Assessing Open Archive OAI-PMH implementations

Emanuele Bellini, Marcel Aime Deussom, Paolo Nesi
Distributed Systems and Internet Technology Laboratory
University of Florence, Department of Systems and Informatics
Http://www.disit.dsi.unifi.it
ebellini@dsi.unifi.it , nesi@dsi.unfi.it

Abstract

The number of works made freely accessible on the Web by institutions joined to Open Archive Initiatives such as libraries, research institutions, scientific and cultural archives, is constantly growing. Until now some international agreements among key players of Open Access approach have been established but the low level of standardization and the lack of a digital resource and repository certification authority prevents any quality control on the works and subsequently their trustable dissemination within the cultural and scientific community. In this scenario, the benefit of having freely accessible resources is lost. The study proposed in this paper aims at providing a statistical analysis of the state of art of OA implementations in order to outline weak practices, propose assessment metrics for self analysis and stimulate a general improvement of the quality of OA implementations.

1 Introduction

The Web has drastically changed the information environment where users of the humanities work and study and the information needs to be more accessible to become collective knowledge. To this end, following the initiatives as Berlin Declaration on Open Access, OA, to Knowledge in the Sciences and Humanities [12], a lot of institutions are implementing open access institutional archives. The main mission of the OA is the dissemination of the digital resources and their metadata. In order to address this aim, a number of software tools is available for implementing an open institutional repository such as dspace [8], Fedora [9], eprints [10], etc. To this end, the Open Archive Initiative has implemented the OAI-PMH protocol [1] for publishing and thus making possible the metadata harvesting among repositories. The OAI architecture identifies two logical roles: "Data Providers" and "Service Providers". Data Providers deal with both the deposit and the publication of resources in a repository they "expose" to provide the metadata about resources in the repository. They are the creators and keepers of the metadata and repositories of corresponding resources (digital items, digital essences, which are the effective files). Service Providers use the OAI-PMH interfaces of the Service Providers to collect and store their metadata as shown in [19] and [25]. They use the collected metadata for the purpose of providing one or more services across all the collected metadata like Pleiadi [20], Citeseer [21]. The types of services which may be offered include a query/search interface, peer-review system, cross linking, etc. Recently, an architectural shift was to move away from only human supporting end-user interfaces for each repository, in favour of both human end-user interfaces and machine interfaces for data collecting.

In this paper, the objective is highlight the features needed to evaluate the effectiveness of data provider implementations in terms of OAI-PMH. The quality of a Data Provider implementation and collected metadata can be evaluated in terms of: completeness and accuracy of metadata description, trust-ability of service, level of standard compliance and adoption, etc.

Therefore, the main goals of this paper are:

- To propose and provide a set of metrics to be used as statistical tool for assessing and monitoring the OAI-PMH service implementations in terms of quality, trustworthiness and standard compliance.
- To achieve a better comprehension on Open Access implementation weakness, stimulating and directing new efforts towards technology, policy, standardization levels since the usage of the current widespread solutions is too vague to be exploited at a reasonable cost in the open world.

In order to assess Open Archive implementations in terms of OAI-PMH services, a preliminary harvesting of a large number of OAs has been performed. This allowed the identification of a number of issues and common drawbacks about these services which could be praiseworthy to discuss about. They are summarised as:

- huge number of OA responding to OAI-PMH protocol, more than 2000. The experiment is based on crawling archives and repositories listed at: http://www.openarchives.org/Register/ListFriends [11];
- a number of different metadata sets for each items in each repository, not always the same, frequently not the same for the whole repository. The used metadata are not 100% compliant with those standards;
- a huge number of total records to be collected (the union of records of all the repositories for all the metadata and formats), more than 18 millions.
- a slow mechanism/protocol to access OAI-PMH repositