habing Thomas G., “Using the Open Archives Initiative Protocols with EAD”, Joint Conference on Digital Libraries 2002, <http://dli.grainger.uiuc.edu/publications/jcdl2002/p14-prom.pdf>

However, if there is no terminology re-construction on service-side according to a thesaurus, retrieval will still be facilitated if writing all thesaurus hierarchy in the plane element, as it is advised in the ARC service "SUBJECT - This is divided into subject and sub subject fields eg. physics/high energy physics."<sup>34</sup>

An important issue was to create crosswalks from EAD records. Indeed, EAD records describe various levels in a single record (collections, sub-collections and items, all together in a single record). When they are to be retrieved together with other types of schemas, representing either a collection or an item, they may need anchors or be divided. A trial has been led within the Illinois-Champaign University to assess the generation of various records from a single EAD record<sup>35</sup>. Then, the repository does not contain EAD records, but rather DC or any other format, unless the EAD records are gathered and clearly processed as finding-aids.

Generally, the mapping task and re-use of existing metadata will lead to assess them and the use of a proper schema will force to a formal quality control on metadata. This step may lead to a better appropriation of metadata held by the institution and improve their use.

#### ▪ Identifiers

The items are "stored" in the repository. It is important to clearly define what they represent (pointer to resources) and how they can be expressed (records generated possibly according to various schemas). The identification of each entity is a key issue in the OAI architecture for three different concepts :

- Identifying a record : the oai-identifier is a unique identifier for each harvested record (XML expression of an item). They must be a valid URI (Uniform Resource Identifier), unique to the repository. It may optionally refer to the OAI identifier scheme to be unique in oai namespace<sup>36</sup>. The record identifiers are then resolvable via a central OAI resolution service available at <http://www.openarchives.org><sup>37</sup>.
- Identifying an item : it represents the resource and may be represented by several records (MARCXML and DC for example). "A *unique identifier* unambiguously identifies an item within a repository; the unique identifier is used in OAI-PMH requests for extracting metadata from the item. Items may contain metadata in multiple formats . The unique identifier maps to the item, and all possible records available from a single item share the same unique identifier. The format of the unique identifier must correspond to that of the URI syntax. Individual communities may develop community-specific URI schemes for coordinated use across repositories."<sup>38</sup>
- Identifying (and possibly locating) the resource. The URL may be used but the DC identifier element does not force to provide a URI. This is an especially important issue since the OAI protocol is asynchronous so that any modification of the repository will not immediately be viewed in the service's structure. There is a clear risk for the service to send the user to several dead links.

The resource identifier is ideally unambiguous, unique and persistent and repositories may consider using persistent identifiers.

In several cases (records describing a resource which is not stored in the institution, such as a Website or a monument), the identifiers issue may lead to difficulties in the de-dupping process (also the case for non unique resources).

PURLs have also been implemented for OAI records to allow human access to repositories records through "cool URLs", rather than the classic OAI-based URIs.

---

<sup>34</sup> Help on advanced search [http://arc.cs.odu.edu:8080/oai/service\\_help.html](http://arc.cs.odu.edu:8080/oai/service_help.html)

<sup>35</sup> see BNSA and Illinois University project

<sup>36</sup> see definition of OAI identifiers at <http://www.openarchives.org/OAI/oai-identifier.xsd>

<sup>37</sup> see definition of OAI identifiers at <http://www.openarchives.org/OAI/oai-identifier.xsd>

<sup>38</sup> <http://www.openarchives.org/OAI/openarchivesprotocol.html>

Hussein Suleiman's tutorial JCDL conference 2002 : among the main problems encountered<sup>39</sup> :

"No unique identifiers

- *Create an independent identifier mapping*
- *Use row numbers for a database*
- *Use filenames for data in files*
- *Use a hash from other fields E.g. author+year+first word in title"*

- **Defining sets for selective harvesting**

The data contained in a repository can also be prepared for the harvest by performing a selection of the data which the service provider may need. If a harvester used all the data and another one only a very small proportion of them, then it shall be possible for the second one to only select the records it needs, which is the role of OAI sets.

A harvester can collect metadata from a repository but never query a repository, which means that there is no possibility for the harvester to select records if the repository is not built for that purpose. The harvester will always collect only changes since last harvest. However, it may be only concerned with part of the records on a specific subject. As an example, this may allow to select only cartographic treasures of the Library of Congress without harvesting the whole repository, the repository can build an OAI set containing only the maps and cartographic treasures. Then, any harvester can connect and collect metadata from one or more sets (selective harvesting).

It is even possible to make hierarchical sets (collections and sub-collections) and an item may belong to various sets but this practice does not seem very widespread.

Consequently, if some harvester is only interested in part of the collection, then it is recommended to build sets. The set must therefore be proposed with a name and an identifier so that a harvester can query the repository by specifying the set(s) it needs : "setSpec: musdibib and setName: LC Dance"<sup>40</sup>.

For a disclosed repository, the sets shall be defined according to harvesters' needs, if the repository has various harvesters. But, in existing systems, mostly shareable ones, sets are often defined by subject classification and sometimes resource types (Periodicals / books...). Indeed, for a shareable repository, the sets may be usefully built by "collection", defined according to subjects or types, possibly a refined definition of collections. It is not made according to one harvester's needs, but rather as a way to define datasets and perform collection level descriptions.

Even if harvesting the whole repository, the definition of datasets may help building specific search interfaces providing the possibility to limit the search to several datasets (ARC). Therefore, it is necessary to publish in human readable format the descriptions of OAI sets defined in the repository, such as done for the "American Memory" repository.

- **Data update**

As stated above, the asynchronic nature of the OAI model creates a responsibility for the data provider to maintain the repository, with up-to-date data. If data are not updated in the repository, the delay for synchronization with the service is even higher, this can lead to unreliability of the service and possibly dead links.

Depending on the type of repository architecture which is implemented, the data update may be an automatic, regular or irregular process. The repository update consists in modifying metadata records, deleting or creating new metadata records and modifying record timestamps adequately (not replace unchanged records then generate artificial modification dates).

If the repository is "integrated" into the original system (directly a view of the database), then the repository is automatically updated. For other configurations, it is necessary to determine when it is suitable to update data. This depends on both the frequency of data modification in the original system and the way in which the service(s) uses the data. For example, most bibliographic records are unlikely to be modified, possibly the IPR element if they are not public domain and this will not create any major inconvenience for a resource discovery service.

<sup>39</sup> [http://www.dlib.vt.edu/projects/OAI/reports/jcdl\\_2002\\_tutorial\\_oai\\_slides.pdf](http://www.dlib.vt.edu/projects/OAI/reports/jcdl_2002_tutorial_oai_slides.pdf)

<sup>40</sup> <http://memory.loc.gov/ammem/oamh/>

New records are created on a regular basis but they are not always integrated immediately in digital libraries. It is therefore possible to update the repository at a similar frequency, when a new collection is fully available (irregularly), or once a month (advised in the PictureAustralia service, by the New Opportunities Fund technical board, planned by the French National Library). For any other type of information, it is necessary to analyse the creation process.

When the repository is updated, solutions are being considered to “push” that information for harvesting data and synchronize data as much as possible<sup>41</sup>.

### 3.2 Technical issues

To properly configure the repository, design the system and deliver data, the system architecture shall take into account the positioning of the repository towards the original system, the way it welcomes harvesters and formats the data.

- **Defining the system and architecture**

The repository can be built in different ways according to the original system. From a filesystem, the metadata are extracted in a different way than from an SQL-based system. Many document management and bibliographic softwares use filesystem-based management tools or a mix between filesystems and relational databases. To make Websites harvestable, it is possible to create metadata per page or per parts of Website and include them in a repository, either manually or automatically (metadata extracted from the META tags of HTML pages).

From those very different structures, specific metadata extraction must be implemented for each application, possibly by using parametrable OAI metadata generators for different softwares.

#### *Various scenarios proposed to data providers<sup>42</sup>*

- If using an application with limited export functionalities, it is not possible to directly extract XML-based schema from mapping existing metadata. Then an intermediary system will be built, gathering exported files and converting them into valid records for the OAI repository.
- If using a database such as MySQL, a repository can be built, including the OAI records directly extracted from the database or the OAI requests can be directly translated into SQL query if all necessary information (including header) are recorded in the database.
- If it is a static Website, an intermediary application will parse the META tags and translate them into the valid schema.

The AMOL project presentation defines the various configurations between the original system and the repository in the following way :

“Repository models :

- independent systems
  - separate from other services
  - periodic updates
- semi-integrated systems
  - read-only connections to databases
  - data providers and data systems not necessarily at same site
- fully integrated systems
  - OAI protocol tightly integrated into database”<sup>43</sup>

<sup>41</sup> see Liu Xiaoming, Maly Kurt, Zubair Mohammad, “Arc - An OAI Service Provider for Digital Library Federation”, in D-Lib Magazine, Volume 7 Number 4, April 2001, <http://www.dlib.org/dlib/april01/liu/04liu.html>

<sup>42</sup> Sévigny Martin, Bourgoïn Christine, “OAI (Open Archives Initiative)”, June 2003 <http://www.ajism.com/projets/pp/technique/oai.html>

<sup>43</sup> Dewhurst Basil, “Enabling Interoperability : Australian Museums Online (AMOL) & the Open Archives Initiative Protocol for Metadata Harvesting “, AMOL/CIMI Institute Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) Workshop June 2002, [http://amol.org.au/oai/files/AMOL\\_CIMI\\_OAI-PMH\\_Workshop\\_BD\\_Enabling\\_Interoperability\\_20020618.pdf](http://amol.org.au/oai/files/AMOL_CIMI_OAI-PMH_Workshop_BD_Enabling_Interoperability_20020618.pdf)

The difficulty of integrated systems is that harvesting requires the original system's capacity and this may create server's overloads for large datasets. But this is an issue only for the first harvest of catalogues, since modifications are not numerous, then a solution could be to manually transfer first harvest. The Celestial system is also a solution to avoid load on repositories, through caching. It harvests various repositories and exposes the results to any harvester, thus avoiding problems due to unavailability of data providers and load of first harvest which is then transferred to the caching system<sup>44</sup>.

The repositories can be implemented as standard functionalities of catalogues and finding aids. The development of those features comes from a very important partnership with major industrial partners of memory organisations for the management of their Web applications<sup>45</sup>. Several products such as Adlib server and ENCompass have already included this option<sup>46</sup>.

The technical implementation of an OAI repository can be tested through a validator, by declaring it to the Openarchives.org. This launches a test on OAI compliance of the repository.

- **Security and access rights**

An OAI request is a simple HTTP request on a repository server. However, in order to design the position of the repository in the global information system of the memory organisation, it is important to take into account the fact that a harvester must enter into the local network, given the existing system architecture (firewalls...). The service provider shall be informed of the necessary parameters to enter and possibly be declared to local network.

A general policy on access rights to metadata shall be implemented, to define whether any harvester can access the repository, whether the repository must be protected by login / password, possibly based on HTTP protection such as in the RDN network or by harvesters' IP control, or whether harvesters visits and the content of their harvest is recorded.

HTTP-based authentication requires an adaptation of harvesters but may be interesting if not willing to expose metadata to any service provider and if not willing metadata to be harvested and represented in any context.

The memory organisation may consider that type of protection as a major feature to keep control on its own metadata. It may also, in disclosed repositories, be willing to define OAI sets according to each harvester, although the protocol does not implement access rights functionalities in datasets and repositories.

Those protections need therefore to be managed for each data provider on service-side. But they bring a better confidence for data providers. Protections can be implemented whether in disclosed or in shareable environments, since the harvester always declare what it is doing and the data provider may always want to keep track of what the harvester is doing.

- **Data delivery**

The data delivery is the main function of OAI repositories. The conditions of that delivery (rights issues and control of data source) and the technical validity of records are major quality indicators for the repository.

The data provider shall clearly define what is possible to do with the data and be clearly identified by the end-user as the data source.

---

<sup>44</sup> Brody Tim, Kampa Simon, Harnad Stevan, Carr Les, Hitchcock Steve, "Digitometric services for Open Archives Environments", in Koch Taugott, Torvik Solvberg Ingeborg, "Research and Advanced Technology for digital libraries" 7<sup>th</sup> European conference, ECDL 2003, Trondheim Norway, August 17-22, 2003, Berlin 2003

<sup>45</sup> see for example current development of OAI-PMH functionality for ENCompass software <http://encompass.endinfosys.com/faq.htm> or Adlib Internet Server [http://www.uk.adlibsoft.com/relnotes/ReleaseNotesWWWOPAC4.7\\_EN.doc](http://www.uk.adlibsoft.com/relnotes/ReleaseNotesWWWOPAC4.7_EN.doc)

<sup>46</sup> see ENCompass, ContentDM, Michigan's DLXS, D-Space in Cole Timothy W., 'Using OAI-PMH to aggregate metadata describing cultural heritage resources', presentation ALA/CLA annual meeting, 22 June 2003, Toronto, <<http://dli.grainger.uiuc.edu/Publications/TWCole/ALA2003OAI/>>

### Data provenance and rights issues

An OAI record is composed of 3 sections, the *header* with the OAI identifier, the *timestamp...*, the harvested *metadata* and the *about* section<sup>47</sup>. The *about* section is used to specify complementary information on the metadata record according to specific schemas :

- the provenance : poorly used, the provenance is very important to specify, especially if the data are collected by an aggregator and “re-harvested”. A specific schema has been created to describe the original repository <http://www.openarchives.org/OAI/2.0/provenance.xsd>.
- rights in case of rights attached to the metadata (eg. Creator) or the limitation to the use of the metadata. There is no existing standard to define the allowed use and the content of this field, though it would be very useful to define one with its community (where existing)

On rights issues, the report “Open Archives and Intellectual Property : incompatible world views”<sup>48</sup> mentions the possibility of copyright on metadata sets, certainly moral rights and rights on databases (a repository may be considered as a database). There is therefore a clear necessity to state rights and usage limitations for individual records as well as for the repository, both when delivering data and when setting the agreement to harvest data.

### Rights issues on metadata<sup>49</sup>

The repository (Data Provider) creates a metadata record	Are there any rights in an individual metadata record?
	If so, who owns them?
	Do DPs wish to assert any rights over either individual metadata records, or data collections?
	If so, what do they want to protect, and how might this be done?
The metadata is disclosed to Service Providers that harvest it.	Do DPs disclose any rights information relating to the documents themselves?
	How do SPs ascertain the rights status of the metadata they're harvesting?
	Do SPs enhance harvested metadata records, creating new IP?
	Do SPs want to protect their enhanced records? If so, how?
	How do SPs make use of any rights information relating to the documents themselves?

*To define the regime of metadata use, the “Creative Commons licensing system” is an option<sup>50</sup> :*

“Rights. All metadata records in the IESR will be freely available and licensed under a Creative Commons Licence: *Attribution Required; Non-Commercial; Share-Alike*. When you submit data to the IESR you are agreeing to this licence on your metadata records.”

### XML formatting

The other important challenge of repositories is to deliver well-formed XML responses, which is a major and very common difficulty for service implementers : “in about 10% of the repositories currently being harvesting, we encountered XML validation errors”<sup>51</sup>. This can impede harvesters to handle the

<sup>47</sup> see <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>

<sup>48</sup> Bide Mark, “Open Archives and Intellectual Property : incompatible world views”, community report for the Open Archives Forum, nov 2002, [http://www.oaforum.org/otherfiles/oaf\\_d42\\_cser1\\_bide.pdf](http://www.oaforum.org/otherfiles/oaf_d42_cser1_bide.pdf)

<sup>49</sup> ROMEO project <http://www.lboro.ac.uk/departments/ls/disresearch/romeo/index.html>

<sup>50</sup> in Apps Ann, “JISC Information Environment Service Registry (IESR) OAI-PMH Service Input Templates”, 2003, <http://www.mimas.ac.uk/iesr/metadata/templates/svcoai-template.html>

<sup>51</sup> Halbert Martin, Kaczmarek Joanne, Hagedorn Kat, “Findings from the Mellon Metadata Harvesting initiative”, in “Research and Advanced Technology for digital libraries” 7<sup>th</sup> European conference, ECDL 2003, Trondheim Norway, August 17-22, 2003, Berlin 2003. See also Kim Hyunki, Choo Chee-Yoong, Chen Su-Shing, “An integrated digital library server with OAI and self-

response. XML encoding requires a specific transcoding task, notably in a multilingual context, since most document management systems do not use Unicode to record metadata.

The inclusion of intermediary aggregators in the framework<sup>52</sup> can help improving the quality of the XML syntax. The implementation of an XML parser at repository level for validity checking and the declaration to Openarchives.org with regular tests on the repository are recommended. But Openarchives.org control may not detect problems in implementation of syntactic transformation for textual information.

*The JCDL 2002 Conference, Hussein Suleiman<sup>53</sup>*

“Data Cleaning

- *Escape special XML characters*
- *Convert to UTF-8 version of Unicode*
- *Convert entity references*
- *Remove extraneous whitespace*
- *Convert CR/LF for paragraphs*
- *URLs*
- */?#=&.;+ must be encoded as escape sequences”*

### *Accessibility*

The protocol should be invisible to the user and repositories are designed to be accessed by machines (harvesters). However, in certain cases, it is useful to read the content of a repository and the OCLC has built a human interface for OAI repositories<sup>54</sup>. It is a convenient way to access all downstream services such as thesaurus which are then logically accessed in a different way than they are through the services.

### **3.3 Communication and management**

The implementation of OAI compliance for cultural heritage resources is a project which is clearly integrated in a political strategy of cooperation and openness of the institution. The project handles such issues as communication, management, consideration for available competencies, budget, and evaluation.

- **Budget issues**

The budget needed may be very different according to the type of service, whether the repository is disclosed, the original database system, the network infrastructure, the available competencies on the OAI-PMH. However, here are several categories of costs which can be considered when setting a repository :

- training / getting informed;
- selection of the tool and installation;
- creation of metadata (for Websites for example or to improve existing metadata);
- mapping of metadata to the selected schema(s);
- creation of the schema(s) (if not existing);
- specifications for the generation of XML formatted metadata from the initial database;
- implementation of the repository, configuration, tests of correct processing of queries and server load testing;
- communication and information on the project, including registering with service providers and Websites which list potential data providers ;

---

organizing capabilities” in same conference and Liu Xiaoming, Maly Kurt, Zubair Mohammad, “Arc - An OAI Service Provider for Digital Library Federation”, in D-Lib Magazine, Volume 7 Number 4, April 2001, <http://www.dlib.org/dlib/april01/liu/04liu.html>

<sup>52</sup> see Liu Xiaoming, Maly Kurt, Zubair Mohammad, “Arc - An OAI Service Provider for Digital Library Federation”, in D-Lib Magazine, Volume 7 Number 4, April 2001, <http://www.dlib.org/dlib/april01/liu/04liu.html>

<sup>53</sup> [http://www.dlib.vt.edu/projects/OAI/reports/jcdl\\_2002\\_tutorial\\_oai\\_slides.pdf](http://www.dlib.vt.edu/projects/OAI/reports/jcdl_2002_tutorial_oai_slides.pdf)

<sup>54</sup> Van de Sompel Herbert, Young Jeffrey A., Hickey Thomas B., “Using the OAI-PMH... differently”, in D-Lib Magazine July/August 2003, vol 9, number 7/8, <http://www.dlib.org/dlib/july03/young/07young.html>

- project management.

The documentary work to normalise terminologies or metadata sets shall be performed on service's side. However, the contribution of the data provider to define the rules to present its content in a relevant context may be resource consuming.

As an example, the French National Library has counted 66 day/person to set up an OAI-DC repository for the Aquitaine Patrimoine project, which includes a small part of its data. The project involved the creation of new metadata (20 days/person) but existing crosswalks, no training, a large communication task and a complex network infrastructure.

- **Available competences**

The experiments of the Mellon Foundation lead to the conclusion that the low take-up of OAI-PMH technology in the heritage sector is highly due to the lack of financial and personal resources. However, when offering to host a repository for various institutions, then the participation was facilitated<sup>55</sup>. The Aquitaine Patrimoine project has adopted a similar strategy with part of the data providers. The repositories are hosted and maintained by the service provider and memory organisations have export their data to the repository on a volunteer basis. However, each processing is made for individual data provider and distinct repositories are built so that, when considering relevant to host its repository, the transfer of the repository will be easy.

This strategy of aggregators partially compensates the lack of competences and investment on the technical implementation. Still all the work on metadata (conversions, mapping design ....) are done in the memory organisation. The memory organisations hold scarce competences in information technologies and the possibility that the repository is outsourced appears a good solution to small institutions.

This mainly fills a lack of qualifications which shall be well identified. The implementation of a repository on a heritage collection may require competences in the following areas :

- information technologies :
  - original system architecture (to define whether the repository shall be integrated, to ensure a relevant update strategy, to avoid system overload ...);
  - scripting to extract metadata from original system and encode according to XML schemas;
  - XML if it appears necessary to write and / or modify metadata schemas;
  - networks to possibly ensure encryption, to control harvesters entrance in the network ...
- metadata :
  - knowledge of the collection(s) exposed, the cataloguing rules practices and the coherence (topics, document types ...) of that collection;
  - metadata quality and mapping procedures to ensure equivalence with XML schemas, standard terminologies, identify the possible problems ...
- Web strategy and access :
  - strategy on rights issues for content dissemination (not necessary by a proper lawyer, but understanding of legal issues for digital content);
  - user knowledge to define a user access strategy and the conditions for the service provider to disseminate and exploit metadata and content.

On other types of repositories, the "user" strategy is slightly different. If the "metadata" are usage logs for example, the knowledge of the collection is held by the server administrator and the person(s) who analyse user access. Metadata competences are content independent.

---

<sup>55</sup> see Halbert Martin, Kaczmarek Joanne, Hagedorn Kat, "Findings from the Mellon Metadata Harvesting initiative", in Koch Taugott, Torvik Solvberg Ingeborg, "Research and Advanced Technology for digital libraries" 7<sup>th</sup> European conference, ECDL 2003, Trondheim Norway, August 17-22, 2003, Berlin 2003

- **Register the repository**

Once a repository is OAI compliant, it can be registered. If it is a shareable repository, then it may be interesting to register it for being harvested by generalist services and main harvesters in the domain (eg. libraries). Here is a list of service providers <http://www.openarchives.org/service/listproviders.html>, and on the OAForum Website [http://www.oaforum.org/oaf\\_db/list\\_db/list\\_services.php](http://www.oaforum.org/oaf_db/list_db/list_services.php).

The OAI registry of the openarchives.org Website also includes a conformance control procedure performed by the registration service<sup>56</sup>. Even if the repository is not accessible, the registration process is a quality label and it may allow to establish further cooperation in the future. The OAForum [http://www.oaforum.org/oaf\\_db/register/reg\\_intro.php](http://www.oaforum.org/oaf_db/register/reg_intro.php) also allows registration of projects so that partnerships and exchange of good practices can be facilitated. The repository explorer <http://oai.dlib.vt.edu/cgi-bin/Explorer/oai2.0/testoai> publishes a list of repositories and the DP9 service<sup>57</sup> allows to make the resources available for Web crawlers.

*A repository has a Vcard which allows to be harvested :*

Element name <sup>58</sup>	element value
Base URL	<a href="http://www.bsz-bw.de/cgi-bin/oai20_send.pl">http://www.bsz-bw.de/cgi-bin/oai20_send.pl</a>
Repository Name	Bibliotheksservice-Zentrum Baden-Württemberg, Germany, Virtueller Medienserver
Protocol Version	2.0
Email	<a href="mailto:oai@bsz-bw.de">mailto:oai@bsz-bw.de</a>
Registration Date	Mon Apr 7 09:13:34 2003
Date Last Validated	Mon Apr 7 15:03:55 2003
OAI Repository ID	bsz-bw.de

- **Publish information on the repository**

In any case, it is important to communicate on OAI compliance and publish information on the repository in human readable format, including sets in order to ensure possible partnerships in the future.

For example, the Library of Congress provides a short description of collections to harvest. The information published about a shareable repository could include preferences and conditions to harvest metadata for example. To see the description of an OAI-PMH data provider, see below "OAI-PMH Service Description Tabular Form"<sup>59</sup>.

This is also a way to communicate on the project and the organisation's strategy. The implementation of an OAI repository in a memory organisation can be considered as :

- a generic approach to international standards;
- an effort to get value out of metadata managed in the institution;
- a commitment to offering new services to user or enlarge its audience<sup>60</sup>;
- a wish to set cooperation with other institutions<sup>61</sup>;
- a strategy of openness of the institution to other institutions and/or sectors.

A communication strategy shall certainly be implemented, not on the technology but rather on the organisational impact it bears. However, there is an ambiguity for the data provider to mention its

<sup>56</sup> see "The Open Archives Initiative – Registering as a data provider", <http://www.openarchives.org/>

<sup>57</sup> <http://arc.cs.odu.edu:8080/dp9/about.jsp#download>

<sup>58</sup> record extracted from the Open Archives Website, <http://www.openarchives.org/Register/BrowseSites.pl>

<sup>59</sup> In this case "service" is used as a Web-based service rather than the OAI terminology of a service provider.

<sup>60</sup> See James Cook University presenting its collaboration with PictureAustralia as a demonstration of efforts to improve access to its resources <http://www.jcu.edu.au/asd/edge/story08.shtml>

<sup>61</sup> See Australian War Memorial Annual report 2000-2001, [http://www.awm.gov.au/corporate/annual\\_report/ann\\_rep00-01.pdf](http://www.awm.gov.au/corporate/annual_report/ann_rep00-01.pdf)

participation to another access service, since it may not be interested in sending its user to another Website to access its own data. This communication appears rather important towards professional counterparts but difficult towards end-users. The reference to the service's Website may then be presented as a possibility to access further resources held by other institutions on a similar subject coverage. For the State Library of Tasmania, "PictureAustralia is also a very useful source of Tasmanian images held in other national and state institutions around Australia."<sup>62</sup>

▪ **User and evaluation**

Finally, as for any project, it is important to define the criteria to consider the OAI repository project is a success and what is a "good" practice of an OAI repository building in the cultural heritage sector according to its way of working and its users.

For a repository, the user is a service harvester and the end-user is shared with the service. The project evaluation must then be considered according to the quality of the data provided and of the data delivery process, as well as the relevance of contextualisation on service side (how data are represented). The metadata quality directly influences the service's quality and the necessary efforts to reprocess content on service-side. Finally, the user experience of both the service and the way back from the service if users reach the data provider's Website shall be considered as end-user requirements.

Several issues to consider for evaluating an OAI repository include :

- information is up-to-date in the repository;
- repository is always available for harvesters or often down....;
- there is an impact on overall consultation of the institution's resources ;
- new users are coming to the data provider's Website or existing users find better service thanks to the possibility of new contexts for information;
- repository is harvested by interesting services;
- appropriation of OAI technology was quick and implementation easy (as the OAI is supposed to be a simple protocol to implement);
- the project has contributed to support openness policy and collaborative actions within a community (and possibly raised additional funding<sup>63</sup>).

On most of those criteria, the repository is highly dependent on the service structure and management. In disclosed frameworks, the data provider is indeed very much involved in the service's design and most of the issues are faced by the whole community which participates to the overall project of setting a new service.

---

<sup>62</sup> <http://www.statelibrary.tas.gov.au/heritage/services.htm>

<sup>63</sup> see Digital Libraries Federation, "DLF evaluation of the Open Archives Initiative", January 2003, <http://www.diglib.org/architectures/testbed.htm>

## 4 Building a service

A service based on cultural heritage data shall rely on a series of data providers among which memory organisations and/or aggregators for which it is particularly important to consider disclosed models. The major difficulty is to ensure interoperability between systems, both semantic and organisational.

### 4.1 A service for cultural heritage resources

The approach of aggregated resources and cross collection services is a major issue for adding value to user access to any type of resource. Still, various types of services can be set up which lead to face different issues.

#### ▪ Services for accessing aggregated content

Different types of services, such a deep portal, a digital library with distributed architecture, an exposition of metadata to search engines, an intermediary service for authority names or thesaurus can be implemented. They may be based on cultural heritage aggregators and they may work with harvesting technologies as well as cross-searching (ZING or Z39.50).

The use of an aggregation model is valuable if the service :

- Brings a new interpretation to the material;
- Reaches new audiences for the content considered;
- Allows to share investments and competencies;
- Provides new access services (functionalities) which are not implemented in the data provider's system.

A major challenge is therefore the representation of content in a new context with new meanings. This access mode shall entail as less information loss as possible and it must take into account metadata author's willing.

The service, by providing access to aggregated content shall set a clear definition of the audience, different from the individual data provider's ones but which must be coherent with the data gathered.

#### ▪ Setting a community

A service can be built with shareable or disclosed repositories or both, though the cultural heritage sector will first need to build coherent access services involving specific data providers, encouraging them to become OAI-compliant.

#### *A charter*

The disclosed model shall rely on a formal agreement of all partners to contribute to the service. Indeed, the service shall be based on common organisational rules and the existence of a community with agreed working processes.

It can be based on sectoral communities (libraries may create content in a similar way, much more similar than with museums although the difference is always large)... Cross-domain communities, including various types of stakeholders shall face several difficulties such as different ways of considering information, creating metadata, ....

An agreement must be found on the way resources are to be represented in the service, including, for example the possibility to display the resources inside the service's template and frames. The Australian War Memorial images are represented in the AMOL Open Collections' frames<sup>64</sup>. The form of the data provider identification on service side is part of the resource representation.

A service shall be based on a community of repositories and it must define the rules with that community :

<sup>64</sup> [http://amol.org.au/collection/collections\\_index.asp](http://amol.org.au/collection/collections_index.asp)

- *One or more metadata schema;*
- *IPR and usage issues for metadata and resources (including allowance to transform metadata on service-side);*
- *Use and possible crosswalks of terminology;*
- *What to do in case of dupping (which record shall be presented in the service...);*
- *Conditions under which the service provider directs the user to the data provider Website;*
- *The way in which the data provider of a record is mentioned (branding, font...);*
- *Commitment to publish basic information on data providers;*
- *Harvesting frequency and repository update frequency;*
- *Guarantee of repository availability;*
- *Provision of thumbnails for example;*
- *Format of resources available;*
- *Respect of community guidelines;*
- *The conditions under which a new member can be included;*
- *Quality criteria in the way the service represents data (Website?) and possibly the data provider's Website if end-user needs to access it<sup>65</sup>.*

Community documents can include :

- *Guidelines to map metadata and on metadata content;*
- *A procedure including level of quality control on metadata : quality commitment on metadata provision*
- *A "charter" or agreement for the use of metadata and all the relations within the partners.*

All those issues shall be dealt with in a formal agreement, whether a contract or a simple charter, such as the one used in the Physnet service.

*Extract of the Physnet charter<sup>66</sup>*

**"6.2 Distributed**

All physics information of PhysNet is kept, stored and maintained by its creators at their local institution's server. PhysNet members retain their ownership and copyright of their data. They agree to make the respective metadata available to the entire PhysNet services.

PhysNet gathers and processes available local information of Physics institutes to make them globally accessible. Searching for the information is provided by the PhysNet providers, which are distributed worldwide according to their expertise and interest. PhysNet is a distributed system with no bias to any center or nation."

*Advantages for the data provider to join the community*

In the case of new services to access content, it is important to well define the advantages for a data provider to collaborate :

- New audience with new user experience;
- Importance for data providers to develop their own future services on their own collections;
- Better standardisation through encouragement to the use of standard terminologies, and resources' formats;
- Benefit for a data providers' community to exchange good practices;
- Openness to other national and international projects' cooperation;
- Opportunity to work in a cross-sectoral environment.

<sup>65</sup> see Quality grid on cultural Website below

<sup>66</sup> <http://www.eps.org/PhysNet/charter.html>

If the service provider allows end-users to reach the data provider's information, it shall provide a feedback on service's usage statistics and information on users' feedback.

*The National Science Digital Library in the US has set up 5 committees for community services, content, technology, educational impact and sustainability*

"The goals of the Sustainability Standing Committee are to:

- Build a collective identity and common vision for the NSDL.
- Identify and promote the value of the NSDL to various communities.
- Establish an economic model for sustainability based on returns from the value provided.
- Develop long-term governance leading to a not for profit Foundation, Cooperative, or Trade Association.<sup>67</sup>

#### *Maintaining the community*

It is necessary to set mechanisms for keeping the community active and ensure its commitment to the maintenance of the service. That commitment may be important to ensure user follow-up for example. It is possible to maintain the community through mailing lists, real meetings, ...

*A community shall first be real – The Picture Australia service maintains an email list with data providers and holds an annual meeting :*

"The annual participants' meeting

The participants' meeting has been held annually since 1999. The meeting is convened during the first quarter of each year. All participants are invited to submit issues to its agenda, preferably before the meeting to allow everyone to consider them in detail. In accordance with its obligations under the Service Level Agreement, the PictureAustralia host provides a summary report of the performance of the service for the preceding year."<sup>68</sup>

#### **4.2 Cross-collection content is always heterogeneous**

The way cataloguing has been performed and the type of resources they contain make collections always heterogeneous from one source to another. The main issues raised by cross-collection services are to foresee the responses of a cross-collection system and the management issues for managing heterogeneous resources and/or metadata.

- **Data analysis**

The data provided from different sources shall be analysed :

- *the original databases / applications used to manage metadata locally, including metadata structure;*
- *if Websites, whether there are metadata available, and how it is possible to harvest them;*
- *metadata formats and syntax : DTD, schemas and cataloguing rules, if XML formatting already implemented;*
- *formats for exporting and publishing metadata;*
- *the OAI sets to harvest;*

<sup>67</sup> <http://sustain.comm.nsdlib.org/>

<sup>68</sup> <http://www.pictureaustralia.org/guide.html>

- *the volume of records it can represent for each harvest;*
- *the updating frequency in the repositories;*
- *the audience of the data providers;*
- *preferences of the data providers for harvesting (eg. a specific hour);*
- *whether resources are online or whether the data provider only can provide metadata;*
- *whether data provider has implemented password protection.*

The repositories content shall also be analysed :

- *whether it is referred to terminologies;*
- *whether there a syntactic control on values (eg. date formatting);*
- *which type of content is recorded, only bibliographic records, if images, multimedia, Websites, .... Whether resources are included in records;*
- *the description level (granularity issue);*
- *the language of metadata;*
- *the accrual status of the repository;*
- *whether modification of records is frequent;*
- *whether there are relations between records;*
- *whether elements refer to external system for example for geo-referencing or thesaurus;*
- *contextual information and interpretive value of the original system;*
- *whether to collect the resources as well.*

The policy defining whether the data provider want to :

- *give access to resources;*
- *collect the resources or low quality versions;*
- *represent resources in another context;*
- *specific information to be published on the service's Website about the data provider;*
- *whether the use of metadata is restricted;*
- *whether there are restrictions in the data providers' conditions according to the community partners.*

All these features help defining how to build a service based on aggregated resources. Then, the major issue is to design a coherent interface for heterogeneous resources.

- **Access points**

The way in which it is possible to deal will cross-collection issues is highly dependent on what the service needs to do with the data. The main challenge for cross-domain collections, is to define how users shall access the metadata and/or resources.

#### *Search interfaces*

This must take into account how users browse collections : "There is not much research available about how people search archives"<sup>69</sup>. Researches are being developed on those issues, especially on cross-domain collections. Still the major difficulty is to find common concepts in all descriptions which represent comprehensive access points for users. The main orientation in a user-centred approach is to implement simple and thin interfaces, with just a few access points. However, intelligent customized applications can implement user search scenarios according to the type of content and audience considered.

To search portals, catalogues and finding-aids, simple keywords box are usually implemented for a simple-search interface, performing query on the full text of the metadata set.

Since data are aggregated, the advanced search interfaces include an access point on the data provenance : the repository, the institution, possibly the set as well, such as in the ARC service. On the metadata content, they are either DC-based interfaces or concept-based interfaces.

---

<sup>69</sup> MacKenzie Georges, Kristiansson Göran, "How Real Archivists can learn to love the OAI - A review of the potential for using the Open Archives Initiative Metadata Harvesting Protocol in conventional archives", OAIForum community-specific report, 2002

*Minerva work on interoperability to build a Dublin Core Culture*<sup>70</sup>

"This picture is also identified in the DIGICULT Report:-

*The controlled vocabularies in use today target the highly specialised and knowledgeable academic community, with the effect, that – if offered online – indexes are rarely used. As Sandy Buchanan, Resource Manager at SCRAN, UK, knows from experience: "80% of our users use text searches, and only 20% make use of structured searches such as indexes. What we need are tools that are comfortable for people. It is not about adapting the users to the Internet, but the other way around."*

DigiCULT ERT, Amsterdam, September 25-26, 2001 in Mulrenin, A., Geser, Guntram & Others, 2001, p. 176.

SCRAN uses the 'Who, What, Where and When' approach, and has found that this is useful for most of the 20% of users who use more advanced search options. The experience of the British Museum COMPASS project is that this approach can be even more successful for users when a thesaurus search is enabled behind this approach, both for the use of non-preferred terms, but also to deal with plurals – users tend to search for 'mummies' rather than 'mummy'."

#### *The Dublin Core Culture : a retrieval-oriented metadata format*

Access points on heterogeneous content are a major challenge for all types of cross-domain applications. The Dublin Core Culture<sup>71</sup> proposed by the Minerva project, on the basis of previous experiments of concept-based access points is a retrieval-oriented metadata set.

The concept-based access points allow to consider access through the user approach and by avoiding references to sector-specific common descriptive standards. The Dublin Core Culture is based on the 'Who, What, Where and When' access points. The traditional DC access points which are useful for retrieval, not for documenting a resource, are organised to match those top level access points. The DC Culture is currently being tested in the UK together with an Open Archives Initiative architecture and it should provide a possible answer to the question of standard access points to cross-domain cultural resources through data aggregation.

By selecting a "retrieval-oriented metadata set", the 24 Hour museum project also guarantees that the intellectual creation of the descriptive metadata are not included in the common schema used in the service, so that no major usage and IPR problem can be raised from the metadata exposition.

#### *Exploiting the wealth of controlled terminologies*

Experiences also demonstrate the necessity to preserve the wealth of controlled vocabularies. "Our experience proves that rich metadata sets not only provide a way to give users a powerful search interface, but also help users to review the search results. Users have the flexibility of sorting and grouping by rank, date stamp, subject or archive"<sup>72</sup>.

The ARC project has implemented an interface builder to search concepts in a dynamic list of subjects<sup>73</sup> which includes the possibility to search either full text or the list of existing subject fields recorded in the database and select the suitable ones. There is no re-building of thesaurus but the categories and sub-categories allow to retrieve the subjects eg. "Nicaragua – History – Flibuster War".

Other possibilities are to build crosswalks between existing terminologies and a standard terminology for the access service and to translate full text search into subject categories. In any case, the number

<sup>70</sup> Dawson David, "Minerva report 4.01 on interoperability", April 2003

<sup>71</sup> <http://www.minervaeurope.org/DC.Culture.htm>

<sup>72</sup> Liu Xiaoming, Maly Kurt, Zubair Mohammad, Qiaoling Hong, Michael L. Nelson\*, Frances Knudson\*\* and Irma Holtkamp, "Federated Searching Interface Techniques for Heterogeneous OAI Repositories", Journal of Digital information, volume 2 issue 4, May 2002, <http://jodi.ecs.soton.ac.uk/Articles/v02/i04/Liu/>

<sup>73</sup> [http://arc.cs.odu.edu:8080/oai/advanced\\_search.jsp#](http://arc.cs.odu.edu:8080/oai/advanced_search.jsp#) and Federated Searching Interface Techniques for Heterogeneous OAI Repositories", Journal of Digital information, volume 2 issue 4, May 2002, <http://jodi.ecs.soton.ac.uk/Articles/v02/i04/Liu/>

of visible categories should be reduced in cross-domain approaches such as demonstrated in the DCHC experiment which has led to reduce document types categories<sup>74</sup>).

- **Display issues**

To display the results of aggregated resources, several difficulties may arise, such as too many responses which make the result incoherent or the lack of focus of responses. Filter and display options can help improving user experience, especially for aggregated data.

*Information displayed*

Interesting functionalities found in existing experiments are :

- *indicate number of results per repository (ARC);*
- *link to other metadata formats (see ARC, SBN-Online digital Library);*
- *always mention the repository (and possibly the dataset) the record comes from;*
- *provide a link to the original system (on data provider's Website);*
- *display the resource type (still image, moving image, bibliographic record...);*
- *display thumbnails;*
- *interface and search customisation.*

*The Illinois University has led a study on student users of cross-domain resource discovery portal. The conclusions were<sup>75</sup> :*

- *Not disappointing user : mention whether there is a resource (image, text ...) or only a description : "learn more" / "view" or precise this is a bibliographic record and propose the possibility to only access where the resource is available online. (See AMOL Open collections for an example).*
- *"Distinction between portal and data provider unimportant to users" and "owning institutions accorded equal credibility"*
- *"Users unable to make use of search results" : too many unranked and advanced search rarely used (simple DC -based search interface)*
- *"Analog only collections excluded"*
- *....*

*Do not disappoint user*

Several types of services mix analogue and digital content. It is then necessary to at least provide a functionality to exclude analogue only content which may be useful to a specific category of users ready to get to the place considered. This was not the case of students testing the Illinois DCHC portal for example so that the implementers have ended up excluding totally descriptions of analogue only content.

*Full text search*

The protocol does not implement any functionality to mix the retrieval system with full text search. This can lead to impoverish retrieval. If willing to use full text as well, the service will need an additional search engine<sup>76</sup>. The problem is then to harvest various types of resources, many of them not usually accessible with classic search engines, and relate the result to standard metadata, given that sometimes, a single OAI record will match many Web pages. Then a solution is to harvest all textual information useful for resource discovery (including text field with textual content for retrieval purpose only) or to send specific robots which may have to index many pages for a single record.

<sup>74</sup> Cole Timothy W., "Using OAI-PMH to aggregate metadata describing cultural heritage resources", presentation ALA/CLA annual meeting, 22 June 2003, Toronto, <<http://dli.grainger.uiuc.edu/Publications/TWCole/ALA2003OAI/>>

<sup>75</sup> Cole Timothy W., "Using OAI-PMH to aggregate metadata describing cultural heritage resources", presentation ALA/CLA annual meeting, 22 June 2003, Toronto, <<http://dli.grainger.uiuc.edu/Publications/TWCole/ALA2003OAI/>>

<sup>76</sup> see Aquitaine Patrimoine implementation

The OAI-PMH is a protocol for resource discovery and an access point to heritage resources may be full text search on the resource (eg. articles or Websites). Then it is possible to use the protocol not only to transfer metadata but also to collect rough text or XML-formatted resources (eg. TEI).

### *Display thumbnails*

The content of the repository can indeed include the resource (text or image) or an extract or low quality version of the resource, such as a thumbnail. It is important to determine what shall be done with the original resources. For example, with images, it shall be interesting to display thumbnails (like the cross-domain RLG CMI<sup>77</sup> service). The thumbnails can be included in the repository and collected together with the metadata so that when the results are displayed, it is easy to display the thumbnail. It is also possible to display the thumbnail from the original system (this increases hits on the data provider's server but access time is larger and the connection insecure). The choice of the 24Hour Museum project in the UK with museum collections was to collect thumbnails and provide access to full screen images on the original Website.

### *Linking to other resources*

When displaying information, links to related resources may be proposed according to a terminological tree or related terms, to other objects from the same collection, the same OAI set or the same data provider.

*Cyclades project*<sup>78</sup>

“Collections

- Collection: a set of metadata records collected together, according to a record selection criteria, from the archives
- Collections are NOT materialized
- Collections represent the information space, while the OAI archives remain hidden to the users and communities
- Users and Communities may define their own collections, e.g. by means of meet, join and refinement of existing ones”

### *Displaying resources in service provider's layer*

Finally it is possible when providing access to a resource to display it in the service provider's Website. But then, integrated access leads to an ambiguity in the responsibilities. The user may send questions to the service provider on the content presented. A clear commonly defined policy must be published, as well as information on copyright issues, usage ... for all data represented, as a general agreement on the aggregated resources.

- **Common metadata schema**

The necessary access points and display functionalities lead to define one or more common schemas to harvest records. It can be a service-specific schema. Indeed, apart from standard metadata sets, the service providers may need specific metadata such as a thumbnail URL or subject-specific terminology. It shall be formally compliant to W3C XMLSchema standard<sup>79</sup> (XML schema validators may ensure the compliance with OAI specifications). This schema must have a persistent location and management of successive versions of the schema shall be foreseen. This probably includes the obligation to ensure compliance of previous version to the new one.

Unqualified Dublin Core is often used as a *pivot* metadata set. The interest of DC is mainly for cross-domain information, to communicate outside its community, though wealthier formats are preferable inside its community for richness of interpretation. Most existing systems only implement unqualified

<sup>77</sup> Research Libraries Group Cultural Material Initiative project <http://www.rlg.org/culturalres/>

<sup>78</sup> Straccia Umberto, "CYCLADES An Open Collaborative Virtual Archive Environment", EC/NSF Digital Library All Projects Meeting, March 2002, Rome, <http://delos-noe.iei.pi.cnr.it/activities/internationalforum/All-Projects/RomeSlides/CYCLADES.pdf>

<sup>79</sup> <http://www.w3.org/XML/Schema>

DC, however, larger metadata formats such as MODS, MARC-XML, SPECTRUM-XML are more and more implemented when dealing with data providers from similar sectors.

The approach of the DC Culture in the 24 Hours Museum is to focus on the necessary data to adequately discover resources rather than describing them. A distinct schema is used to display results, including a summary of the resource.

- **The granularity of descriptions**

The metadata aggregated and searched shall describe similar entities. "The issue of *defining Item and Collection* (and whether a resource is one or the other) may require a recommended set of definitions that a site can ignore if the site believes that the standard definitions do not apply to it."<sup>80</sup>

The OAForum community report on OAI for archives mentions the difficulty of providing access to EAD descriptions and other types of descriptions for this reason. The Illinois University has tested the implementation of OAI-PMH for heterogeneous resources, including EAD. Major difficulties have been encountered to relate them to other types of descriptions, especially granularity level of description and the very nature of EAD records as finding aids rather than descriptive metadata. The EAD records and various granularity levels are challenges to cross-domain access.

The description level of records must be mentioned. Most records coming from memory organisations are either collection level descriptions or more often item level description. An EAD record might include a collection description, sub-collections descriptions and item descriptions.

A mapping frame has been developed within the Illinois University OAI projects to generate multiple OAI-DC files out of a single EAD finding-aid. It raises a major problem which is the multiple guidelines available for EAD encoding, that "EAD encourages a wide range of possible tagging practices"<sup>81</sup> and that "further standardization of EAD markup practices would enhance interoperability"<sup>82</sup>. A terminology is clearly needed for the LEVEL tagging. The mapping task includes division of the EAD file and inclusion of several top-level elements in the item-level or sub-collection level descriptions. The University of Illinois-Champaign is currently building a tool to generate OAI-DC records out of EAD files.

---

<sup>80</sup> Giersch Sarah, Jones Casey, Sumner Tammy, "NSDL Evaluation Pilot - Preliminary Report of Collections Data & Users and Usage Data", sept 2002, [http://eduimpact.comm.nsdlib.org/doc\\_tracker/docs\\_download.php?id=230](http://eduimpact.comm.nsdlib.org/doc_tracker/docs_download.php?id=230)

<sup>81</sup> Prom Christopher J., Habing Thomas G., "Using the Open Archives Initiative Protocols with EAD", Joint Conference on Digital Libraries 2002, <http://dli.grainger.uiuc.edu/publications/jcdl2002/p14-prom.pdf>

<sup>82</sup> Prom Christopher J., Habing Thomas G., "Using the Open Archives Initiative Protocols with EAD", Joint Conference on Digital Libraries 2002, <http://dli.grainger.uiuc.edu/publications/jcdl2002/p14-prom.pdf>

*EMII-DCF report Content Creation Analysis, 2002 – RLG-CMI project, interview of Tony Gill*

“The granularity is not ambiguous. We would not load collection-level records unless their relationship to their member items/groups/collections was clear and unambiguous. We only *require* two levels of hierarchy; bottom-level item-level descriptions must have a top-level parent collection description (Diagram 1). However, we can also support group-level records, and any level of nested collections (Diagram 2) in between, if they are found within the contributed descriptive data (for example, EAD finding aids often have many levels of hierarchy).

Diagram 1: Items *must* belong to a parent collection:

Collection

    L Items

Diagram 2: Items *may* belong to groups, and collections *may* be nested to an arbitrary number of levels:

Collection

    L Collection

        L (...)

            L Collection

                L Group

                    L Items“.

Access to parts of an EAD record shall be possible, either by links to EAD nodes or through EAD file division to provide access to it together with other types of records (collection or item-level descriptions only).

- **Metadata crosswalks**

Once the common schema has been agreed to be used by the service provider, each data provider shall adequately make its records available in this format. The use of that schema, the concepts it includes shall be well defined. Mapping guidelines to the common schema must be designed so that all partners can use it in a similar way for they mapping task. This is especially important for a service-specific schema.

The guidelines may be as standard as possible, especially because if re-using DC mapped descriptions within another service, the mapping guidelines to OAI DC will not be possible to modify.

Since a service based on the OAI-PMH architecture relies on metadata quality, the quality of the final service also depends on the use made of the metadata elements. A proper table on elements used by different data providers will help build better search routines, possibly normalize content and establish proper guidelines.

## DCHC project analysis

Variations in DC element usage<sup>83</sup>

	Subject	Description
Dig lib	78%	36%
Museums, hist. Societies	93%	93%
Academic libraries	15%	13%

For each collection, the presence or absence of each element of the pivot schema has an influence on end-user search results. For each element, the service shall define :

- *whether element is possible to automatically extract ;*
- *language of element;*
- *whether it refers to terminology;*
- *whether it must be normalize (reprocess the value);*
- *whether it has relations (necessity to re-build relationships on service-side);*
- *whether it matches cataloguing rules.*

### ▪ Terminology issues

As previously stated about access points, it is a major challenge to take advantage of controlled terminologies used in original systems of data providers. Depending on the type of service which is set up, the approach of thesaurus mapping may be avoided. A major difficulty can be that in archives, libraries and museums, the concept of a "subject" for example (DC:Subject) may be slightly different. Within the same community, it may be possible to adopt a similar terminology but it is always a challenge :

#### *Management of subject element which is a different issue in all sectors in Digital Cultural Heritage Community project<sup>84</sup>*

In order to cope with the difficulty of defining subject as different concepts, the Illinois DCHC experiment has led to an intermediary format of qualification of DC:subject :

"In using the Dublin Core metadata schema for the framework of the database, several of the fields were qualified. In particular, subject fields were qualified with sub-fields, including keywords. This scheme helped to accommodate the different vocabularies and descriptive structures used at the various partner institutions as well as to include reference to the teachers' curriculum units and the Illinois Learning Standards. In turn, these sub-fields can be easily mapped into the "Subject" field to accommodate cross-repository searching, which is now implemented on a preliminary basis using Open Archives Initiative (OAI) protocols."

"Although the original DCHC concept contained no provisions for an interpretation field, after persuasive argument from the museum participants, the database fields were modified to include a separate interpretive sub-field in the subject field.[...] As a result of the DCHC project, the library community as well as the other partners understand that the significance of having an interpretive field in the database includes both ensuring proper historic documentation and appropriate integration into the curricular unit."

The "interpretive field" helps underlining the educational value of the material.

<sup>83</sup> Cole Timothy W., "Using OAI-PMH to aggregate metadata describing cultural heritage resources", presentation ALA/CLA annual meeting, 22 June 2003, Toronto, <<http://dli.grainger.uiuc.edu/Publications/TWCole/ALA2003OAI/>>

<sup>84</sup> Bennet Nuala, Sandore Beth, Pianfetti Evangeline, « Illinois Digital Cultural Heritage Community – Collaborative Intercations among libraries, museums and elementary schools », in D-Lib magazine vol 8-1, January 2002, <http://www.dlib.org/dlib/january02/bennett/01bennett.html>

Possible approaches include :

- A *pivot* or common intermediary terminology can be used, whether created from scratch or adopted from a standard<sup>85</sup> : every subject system must be mapped to the pivot terminology : drawback is that it is time-consuming and must be done as an obligation for newcomers, but it must only be done once. New concepts are difficult to add to the pivot terminology, so that it is much preferable to use standard, largely experienced terminologies to avoid such a process.
- Many to many terminology mapping<sup>86</sup> : does not use *pivot* terminology but rather creates equivalences between a terminology with each of the others. It is time-consuming, each new collection aggregated which uses a new terminology will entail a new mapping task by all participants but results may be much more valuable.
- No terminology mapping : subject is searchable as free text, then the full thesaurus tree shall be included in the metadata value.
- Free text equivalences : no thesaurus is used but the system is able to interpret user query according to the thesaurus, when the user asks for *cemeteries* to propose a record which only contains *funeral architecture* as a subject. Terminology mapping can be incomplete. User queries are mapped to terminologies, the semantic issues may be very important and the work very heavy, can be only based on the mapping to a simple pivot terminology. It is invisible to the user and may create confusions (natural language and ontology-driven systems). The system may be constantly improved through recording user queries.
- Gateway style mapping : does not try to build a new *pivot* terminology equivalent to the original ones, but rather tries to include gateway-style categories in order to roughly present and organise the content but they do not fill the same role.

*A methodology to map metadata in Illinois university*<sup>87</sup> :

- *“extract and analyse the element values”, for example see whether DC elements are equally used in collections (see above table, previous section)*
- *“determine how each element is interpreted and what controlled vocabulary if any, is used”*
- *“determine focus and vocabulary for normalization”*
- *“normalizing the data” : mapping and determine processing of new terms*
- *“providing services based on the normalization process”*
- *end-user search interface : design search scenarios but stay simple*

On subject, the DCHC has preserved concepts by a qualified DC stage : eg. DCQ Date.created, DCQ.coverage.spatial, DCQ.coverage.temporal, DCQ.coverage.cultural

The consideration for terminology issues also depends on the type of data and the user needs. A major criteria for user search is geographic coverage, then the relation to a thesaurus or a geographic information system may be very useful (eg. in the Aquitaine Patrimoine regional portal). All the same, to indicate to the user what he/she will find behind metadata, the standard DC Type or extended versions can be proposed (including distinction between still and moving images for example). A normalization work can be done to allow a clear mapping of to a standard terminology in those cases.

<sup>85</sup> eg. The Getty AAT, the RAMEAU terminology in French libraries. The Illinois experience on the DCHC project shows that the mapping task to standard terminology (AAT , LCSH) has proved very difficult to implement (see Bennet Nuala, Sandore Beth, Pianfetti Evangeline, « Illinois Digital Cultural Heritage Community – Collaborative Interactions among libraries, museums and elementary schools », in D-Lib magazine vol 8-1, January 2002, <http://www.dlib.org/dlib/january02/bennett/01bennett.html>). For cross-domain information, no existing terminology has been reached the point to be a standard. Several attempts, such as the Cornucopia museum terminology have been assessed by other CH sectors : <http://www.cornucopia.org.uk/term.html> and used as a basis to build the cross-roads used in the EnrichUK portal : <http://www.enrichuk.net/browse/?browsehow=subject>.

<sup>86</sup> MACs project choice for multilingual heterogeneous thesauri of 3 European national libraries

<sup>87</sup> Cole Timothy W., Kaczmarek Joanne, Marty Paul F., Prom Christopher J., Sandore Beth, Shreeves Sarah, “Now that we’ve found the ‘hidden Web’, what can we do with it?”, Museums and the Web 2002 conference, <http://www.archimuse.com/mw2002/papers/cole/colefig1.html>

For other metadata elements, a simple syntactic control can be very useful, for example on temporal coverage, by converting periods into time intervals (proper dates). Normalization on names can be done according to Anglo-American cataloguing rules (AACR2)<sup>88</sup> recommended in the CIMI guide to best practices of DC, they shall be related to the LEAF<sup>89</sup> service which connects name authority records in Libraries and archives, based on EAC and ISAAR (International Standard Archival Authority Record) rules<sup>90</sup>.

- **A multilingual environment**

Terminology issues are especially important with multilingual content. In a European context, the service may have to send users to resources in various languages and handle records in various languages. The language of the metadata values shall be defined in a machine readable format so that the service application can process the content as a specific language. This information can be mentioned in the “about” section of the record but it implies that the full repository has descriptions in the same language. Language can also be specified within an element of the common schema. Finally, another solution is, for each element of the schema used, to define the language of the value and possibly the terminology it refers to and use the xml:lang attribute.

Solutions to access information in various languages include cross-language information retrieval, mainly focused on unstructured data<sup>91</sup>, terminology mapping like in the TEL service which uses the MACS authorities in three languages for several subjects and on the fly translations using automatic translation tools to make descriptions accessible.

The language problem is a growing challenge for European heritage stakeholders, both to process metadata sets and to ensure the accessibility of results (descriptions). The issue is very similar to the standard terminology issue, with the need to clearly mention which language is used, all the same as a value of standard terminology shall be identified in a machine-readable format.

#### **4.3 Resource aggregation**

The interoperability issues considered above shall be taken into account on service-side through systems to handle all these types of content and the specificity of each collection. The aggregation process leads to record information on the conditions for harvesting repositories and the specific needs to implement tools for reprocessing content.

- **Technical features**

Various open source harvesters are available and registered on the Openarchives.org Website. Flows of records when harvesting shall be evaluated and the harvester can divide the harvest in order to avoid overloads. The transfer of full EAD files, for example, may entail specific methods due to their possible size.

Repositories' availability must be tested when harvesting, with a procedure to launch in case the repository is not available (eg. come back in an hour, three times, then declare connection failed). A system can be implemented to detect whether a repository is too often down so that it does not damage the service's quality.

- **Harvesting frequency to improve synchronization**

The management of an OAI service relies on harvests. The metadata handled by the service shall be synchronized with the repository content, this depends on frequency of both data update and harvesting.

“The nature of the data provider can influence how often records are modified or updated. E-print type data providers are likely to have a small but steady stream of ongoing daily or weekly updates.

---

<sup>88</sup> <http://www.nlc-bnc.ca/6/18/s18-219-e.html#AUTH>

<sup>89</sup> <http://www.crxnet.com/leaf/index.html>

<sup>90</sup> [http://www.hmc.gov.uk/icacds/eng/ISAAR\(CPF\)2.pdf](http://www.hmc.gov.uk/icacds/eng/ISAAR(CPF)2.pdf)

<sup>91</sup> see the Cross Language Evaluation Forum at European level, <http://clef.iei.pi.cnr.it:2002/>

Museum or historically oriented archives will have an initial bursty period of accession (perhaps all at once), but then are likely to trickle down to just infrequent error corrections or edits.”<sup>92</sup>.

Harvesting frequency can be adapted to modification dates of records by implementing a functionality to measure changes, such as demonstrated in the ARC graph on ePrint archives<sup>93</sup>). “Although not currently implemented by any data providers, if a data provider allowed the metadata to change based on usage, annotations or reviews, the required harvesting would likely become significant.”<sup>94</sup>.

Other solutions are considered to mix push/pull mechanisms for the data provider to require harvests when modifying data. This could improve data reliability and possibly increase security control by delivering data outside the firewall instead of expecting harvests<sup>95</sup>.

The harvest should follow the repository update frequency. If it is automatic, then once a month seems commonly admitted<sup>96</sup> for catalogues and finding aids, based on existing digital libraries updates<sup>97</sup>. This avoids any problem with datestamp granularity (day-level).

#### ▪ **Harvesting profiles**

Those parameters must be taken into account to manage the service, but they may differ from one repository to the other. The service provider shall therefore record and manage information on how to handle each data provider’s repository according to its features :

- *Which OAI sets;*
- *Which volume of records;*
- *Which harvesting frequency;*
- *Specific harvesting time / day if preferences;*
- *Specific password protection.*

Those harvesting profiles can be completed by all other types of information on the way content can be processed, displayed and accessed (whether specific information shall be published on the service’s Website, whether metadata use is restricted, IPR and usage, branding ...), according to initial data analysis.

Finally, data on harvests can be stored : dates of harvests, non accessibility of a repository ...

*The NSDL portal mixes various data types and various metadata collection methods*

*“TECHNICAL OPTIONS FOR RESOURCE INGEST INTO THE NSDL COLLECTION [...]*

- Method A (OAI harvesting of metadata with automated checking) will be the standard method for ingest of collections that meet the requirements of metadata records for all resources and support of an OAI server.
- Method B (OAI static repository) will be an alternative to method B, for small collections.
- Method C (crawled collections of textual materials with some level of manual selection) will be the standard method for ingest of collections that satisfy key criteria (open-access, crawlable, and primarily textual), but do not meet the criteria for method A or B.”<sup>98</sup>

<sup>92</sup> Liu Xiaoming, Maly Kurt, Zubair Mohammad, “Arc - An OAI Service Provider for Digital Library Federation”, in D-Lib Magazine, Volume 7 Number 4, April 2001, <http://www.dlib.org/dlib/april01/liu/04liu.html>

<sup>93</sup> Liu Xiaoming, Maly Kurt, Zubair Mohammad, “Arc - An OAI Service Provider for Digital Library Federation”, in D-Lib Magazine, Volume 7 Number 4, April 2001, <http://www.dlib.org/dlib/april01/liu/04liu.html>

<sup>94</sup> Liu Xiaoming, Maly Kurt, Zubair Mohammad, “Arc - An OAI Service Provider for Digital Library Federation”, in D-Lib Magazine, Volume 7 Number 4, April 2001, <http://www.dlib.org/dlib/april01/liu/04liu.html>

<sup>95</sup> according to the Gnutella system mentioned in Liu Xiaoming, Maly Kurt, Zubair Mohammad, “Arc - An OAI Service Provider for Digital Library Federation”, in D-Lib Magazine, Volume 7 Number 4, April 2001, <http://www.dlib.org/dlib/april01/liu/04liu.html>

<sup>96</sup> <http://www.pictureaustralia.org/guide.html>

<sup>97</sup> The considered frequency of repository update in the French National Library is once a month, just like the update frequency of the Gallica digital library

<sup>98</sup> Saylor John M., “Draft – NSDL collection development policy draft v030715”, July 2003, [http://content.comm.nsdlib.org/doc\\_tracker/docs\\_download.php?id=452](http://content.comm.nsdlib.org/doc_tracker/docs_download.php?id=452)

OAI-PMH standard service description can help recording all necessary information about data providers and such a task is currently being performed. Those information may be recorded at set level rather than for each repository.

*OAI-PMH Service Description Tabular Form<sup>99</sup>*

Name	EncodingScheme/Language	Value
<b>Title</b>	Language. Optional	Name of the service. Required Not Repeatable
<b>identifier</b>		Data provider's identifier of the metadata record of this service. Required Not Repeatable
<b>description</b>	Language. Optional	A free text summary description of the service. A service description should be provided only for a transactional service (not associated with a collection). Optional. Not Repeatable
<b>Locator</b>		The URI of the access point for the service. Required. Not Repeatable
<b>accessType</b>	AccMthdList	oai-pmh
<b>serviceType</b>	SvcTypeList	Type of the service. Optional. Repeatable
<b>accessRights</b>	AuthList	Access control for the service. Required. Repeatable
<b>accessDomain</b>	DNSDomain	Domain where service is available. Optional. Repeatable
<b>supportsStandard</b>	StdsList	Indication of OAI-PMH version supported by the service. Optional. Repeatable
<b>seeAlso</b>		The global identifier of a document that provides more information about using the service. Optional. Repeatable
<b>administrator</b>		Data provider's identifier of the metadata record for the agent that has responsibility for the electronic environment in which the collection is held. Required. Repeatable
<b>Creator</b>		Person or organisation who created the metadata record. Optional. Repeatable
<b>contributor</b>		Person or organisation who modified the metadata record. Optional. Repeatable
<b>Created</b>	W3CDTF	Date when metadata record was created (or manually updated). Optional. Repeatable
<b>publisher</b>		Organisation or project that published this metadata record. Optional. Repeatable
<b>metaLanguage</b>	RFC3066	Language of the metadata record. Optional. Not Repeatable
<b>Source</b>		URL of source record from which this metadata record was derived. Optional. Repeatable

- **Reprocessing content**

When collected, records shall be selected, according to a quality check on their values (possibility to reject incomplete records), XML validity, and de-dupping.

A normaliser (re-processing application) can improve their quality for the specific use of the service. Syntactic control can be performed, values added and the records can be re-organised to be included in the service application.

The de-dupping task is not trivial and well-known in Z39.50 gateways. This functionality is indeed rarely implemented in online environments "due to the complexity and expensiveness of the duplicate detection algorithms"<sup>100</sup>. The University of Michigan currently focuses efforts on "determining a method for handling duplicate records"<sup>101</sup>.

<sup>99</sup> Apps Ann, "JISC Information Environment Service Registry (IESR) OAI-PMH Service Input Templates », 2003, <http://www.mimas.ac.uk/iesr/metadata/templates/svcoai-template.html>

<sup>100</sup> Sfakakis Michalis, Kapidakis Sarantos, "An architecture for online information integration on concurrent resource access on a Z39.50 environment" in Koch Taugott, Torvik Solvberg Ingeborg, "Research and Advanced Technology for digital libraries" 7<sup>th</sup> European conference, ECDL 2003, Trondheim Norway, August 17-22, 2003, Berlin 2003.

<sup>101</sup> Halbert Martin, Kaczmarek Joanne, Hagedorn Kat, "Findings from the Mellon Metadata Harvesting initiative", in Koch Taugott, Torvik Solvberg Ingeborg, "Research and Advanced Technology for digital libraries" 7<sup>th</sup> European conference, ECDL 2003, Trondheim Norway, August 17-22, 2003, Berlin 2003

It is also possible to re-create relationships between records, and between a record and a thesaurus. The thesaurus is external and managed outside the protocol (or as a different set), it is not transferred at the same time as the record. However, it can be related to a crosswalk table for example or to an external service (such as the MACs service in the TEL portal).

Nevertheless, it is not always necessary to re-create the relationships between records on service-side if the resources are accessed on the data provider's Website (or are analogue only resources). The single purpose of resource discovery only requires that from the description of the collection, the user can access the right part of a database for example, he/she will then be able to access its components from the data provider's Website. The service provider only needs to consider records in the same way, whether collection-level or item-level description.

- **Managing aggregated collections**

The data collected shall be managed and presented to the service's user in a comprehensive way.

Collections, terminologies in use, provenance, possibly metadata creators, original system (publisher)... shall be presented to the end-user. Collection level description<sup>102</sup> may be used, according to a definition at institution level (one description per partner institution) or as logical datasets.

When including the metadata in the information system, a normalisation process may be necessary, but in any case, their provenance shall be recorded. The ARC for example, for each record, stores information on the institution (data provider) and the OAI set corresponding to the record.

The aggregated resources will then be handled within the service, in a similar way. The identity of the service provider highly depends on its collection and this is a major challenge for services based on aggregated resources, since they do not control data sources.

Service's metadata collection management implies that the service provider is aware of the data providers' repositories evolution. It shall take into account new data providers' collections. Each individual repository may have growing or changing collections. The overall collections made available (or stored) by the service provider may be modified and its identity change according to the new data providers and the more or less growing status of the data providers' collections. This can be an important management issue for community-based services. For example, for the "Aquitaine Patrimoine" project of a regional portal on heritage, the available resources shall provide an overview of the region on the Internet and its heritage will be perceived according to the resources available in the portal. If various data providers on a specific subject have very quickly growing collections or if all new partners are focused on a specific subject, for example industrial heritage, the overall value of the service, its audience and coherence may change. This is a major challenge for several services<sup>103</sup>.

Experiments are also led to facilitate the management of aggregated resources through "self organising maps" of the metadata contained on service-side<sup>104</sup>. This automatically organises the data in clusters and sub-clusters as a new collection.

#### **4.4 Displaying information on the service**

The use of OAI-PMH is invisible to the end-user. However, the concept of data aggregation shall be presented and it is the responsibility of the service provider to publish information on the collections and data sources on the one hand, on the other hand on the relation between the records found on the service and those on the original system's Website.

- **Information reliability**

The resources available through the service must indeed be clearly identified by the user. A short description of collections harvested must be published and for each record, its provenance (data

---

<sup>102</sup> see RSLP model and work on DC collection, <http://www.ukoln.ac.uk/metadata/rsip/schema/>

<sup>103</sup> see also the collection issue dealt with in the NSDL evaluation pilot report. Giersch Sarah, Jones Casey, Sumner Tammy, "NSDL Evaluation Pilot - Preliminary Report of Collections Data & Users and Usage Data", sept 2002, [http://eduimpact.comm.nsdlib.org/doc\\_tracker/docs\\_download.php?id=230](http://eduimpact.comm.nsdlib.org/doc_tracker/docs_download.php?id=230)

<sup>104</sup> Kim Hyunki, Choo Chee-Yoong, Chen Su-Shing, "An integrated digital library server with OAI and self-organizing capabilities", in Koch Taugott, Torvik Solvberg Ingeborg, "Research and Advanced Technology for digital libraries" 7<sup>th</sup> European conference, ECDL 2003, Trondheim Norway, August 17-22, 2003, Berlin 2003

provider), possibly through branding (data provider logo displayed with each resource reference). Those information shall be transferred together with the record if it is transferred to another service.

Since data update is a major challenge of OAI models, the user may find more information on the data provider's Website, or there may be slight differences between the records displayed on service-side and the ones accessible on data provider's application. Then, in order to improve data reliability, the records and/or the descriptions of collections shall be published together with last update and/or harvesting frequency.

- **Integrated access to resources**

Integrated access to resources also leads to the creation of a new environment for representing resources. This environment shall be clearly identified through policy statements, responsibility statements, a particular identity and audience, different from the ones of each data provider.

The access service sets its own policy on IPR, data usage, data update, selection criteria for aggregated resources (eg. avoid pornographic content, especially for services based on shareable repositories)<sup>105</sup> and makes them available for agreement by data providers. As an example, the US Research Libraries Group Cultural Material Initiative which collects metadata and thumbnails from its partners and provides access to original Websites for full screen images ensures usage policy for the whole datasets through common "terms of use"<sup>106</sup>.

Due to integrated access, users may not understand clearly the share of responsibilities as regards to the content presented. Users may therefore send questions to the service administrator on content, especially if presented inside service provider's frames and layer. The Picture Australia service administrators answer questions and forwards to the data providers' mailing list whenever it cannot ensure good user question follow up<sup>107</sup>.

The user browsing experience from the service provider's Website to the data provider's Website shall be comprehensible and coherent. If the resources are displayed in the data provider's frame, then a way back to the service provider's Website must be ensured (if not a new window).

The charter can take those issues into account and include quality level of data providers' Websites as well, such as in the PictureAustralia project.

*Picture Australia recommendations for data provider browsing*

"It is also necessary to give some thought to the links provided to the user when they reach the image provider's site. The links should provide access where available to:

- copyright conditions,
- an ordering capability,
- the local image service, for further searching, and reference assistance."<sup>108</sup>

It is possible here to refer to common quality criteria for Web applications to ensure that the user will find a similar level of quality as for the information it will access, the definition of an IPR policy and copyright conditions of material displayed for example. This is valuable for both the service and the data provider, and it is even more necessary within a clearly defined community building a service, as part of the charter, agreement or contract.

<sup>105</sup> see Saylor John M., "Draft – NSDL collection development policy draft v030715", July 2003, [http://content.comm.nsdlib.org/doc\\_tracker/docs\\_download.php?id=452](http://content.comm.nsdlib.org/doc_tracker/docs_download.php?id=452)

<sup>106</sup> <http://www.rlg.org/agreements/termsrcm.html>

<sup>107</sup> see the PictureAustralia mailbox section <http://www.pictureaustralia.org/guide.html>

<sup>108</sup> <http://www.pictureaustralia.org/guide.html>

*Analysis of an OAI-PMH service for cross-collection heritage information retrieval according to the Minerva quality criteria for cultural Web applications<sup>109</sup> :*

Quality grid for cultural Websites	OAI services issue
<b>Definition of</b>	
institution responsibility	<i>service provider's responsibility</i>
services proposed	<i>e-commerce, full screen images display, ....</i>
who the service represents	<i>definition of the community : data providers descriptions and data providers' responsibility</i>
audience	<i>of the service provider, as opposed to the data provider's ones</i>
representation of the identity of the service	<i>of the service provider, as opposed to the data provider's ones</i>
<b>Set a policy on</b>	
IPR	<i>common policy statement on service's Website and possibly commitment to an IPR policy set on data provider's Website if user gets there</i>
maintenance	<i>harvesting frequency, effort for best synchronization</i>
content selection	<i>define coverage of collections, set selection criteria, take into account overall services' collections evolution when adding data providers and when data providers' collections grow up</i>
<b>Usability : Navigation</b>	
context evidence : where user is	<i>on which Website (data provider's one or service provider's one), searching record, viewing thumbnails...</i>
link evidence : where user is goes	<i>be clear on the access to a resource or a bibliographic record, on whether user will stay on the service provider's Website or not</i>
backtracking soundness	<i>ensure that if the user is sent to the data provider's Website, he/she will be able to get back to the list of results</i>
<b>Identification of</b>	
accessibility : access conditions, necessary plug-ins if any...	<i>access to all resources may not be free, reference to interactive resources may require specific plug-ins or browser level ... which data types (moving image, sound...) and WAI compliance</i>
content sources	<i>always mention data provenance, possibly branding</i>
currency	<i>last harvest on the set / repository shall be mentioned</i>

▪ **Aggregation procedure for newcomers**

The service shall publish an information on the OAI model and possibly invite new members in the community. The OAI-PMH compliance is a very important information to display for professional use if the service's community is to be extended : the "Contributing to PictureAustralia" section on Picture Australia Website is providing for example the condition to be a Picture Australia data provider<sup>110</sup>.

An aggregation procedure for new data providers must be designed within the charter, especially for services heavily based on a formal community, including compliance possible acceptance by other data providers and compliance with all technical and documentary conditions defined in the charter.

<sup>109</sup> see Minerva project, <http://www.minervaeurope.org/>

<sup>110</sup> <http://www.pictureaustralia.org/join.html>

*"Requirements for participating in PictureAustralia*

PictureAustralia seeks the following obligations from a participating agency:

- signatory to a Service Level Agreement, accompanied by payment of an annual fee,
- provision of appropriate metadata,
- designated participant contact details and referral contact details for mailbox queries,
- secure access to the host server of the metadata and associated image files,
- creation of thumbnail images to consistent dimensions for display purposes,
- a digitised copy of the agency's corporate logo and, if necessary, the related image service logo,
- descriptive text outlining the scope of the agency's image contributions to the PictureAustralia service,
- a selection of representative images for digital and printed reproduction purposes, to use in publicity materials, and
- dissemination of PictureAustralia publicity materials on a continuous basis."<sup>111</sup>

- **User approach and service evaluation**

Major difficulties when aggregating heterogeneous resources are the relevance of search results, usability and data update. Then, the evaluation of the project use indicators on operational efficiency (management of harvesting procedures), end-user evaluation and data provider satisfaction. Those include :

- the way the collections are growing and the way this is handled;
- whether the OAI protocol lowers barrier for data providers to participate to the service;
- whether it facilitates conversion and aggregation of collections, in the specific domain concerned ;
- whether the service has contributed to build new communities.

The cultural heritage service relying on harvested data shall therefore foster heritage communities and lead them to work with other communities. The agreement between service's partners (data providers) is an opportunity to set criteria for user experience, presentation of resources, quality of metadata and Web application and to harmonise the way partner institutions handle digital resources. This brings strong organisational impacts on management of digital heritage in memory organisations.

---

<sup>111</sup> <http://www.pictureaustralia.org/guide.html>

## 5 Conclusion

The main interest of aggregated resources seems to lie in the possibility to build cross-collection services based on cultural heritage resources. However, synchronization of intermediary services, such as on thesaurus, name authorities or even OpenURL registries and usage records (Usage logs project) may be easier to organise since the documentary difficulties are more focused<sup>112</sup>. Both types are currently being developed and they already demonstrate the potentiality of the technology to lower traditional interoperability barriers and facilitate collective services. The OAI environment appears very well adapted to the major evolutions of the services offered from cultural heritage resources, including data reusability and distributed content.

The OAI model then brings an opportunity for memory organisations to build cooperation policies but it also contains the “roles” of new actors of the digital library field, whether aggregators or service providers which may not be traditional memory organisations. The OAI architecture can facilitate the collaboration between institutions to for cross-domain resources discovery. This is a major opportunity for the integration of cultural heritage institutions in the digital networks.

Generally, the use of the OAI technology shall be invisible to the end-users who will only benefit from the development of new services based on cultural heritage resources. For data providers, that technology can constitute a rather simple solution to technical interoperability, but it is much more an organisational model which raises the opportunity to get involved in community-based services. The service providers which use OAI repositories and face the challenge to show advantages of their service and community to the data providers and to the end-users.

The experiences related to cross domain services have demonstrated indirect organisational impact :

“Curators and librarians indicated that they were motivated to join the project because it provided them with the impetus to do a number of things that they considered were institutional priorities but often had been un-funded mandates, including:

- Focusing on a community outreach project;
- Forming new partnerships with previously un-served or under-served groups; and
- Identifying and assessing collections for digitization.”<sup>113</sup>

The benefit for the data providers is also to increase the visibility of their resources. In the State Library of New South West, in 2002/3 there were 2,960,665 requests for pages referred by PictureAustralia, representing 37% of the total requests for pages to State Library image server. The Australian National Archives also recognize this impact : “From September 2000, website visits were boosted by the inclusion of our photographic database, PhotoSearch, in the PictureAustralia website maintained by the National Library. This website offers a single point of access to some of Australia's largest pictorial collections. In the first month of operation, almost 100 000 visitors accessed PhotoSearch on our website via PictureAustralia.”<sup>114</sup>. Those experiments clearly show the efficiency of building or participating to services based on aggregated resources to reach a broader audience<sup>115</sup>.

To ensure the benefit of the technology to small institutions, the implementation of aggregators seems to compensate the lack of technical skills and financial opportunities. Thus, to spread the use of the protocol in the heritage sector, funding bodies can facilitate the development of aggregators, possibly built by the service providers themselves<sup>116</sup>.

The National Representatives Group on digitisation of cultural and scientific heritage (NRG), gathering representatives of all European ministries of culture has demonstrated a clear interest for the OAI-PMH technology and the Minerva project in charge of implementing the political decisions of the NRG, examines the way interoperability is handled in various countries and disseminates common standards for a European Information Environment, with notably a clear recognition of the major interest of the

<sup>112</sup> as shown in Van de Sompel Herbert, Young Jeffrey A., Hickey Thomas B., “Using the OAI-PMH... differently”, in D-Lib Magazine July/August 2003, vol 9, number 7/8, <http://www.dlib.org/dlib/july03/young/07young.html>

<sup>113</sup> Bennet Nuala, Sandore Beth, Pianfetti Evangeline, « Illinois Digital Cultural Heritage Community – Collaborative s among libraries, museums and elementary schools », in D-Lib magazine vol 8-1, January 2002, <http://www.dlib.org/dlib/january02/bennett/01bennett.html>

<sup>114</sup> National Archives of Australia “Annual report 2001”, [http://www.naa.gov.au/publications/corporate\\_publications/ar2001/outcomes\\_outputs\\_reports2.html](http://www.naa.gov.au/publications/corporate_publications/ar2001/outcomes_outputs_reports2.html)

<sup>115</sup> Davidson Stephen, “The Open Archives Initiative (OAI) Sheet music project - a gateway to sheet music collections on the Web”, Sheet Music Roundtable, Music Library Association 2003, Austin, <http://unitproj.library.ucla.edu/music/oaisheetmusic/mla.ppt>

<sup>116</sup> such as in the Metasearch and Aquitaine Patrimoine projects

OAI-PMH for digital resources discovery. The development of that model in the cultural heritage sector could be implemented through the following suggestions raised by the Illinois tests<sup>117</sup> to integrate the OAI-PMH discovery architecture in the overall process of digital content creation :

- Build registry of collections with digital content;
- Guide funded projects to make their metadata available with OAI-PMH;
- Build a repository and search & discovery tools for integrated access to the content of those collections;
- Research best practices for sharing metadata about heterogeneous content in various user communities.

The best practices inventory is certainly needed for such an emerging technology, still not very widespread in the cultural heritage sector but rather promising. Forums such as the OAForum, the cultural content Forum<sup>118</sup>, Digicult<sup>119</sup> and the Minerva project<sup>120</sup> shall contribute to exchange and disseminate good practices.

The present report aims at describing interesting good practices, however, many issues are still to be more widely experimented and tools to be developed. The spreading of the OAI-PMH may allow such experiments to be launched and tools and procedures to be exchanged. Notably guidelines for the use of DC for cross-domain cultural heritage metadata and guidelines for crosswalks, schemas for all metadata model only encoded as DTDs, schemas for the use of rights on metadata, EAD splitting or tagging for cross-domain applications (item-level or collection level), information retrieval through a mix of classical search engines and harvested content, cross-language heritage resource discovery, representation of data in various schemas and data re-usability, use of push technologies to launch harvesters, and moreover, user interfaces to heterogeneous and aggregated content and validation of new types of agents, roles, financial models and users in the heritage information environment.

For ensuring a *secure* development of the technology in the cultural heritage sector, based on the experiences of all implementers, programme leaders and institutional stakeholders shall ensure the impact is improved and the metadata and terminologies used are mostly based on international standards.

Indeed, the OAI model also helps disseminating regional and national standards. As an example, the Picture Australia service encourages the use of the Australian pictures thesaurus. It is important that this standardization be related to international standards and avoid any effort to set a local standard between partners of a service where the mapping work will not be re-usable. The standardisation authorities have a clear role to facilitate this evolution.

This is a challenge for funders and institutional stakeholders of digital heritage, to include the OAI model within the digital content creation framework, to anticipate the major initiatives for standardization within the cultural heritage sector and ensure the involvement of major industrial heritage partners to implement OAI repositories for document management applications.

---

<sup>117</sup> Cole Timothy W., 'Using OAI-PMH to aggregate metadata describing cultural heritage resources', presentation ALA/CLA annual meeting, 22 June 2003, Toronto, <<http://dli.grainger.uiuc.edu/Publications/TWCole/ALA2003OAI/>>

<sup>118</sup> <http://www.culturalcontentforum.org> and Miller Paul, Dawson David, Perkins John, "Towards a Digital Cultural Content Forum", in Cultivate Interactive, July 2002, <http://www.cultivate-int.org/issue7/washington/>

<sup>119</sup> <http://www.digicult.info>

<sup>120</sup> <http://www.minervaurope.org/>

## Annex 1 – Glossary

### Dublin Core Metadata Set

"The Dublin Core metadata element set is a standard for cross-domain information resource description."<sup>121</sup> It is composed of 15 elements, namely : Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Source, Identifier, Language, Relation, Coverage, Rights

### aggregated resources

Aggregated resources are gathered from various sources within a single location and services can be built on that basis.

### harvester

"A *harvester* is a client application that issues OAI-PMH requests."<sup>122</sup>

### integrated access

an access node at the user premises that can simultaneously deliver information from various sources

### item

"An *item* is a constituent of a repository from which metadata about a resource can be disseminated."<sup>123</sup>

### memory organisations

Institutions or departments which preserve and disseminate heritage, through a policy and a strategy specifically designed for organisation the "memory". The cultural heritage sector is composed of archives, libraries, museums, monuments, archaeological sites, galleries ...

### original system

internal to the data provider, the one which is already used out of OAI structure. This is the data source, from which the repository takes its content

### pivot schema or terminology

An intermediary standard used to create equivalences between distinct schemas or terminologies used in various systems.

### record

"A record is metadata expressed in a single format. A record is returned in an XML-encoded byte stream in response to an OAI-PMH request for metadata from an item"<sup>124</sup>.

### repository

"A repository is a network accessible server that can process the 6 OAI-PMH requests in the manner described in this document"<sup>125</sup>

### repositories shareable/ disclosed

OAI shareable repositories are built to expose their metadata for being largely harvested. Therefore, they are built to be useful to as many harvesters as possible. Disclosed (or coordinated) repositories are built for the only use of a service and their content defined according that or those service provider(s).

### resource

the cultural heritage object described by metadata. It is represented by an item in the OAI repository

### set

"A *set* is an optional construct for grouping items for the purpose of selective harvesting"<sup>126</sup>

---

<sup>121</sup> <http://www.dublincore.org/documents/dces/>

<sup>122</sup> <http://www.openarchives.org/OAI/openarchivesprotocol.html#harvester>

<sup>123</sup> <http://www.openarchives.org/OAI/openarchivesprotocol.html#Item>

<sup>124</sup> <http://www.openarchives.org/OAI/openarchivesprotocol.html#Record>

<sup>125</sup> <http://www.openarchives.org/OAI/openarchivesprotocol.html#Repository>

<sup>126</sup> <http://www.openarchives.org/OAI/openarchivesprotocol.html#Set>

## Annex 2 – Bibliography

- “A framework of guidance for building good digital collections”, IMLS Digital Library Forum Report, 2001, <http://www.ims.gov/pubs/forumframework.htm>
- “Australian War Memorial Annual report 2000-2001”, [http://www.awm.gov.au/corporate/annual\\_report/ann\\_rep00-01.pdf](http://www.awm.gov.au/corporate/annual_report/ann_rep00-01.pdf)
- “EAD “Index Record” Proposal”, <http://www.columbia.edu/cu/libraries/inside/projects/findingaids/model/>, Columbia Working Documents for the “Greene & Greene Virtual Archive”
- “Progress report of the National Representatives Group: coordination mechanisms for digitisation policies and programmes 2002.”, European Commission : The Information Society Directorate-General, 2003, <http://www.minervaeurope.org/publications/globalreport.htm>
- Ahlborn Benjamin, Nejd Wolfgang, Siberski Wolf, “OAI-P2P : A peer-to-peer network for open archives”, 2002, <http://projekte.learninglab.uni-hannover.de/pub/bscw.cgi/d7694/OAI-P2P%3a%20A%20Peer-to-Peer%20Network%20for%20Open%20Archives>
- Apps Ann, “JISC Information Environment Service Registry (IESR) OAI-PMH Service Input Templates”, 2003, <http://www.mimas.ac.uk/iesr/metadata/templates/svcoai-template.html>
- Arms Caroline, “Implementing the OAI Protocol for Metadata Harvesting at the Library of Congress”, Fall Forum 2002, <http://www.diglib.org/forums/fall2002/oidlf6.htm>
- Baxter Robert, Blomeley Frances, Kemsley Rachel, “The AIM25 Project”, in Ariadne issue 31, March/April 2002, <http://www.ariadne.ac.uk/issue31/aim25/>
- Bennet Nuala, Sandore Beth, Pianfetti Evangeline, “Illinois Digital Cultural Heritage Community – Collaborative Interactions among libraries, museums and elementary schools”, in D-Lib magazine vol 8-1, January 2002, <http://www.dlib.org/dlib/january02/bennett/01bennett.html>
- Bergman K. Michael, “The Deep Web: Surfacing Hidden Value”, The Journal of electronic publishing, University of Michigan Press, 2001, <http://www.press.umich.edu/jep/07-01/bergman.html>
- Bide Mark, “Open Archives and Intellectual Property : incompatible world views”, report for the Open Archives Forum, nov 2002, [http://www.oaforum.org/otherfiles/oaf\\_d42\\_cser1\\_bide.pdf](http://www.oaforum.org/otherfiles/oaf_d42_cser1_bide.pdf)
- Brody Tim, Kampa Simon, Harnad Stevan, Carr Les, Hitchcock Steve, “Digitometric services for Open Archives Environments”, in Koch Taugott, Torvik Solvberg Ingeborg, “Research and Advanced Technology for digital libraries” 7<sup>th</sup> European conference, ECDL 2003, Trondheim Norway, August 17-22, 2003, Berlin 2003
- Campbell Debbie, “The National Library Experiments with Resource Discovery », National Library of Australia Gateways, June 2002, <http://www.nla.gov.au/ntwkpubs/gw/57/p16a01.html>
- Carpenter Leona, Heery Rachel, “OAI and its value to the cultural heritage sector”, in Digicult Info 3, Feb. 2003, [http://www.digicult.info/downloads/digicult\\_info3\\_low.pdf](http://www.digicult.info/downloads/digicult_info3_low.pdf)
- Chanay Daniel, “Centre pour la Communication Scientifique Directe“, dec. 2001 [web.ccr.jussieu.fr/urfist/presse/ccsd\\_charnay.ppt](http://web.ccr.jussieu.fr/urfist/presse/ccsd_charnay.ppt)
- CIMI – “Guide to Best practices Dublin Core DC 1.0 RFC 2413, version 1.1”, 21 April 2000, [http://www.cimi.org/public\\_docs/meta\\_bestprac\\_v1\\_1\\_210400.pdf](http://www.cimi.org/public_docs/meta_bestprac_v1_1_210400.pdf)

- Cliff Pete, "Building ResourceFinder", in Ariadne issue 30, dec. 2001, <http://www.ariadne.ac.uk/issue30/rdn-oai/>
- Cole Timothy W., "Using OAI-PMH to aggregate metadata describing cultural heritage resources", presentation ALA/CLA annual meeting, 22 June 2003, Toronto, <http://dli.grainger.uiuc.edu/Publications/TWCole/ALA2003OAI/>
- Cole Timothy W., Kaczmarek Joanne, Marty Paul F., Prom Christopher J., Sandore Beth, Shreeves Sarah, "Now that we've found the 'hidden Web', what can we do with it?", Museums and the Web 2002 conference, <http://www.archimuse.com/mw2002/papers/cole/colefig1.html>
- Cole Timothy W., "A framework of guidance for building good digital collections", IMLS Digital Library Forum Report, Web-Wise 2002, [http://dli.grainger.uiuc.edu/publications/twcole/WebWise2002/Cole\\_IMLS\\_DLF\\_Framework.ppt](http://dli.grainger.uiuc.edu/publications/twcole/WebWise2002/Cole_IMLS_DLF_Framework.ppt)
- Cole Timothy W., "OAI provider & harvesting services at the University of Illinois", in Joint Conference on Digital Libraries 2001, [http://dli.grainger.uiuc.edu/Publications/TWCole/JCDL2001/JCDL01\\_OAI.ppt](http://dli.grainger.uiuc.edu/Publications/TWCole/JCDL2001/JCDL01_OAI.ppt)
- Cole Tim (guest editor, collective), "Open Archives Initiative metadata harvesting", Library High Tech, vol 21, Number 3, 2003, Emerald
- Davidson Stephen, "The Open Archives Initiative (OAI) Sheet music project - a gateway to sheet music collections on the Web", Sheet Music Roundtable, Music Library Association 2003, Austin, <http://unitproj.library.ucla.edu/music/oaisheetmusic/mla.ppt>
- Dawson David, "Open Archives Initiative, Metadata Harvesting and the NOF portal – an information paper from the NOF technical advisory service", <http://www.ukoln.ac.uk/nof/support/help/papers/oai-pmh>
- Dawson David, "Minerva report 4.01 on interoperability", April 2003
- Dewhurst Basil, "Enabling Interoperability : Australian Museums Online (AMOL) & the Open Archives Initiative Protocol for Metadata Harvesting ", AMOL/CIMI Institute Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) Workshop June 2002, [http://amol.org.au/oai/files/AMOL\\_CIMI\\_OAI-PMH\\_Workshop\\_BD\\_Enabling\\_Interoperability\\_20020618.pdf](http://amol.org.au/oai/files/AMOL_CIMI_OAI-PMH_Workshop_BD_Enabling_Interoperability_20020618.pdf)
- Digital Libraries Federation, "DLF evaluation of the Open Archives Initiative", January 2003, <http://www.diglib.org/architectures/testbed.htm>
- Dobratz Susanne, Matthaei Birgit, "Open Archives Activities and Experiences in Europe - An Overview by the Open Archives Forum", in Dlib magazine, vol. 9, number 1, Jan. 2003, <http://www.dlib.org/dlib/january03/dobratz/01dobratz.html>
- Dobratz Susanne, Schimmelpfennig, Schirnbacher Peter, "The Open Archives Forum", in Ariadne issue 31, March/April 2002, <http://www.ariadne.ac.uk/issue31/open-archives-forum/>
- Dobratz Susanne, Matthaei Birgit, Wang Jing Yuan, Castelli Donatella, Heery Rachel, Carpenter Leona, "Interim review of technical issues", OAF forum deliverable 2.2, [http://www.oaforum.org/otherfiles/oaforum\\_d22\\_technical1.pdf](http://www.oaforum.org/otherfiles/oaforum_d22_technical1.pdf)
- Erway Ricky, "Creating New Knowledge through RLG Cultural Materials", RLG annual meeting, Amsterdam 2002, <http://www.rlg.org/annmtg/erway02.html>
- Foulonneau Muriel, "EMII-DCF report Content Creation Analysis" , 2002
- Foulonneau Muriel, "Le protocole OAI-PMH : une opportunité pour le patrimoine numérique", jan 2002, <http://www.culture.gouv.fr/culture/mrt/numerisation/fr/technique/documents/oai.pdf>

- Freeman Danielle, « PictureAustralia - a collaborative digital project, Images of Australiana – cultural agencies cooperate to bring their pictorial collections together at the one web site”, [http://www.natlib.govt.nz/files/forum/freeman\\_files/frame.htm](http://www.natlib.govt.nz/files/forum/freeman_files/frame.htm)
- Giersch Sarah, Jones Casey, Sumner Tammy, “NSDL Evaluation Pilot - Preliminary Report of Collections Data & Users and Usage Data”, sept 2002, [http://eduimpact.comm.nsdlib.org/doc\\_tracker/docs\\_download.php?id=230](http://eduimpact.comm.nsdlib.org/doc_tracker/docs_download.php?id=230)
- Halbert Martin, Kaczmarek Joanne, Hagedorn Kat, “Findings from the Mellon Metadata Harvesting initiative”, in Koch Taugott, Torvik Solvberg Ingeborg, “Research and Advanced Technology for digital libraries” 7<sup>th</sup> European conference, ECDL 2003, Trondheim Norway, August 17-22, 2003, Berlin 2003
- Johnston Pete, Dawson David “Collections et services : construire un environnement informationnel pour l’Europe” in Culture & Recherche, Paris, 2002, <http://www.culture.gouv.fr/culture/doc/index.html>
- Kim Hyunki, Choo Chee-Yoong, Chen Su-Shing, “An integrated digital library server with OAI and self-organizing capabilities”, in Koch Taugott, Torvik Solvberg Ingeborg, “Research and Advanced Technology for digital libraries” 7<sup>th</sup> European conference, ECDL 2003, Trondheim Norway, August 17-22, 2003, Berlin 2003
- Liu Xiaoming, Maly Kurt, Zubair Mohammad, “Arc - An OAI Service Provider for Digital Library Federation”, in D-Lib Magazine, Volume 7 Number 4, April 2001, <http://www.dlib.org/dlib/april01/liu/04liu.html>
- Liu Xiaoming, Maly Kurt, Zubair Mohammad, Qiaoling Hong, Michael L. Nelson\*, Frances Knudson\*\* and Irma Holtkamp, “Federated Searching Interface Techniques for Heterogeneous OAI Repositories”, Journal of Digital information, volume 2 issue 4, May 2002, <http://jodi.ecs.soton.ac.uk/Articles/v02/i04/Liu/>
- Lusso Salvatore, “OpenHeritage: Developing cultural tourism in lesser-known regions”, in Cultivate Interactive issue 9, Feb. 2003, <http://www.cultivate-int.org/issue9/openheritage/>
- MacKenzie Georges, Kristiansson Göran, “How Real Archivists can learn to love the OAI - A review of the potential for using the Open Archives Initiative Metadata Harvesting Protocol in conventional archives”, OAForum community-specific report, 2002
- Maly K., Zubair M., Nelson M., Liu X., Anan H., Gao J., Tang J., Zhao Y., “Archon – a digital library that federates physics collections”,
- Mayer Constance, Munstedt Peter, Dunn Jon, “Sharing digital scores : will the Open Archives Initiative Protocol for Metadata Harvesting provide the key?”, Feb 2003, <http://www.dlib.indiana.edu/~jwd/scores-mla2003.ppt>
- Miller Paul, Dawson David, Perkins John, “Towards a Digital Cultural Content Forum”, in Cultivate Interactive, July 2002, <http://www.cultivate-int.org/issue7/washington/>
- Mulrenin Andrea, “The DigiCULT Report Technological Landscapes for Tomorrow’s Cultural Economy Unlocking the Value of Cultural Heritage”, European Communities, 2002, [ftp://ftp.cordis.lu/pub/ist/docs/digicult/executive\\_summary\\_en.pdf](ftp://ftp.cordis.lu/pub/ist/docs/digicult/executive_summary_en.pdf)
- National Archives of Australia, “Annual report 2001”, [http://www.naa.gov.au/publications/corporate\\_publications/ar2001/outcomes\\_outputs\\_reports2.html](http://www.naa.gov.au/publications/corporate_publications/ar2001/outcomes_outputs_reports2.html)
- Nelson Michael L., “U.S. Government Use of the OAI-PMH”, ISTE/NSF Ibero-American digital library Joint Project development symposium, Brazil, 2003, <http://www.cs.odu.edu/~mln/pubs/oai-campinas/nelson.ppt>

- Perkins John, "Disclosing Digital Cultural Wealth: Museums and the Open Archives Initiative", in *Cultivate Interactive* issue 6, Feb 2002, <http://www.cultivate-int.org/issue6/cimi/>
- Perkins John, "A new way of making cultural information resources visible on the Web: museums and the open archives initiative", museums and the Web 2001 papers, <http://www.archimuse.com/mw2001/papers/perkins/perkins.html>
- Plante Ray, "An evaluation of the Open archives initiative for VO registries", 2003, <http://www.ivoa.net/internal/IVOA/RegistryRequirements/evaloai.html>
- Powel Andy, "An OAI approach to sharing subject gateway content", Tenth International World Wide Web Conference, May 1-5, 2001, <http://www.www10.org/cdrom/posters/1097.pdf>
- Prom Christopher J., Habing Thomas G., "Using the Open Archives Initiative Protocols with EAD", Joint Conference on Digital Libraries 2002, <http://dli.grainger.uiuc.edu/publications/jcdl2002/p14-prom.pdf>
- Richardson Steve, Powell Andy, "Exposing information resources for e-learning - Harvesting and searching IMS metadata using the OAI Protocol for Metadata Harvesting and Z39.50", in *Ariadne* issue 34 dec 2002/Jan 2003, <http://www.ariadne.ac.uk/issue34/powell/>
- Saylor John M., "Draft – NSDL collection development policy draft v030715", July 2003, [http://content.comm.nsdlib.org/doc\\_tracker/docs\\_download.php?id=452](http://content.comm.nsdlib.org/doc_tracker/docs_download.php?id=452)
- Sévigny Martin, Bourgoüin Christine, "OAI (Open Archives Initiative)", June 2003 <http://www.ajlsm.com/projets/pp/technique/oai.html>
- Sfakakis Michalis, Kapidakis Sarantos, "An architecture for online information integration on concurrent resource access on a Z39.50 environment" in Koch Taugott, Torvik Solvberg Ingeborg, "Research and Advanced Technology for digital libraries" 7<sup>th</sup> European conference, ECDL 2003, Trondheim Norway, August 17-22, 2003, Berlin 2003
- Shreeves Sarah L., Kirkham Christine, Kaczmarek Joanne, Cole Timothy W., "Utility of an OAI service provider search portal", Joint Conference on Digital Libraries 2003, May 2003, <http://dli.grainger.uiuc.edu/Publications/JCDL2003/ShreevesJCDL.ppt>
- Straccia Umberto, "CYCLADES An Open Collaborative Virtual Archive Environment", EC/NSF Digital Library All Projects Meeting, March 2002, Rome, <http://delos-noe.iei.pi.cnr.it/activities/internationalforum/All-Projects/RomeSlides/CYCLADES.pdf>
- Suleman Hussein, "Building interoperable digital libraries : a practical guide to creating open archives", JCDL conference 2002, [http://www.dlib.vt.edu/projects/OAI/reports/jcdl\\_2002\\_tutorial\\_oai\\_slides.pdf](http://www.dlib.vt.edu/projects/OAI/reports/jcdl_2002_tutorial_oai_slides.pdf)
- Suleman Hussein, Fow Edward A., "Beyond harvesting : digital library components as OAI extensions", 2002, [http://oai.dlib.vt.edu/odl/pubs/cstr\\_2002\\_odl\\_1.pdf](http://oai.dlib.vt.edu/odl/pubs/cstr_2002_odl_1.pdf)
- Van de Sompel Herbert, Lagoze Carl, "Notes from the interoperability front : a progress report on the Open Archives Initiative", ECDL 2002 submission draft, <http://www.openarchives.org/documents/ecdl-oai.pdf>
- Van de Sompel Herbert, Young Jeffrey A., Hickey Thomas B., "Using the OAI-PMH... differently", in *D-Lib Magazine* July/August 2003, vol 9, number 7/8, <http://www.dlib.org/dlib/july03/young/07young.html>
- Van Veen Theo, "The European Library:opportunities for new services", Oaforum workshop "Open Access to Hidden Resources", 2002, [http://www.oaforum.org/otherfiles/lisb\\_tel.ppt](http://www.oaforum.org/otherfiles/lisb_tel.ppt)
- Waters Donald J., "The metadata harvesting initiative of the Mellon Foundation", in *ARL bimonthly report* 217, August 2001, <http://www.arl.org/newsltr/217/waters.html>

### Annex 3 – Tricks, traps and issues

	Tricks	Traps	Issues
<b>Data provision</b>			
<b>Budget issues</b>	<ul style="list-style-type: none"> <li>▪ consider into details the terminology and/or metadata mapping necessary</li> <li>▪ take into account possible restrictions and difficulties of the technical architecture</li> </ul>	Do not measure the necessary documentary effort	Costs efforts for building a repository, for metadata and technical implementation
<b>Communicating on a repository</b>	<ul style="list-style-type: none"> <li>▪ towards users, communication strategy on improved access (access to related resources from other institutions)</li> <li>▪ towards professionals, communication strategy on institution openness</li> </ul>	Send users to services' Website or focus on the technology issue	How to publish information on OAI compliance, the issue is different whether it is shareable or disclosed repository
<b>Datasets definition</b>	<ul style="list-style-type: none"> <li>▪ define OAI sets according to harvesters if disclosed repository</li> <li>▪ otherwise, many examples of sets by subject and by document type</li> <li>▪ publish the datasets in human readable format</li> </ul>	Difficulty to determine the opportunity to build datasets and avoid spending much work for collection level description.	Define OAI sets for harvesters to select what they need to harvest.
<b>Interoperability issues</b>			
<b>Hierarchical information</b>	<ul style="list-style-type: none"> <li>▪ use the DC:relation element for example with standard qualifiers HasPart and IsPartOf</li> <li>▪ the service can use a system to rebuild hierarchies, but this may not be useful if only for resource discovery</li> <li>▪ Clearly identify the level of information (for example in the LEVEL element of EAD or the DC:type element), with a comprehensive terminology</li> <li>▪ Use standard links (Xpointers, OpenURL)</li> </ul>	Lose meaning if do not re-build the full structure.	Hierarchical information, notably with EAD, to be handled on service-side.

<b>Terminology empowerment</b>	<ul style="list-style-type: none"> <li>▪ Possibility to “normalize” terminologies on service side, by relating elements to the terminology.</li> <li>▪ The service may re-build the relation between values and the terminology. Otherwise, the full path of a thesaurus shall be recorded as the value, eg. “Nicaragua – History – Flibuster War”.</li> <li>▪ Terminology search tools can be implemented in user interface.</li> </ul>	Empowerishment of search tools and absence of any possible multilingual search functionality	Necessity to map terminologies when aggregating heterogeneous content
<b>Granularity</b>	Defining what it is : a collection, an image, a Website, a finding aid <sup>127</sup>	irrelevant results : not comprehensible what I can find and what I have in front	Mixing items, collections and finding aids
<b>Identifiers</b>	Use of persistent identifiers for resources. PURL implementation for records to provide “cool URLs”	URL may change for resource	Issue of persistent identifiers
<b>Metadata Schema</b>			
<b>Valid XML</b>	<ul style="list-style-type: none"> <li>▪ implement quality check on responses delivered</li> <li>▪ quality check on repository when declaring repository to Open Archives</li> <li>▪ aggregators improve the quality of the data delivered as demonstrated by the ARC experience</li> </ul>	<p>Risk that the service provider is unable to handle the record since this XML is an OAI requirement.</p> <p>This issue is especially important for various languages encoding in XML.</p>	Not all data providers actually provide valid XML files
<b>Create a schema</b>	<ul style="list-style-type: none"> <li>▪ The standardization process under the form of XSD is to be encouraged and/or the OAI specification may not be respected for several metadata standards to be used.</li> <li>▪ The standardization authorities can maintain XSD permanently accessible but for service-specific ones, this can be a challenge if the OAI records validity is checked on a regular basis according to the declared schema.</li> </ul>	The schema may not be persistently accessible, not sufficiently accepted within the community, not clearly defined for the community, not existing for a given metadata standard, notably the EAD DTD.	<p>Location and sustainability of the schema. The schema is a requirement and must be persistently accessible.</p> <p>All metadata standards are represented by schemas.</p>
<b>Mapping to OAI DC</b>	<ul style="list-style-type: none"> <li>▪ Guidelines, including or excluding terminology values</li> <li>▪ only valuable for heterogeneous / cross-domain content, otherwise, better use richer system</li> <li>▪ The XOAI-PMH does include the DC requirement, in any case DC :identifier and DC :title is enough to be DC compliant</li> </ul>	<p>DC elements differently filled according to the type of resources : eg. Only 10% of resources have filled the DC:subject, then this does not mean anything for searching.</p> <p>Inconsistency of DC mapping in several cases.</p> <p>Absence of standard crosswalks.</p>	DC may be used as the highest common denominator for cross-domain content. It is a requirement of OAI 2.0 specifications but it is not meaningful for all types of content.

<sup>127</sup> see RLG which always defines the granularity level

<b>Mapping to an Xschema schema</b>	<ul style="list-style-type: none"> <li>▪ Necessity to map to standard XML schemas as far as possible.</li> <li>▪ Existing mapping procedures including guidelines.</li> <li>▪ Define new mapping guidelines according to the guidelines of the schema.</li> </ul>	From the database structure, it is necessary to extract XML records. This may entail a heavy work.	Usually catalogues are not conceived as XML records.
<b>Modifying the common pivot schema</b>	Compatibility of the new schema with previous versions.	Records built according to old version rejected.	Create a new version of the schema.
<b>Professional schema</b>	<ul style="list-style-type: none"> <li>▪ For community specific services or for display. Community-specific simplified schema can be used (eg. DC Library).</li> <li>▪ Otherwise, full community-specific records can be exchanged, with the condition of the creation of a schema if not existing</li> </ul>	Information loss if only using DC. DC mapping is sometimes irrelevant.	DC not enough to provide full information on resources : possibilities to use EAD, MARC-XML, CIMI-DTD/SPECTRUM-XML
<b>High quality service provision</b>			
<b>Building portals</b>	<ul style="list-style-type: none"> <li>▪ A portal can mix cross-searching (Z39.50) and harvesting (OAI-PMH).</li> <li>▪ An aggregator can collect metadata through cross-searching and expose them for harvest.</li> </ul>	Not all data providers are OAI-compliant, some have already implemented Z39.50 gateway and may not understand the need to implement a new protocol.	When necessity of cross-searching or existing Z39.50 server, is necessary to build an OAI repository?
<b>Limited investment and competences in a memory organisation</b>	<ul style="list-style-type: none"> <li>▪ Aggregators set up at institutional level or through great institutions (National Library, National Archives, Regional aggregators) to expose data extracted from small institutions systems.</li> <li>▪ Great institutions, funding programmes and government organisations should ensure that heritage treasures held in small institutions can be inventories and harvested.</li> </ul>	Small institutions not able to implement repositories.	A memory organisation may not be or feel able to lead an OAI repository project.
<b>Danger of misuse of metadata</b>	<ul style="list-style-type: none"> <li>▪ Defining an IPR policy and possibly add an HTTP-based protection with login/password to control harvesters</li> <li>▪ Defining a terminology for rights on metadata harvested and usage [see ROMEO project]</li> <li>▪ Controlling harvesters : recording which one is doing what</li> <li>▪ Use a retrieval-oriented metadata set (DC Culture) which contains intellectually impoverished description for retrieval only.</li> </ul>	The memory organisation when exposing its metadata to a harvester, may not abandon the right to agree on the way in which the service provider uses records.	Metadata creation lead to copyright, moral rights on records as they are the memory organisations' intellectual asset.  Copyright may also exist on terminologies (thesaurus).  A repository is also a database and it may be protected by copyright.

<p align="center"><b>Publishing information from data providers</b></p>	<p>Displaying information on data sources in a specific section and referring records to it :</p> <ul style="list-style-type: none"> <li>▪ Repositories harvested;</li> <li>▪ Frequency of data update in repository and harvesting;</li> <li>▪ When giving a result, propose the link to the original catalogue;</li> <li>▪ data branding (data provider's logo).</li> </ul>	<p>User may not understand the difference between the responsibilities of the service providers and the ones of the data providers. User may not be interested in knowing where the resource comes from as long as he/she successfully accesses it (DCHC project evaluation. However, he/she may be interested in knowing about reliability of information and if/ that, by going to the repository Website (if any), he/she may find additional information.</p>	<p>Necessity to mention clearly record's provenance for the user.</p>	
<p align="center"><b>Representing the resource</b></p>	<ul style="list-style-type: none"> <li>▪ Collecting thumbnails but providing access to full resources on the original system's Website.</li> <li>▪ Textual resources may be transferred for information retrieval (as an access point to the resource).</li> </ul>	<p>Need to leave benefit to the data provider (hits) but if displaying various resources in the same page from various Websites, the time lag and chance to have a server failure are multiplied.</p>	<p>It is technically possible to the resource together with the metadata. Alternatively an extract of the resource may be transferred and displayed on the service's Website.</p>	
<p align="center"><b>Integrated access to heterogeneous content metadata</b></p>	<p>Build indexes according to resources types, rather than provenance.</p>	<p>Access centralised rather than integrated, differences in data processing kept in service after initial aggregation procedure.</p>	<p>Providing integrated access to content from various provenances.</p>	
<p align="center"><b>Users have a single entry point to aggregated resources</b></p>	<ul style="list-style-type: none"> <li>▪ Branding descriptions of resources or clearly mentioning to the user that access is granted to external Website</li> <li>▪ Setting a follow up service including participation of data providers' community</li> <li>▪ Clearly publish the service's IPR policy, selection policy, terms of use .... Agreed by data providers</li> </ul>	<p>User will send questions to service provider.</p>	<p>Especially if presenting results in service's frame, user will consider the resource comes from the service provider.</p>	
<b>Community issues</b>				
<p align="center"><b>Maintaining a community</b></p>	<ul style="list-style-type: none"> <li>▪ an email discussion list?</li> <li>▪ real meetings?</li> </ul> <p>See Picture Australia</p>	<p>Participation of data providers may be ensured for repositories updates, common management of problems...</p>	<p>Service shall be based on a community to include common rules ...</p>	
<p align="center"><b>New data providers</b></p>	<ul style="list-style-type: none"> <li>▪ Clear, simple and formalised procedure, including metadata quality issues, mapping guidelines .... for consistent practice of metadata authoring</li> <li>▪ Use of a charter</li> </ul>	<p>Too many constraints to get harvested</p>	<p>Be harvested</p>	

<b>Architecture</b>			
<b>Repositories updates</b>	<ul style="list-style-type: none"> <li>▪ Publish frequency of data update in repository and harvesting, not apply OAI-PMH if too many and frequent updates.</li> <li>▪ For catalogues, repositories should be updated either once a month, or to include new records</li> <li>▪ Maintain profiles according to data providers to take preferences and features (security?) into account</li> <li>▪ Solutions are considered to implement push signal when repository is updated (Web services)</li> </ul>	Dead links and unreliable services	<ul style="list-style-type: none"> <li>▪ Cultural heritage repositories may not be updated very often</li> <li>▪ Harvesting frequency depends on the information contained.</li> <li>▪ However, as long as we are dealing with cultural heritage catalogues, few editing is expected, numerous new records.</li> </ul>
<b>Mapping and re-processing issues</b>	<ul style="list-style-type: none"> <li>▪ Do not map terminology and metadata at repository level, but rather on service-side</li> <li>▪ Quality control on XML and / or data formatting can be done at repository level, since this increases the quality of data without entailing meaning loss.</li> </ul>	Data may be re-used and if mapping metadata content initially, another service will not harvest the original record, but rather a record already mapped, which will therefore have lost part of its interpretive value	Service-side or repository-side, when building a strict service based on a strict community.
<b>Harvesting overloads</b>	<ul style="list-style-type: none"> <li>▪ Build OAI sets on data provider side</li> <li>▪ Divide transfers</li> <li>▪ Compress transfers</li> </ul>	Too many records according to service's needs. Inability to query a repository.	The repository contains lots of records and the service does not need all of them.
<b>Servers' overload</b>	<ul style="list-style-type: none"> <li>▪ Build non integrated repositories (different from the original system)</li> <li>▪ First harvest out of harvesting protocol (manual transfer)</li> <li>▪ Harvests only through an OAI Caching system such as Celestial</li> </ul>	Repository server overload when collecting data.	The harvesting task requires too much resources from the data provider's server because of the amount of data to generate and transfer.
<b>Service search functionalities</b>			
<b>Multilingual information</b>	<ul style="list-style-type: none"> <li>▪ the language of the metadata values must be recorded in a machine-readable way.</li> <li>▪ metadata in the schema or at collection level and build multilingual information retrieval functionalities on the service side.</li> <li>▪ for each element of the schema, defining the language of value.</li> </ul>	<p>If no specific processing is implemented, search functionalities on content are not efficient.</p> <p>If specific processing are implemented (reference to a multilingual thesaurus for example), the system still needs to know the language of the metadata value.</p>	Metadata values in various languages, this is especially important to build European-wide services.

<b>Harvesting frequency</b>	<ul style="list-style-type: none"> <li>▪ Harvesting frequency must be synchronised with repository update frequency. Once a month for cultural heritage resources are commonly accepted.</li> <li>▪ The service can manage harvesting profiles for each data provider.</li> <li>▪ Possibility to set a mix push/pull mechanisms to launch harvests.</li> </ul>	Information available on service side is not reliable and the user may better search information on data provider's Website or the information available becomes irrelevant since not up-to-date.	Data update is a core issue for an OAI architecture. The service provider's reliability depends on harvesting frequency and repositories update frequency.
<b>Number results</b>	<ul style="list-style-type: none"> <li>▪ Refine search functionalities</li> <li>▪ Clear selection of collections harvested</li> <li>▪ User profiles</li> </ul>	Too many unranked results, user does not know how to handle them.	Mixing heterogeneous content from various specialised catalogues may lead to a very large amount of results and a few users actually use advanced search.
<b>Open access to metadata on physical resources</b>	<ul style="list-style-type: none"> <li>▪ not disappoint user, explicitly mention the type of resource which is accessed through the link ("view" / "learn more" or "bibliographic record" / "image" / "sound").</li> <li>▪ possibility to exclude analog only resources from results.</li> </ul>	Open access means something different to user and to professional (see Tim Cole's slides)	Open access to physical only resources means that the end-user will only access bibliographic files or finding aids.
<b>Access points for heterogeneous content</b>	<ul style="list-style-type: none"> <li>▪ Comprehensive interfaces, according to user search scenarios</li> <li>▪ Few access points and limit terminology values</li> <li>▪ DC culture with CIMI high level access points, or simple DC-based search interfaces</li> </ul>	Complex structure and interface, difficult to understand for user.	Accessing heterogeneous content, created in distinct institutions, according to different sectoral rules.
<b>Digital content creation</b>			
<b>Facilitating new value-added services</b>	<ul style="list-style-type: none"> <li>▪ build registries of collections created</li> <li>▪ assist and support availability of item-level metadata</li> <li>▪ build tools for open access to digital content created</li> </ul> <p>=&gt; This is a political issue on overall policy of digital content creation</p>	Content is heterogeneous content and integrated access very difficult to implement in a comprehensive way.	Necessity to Improve access to all digital resources created.

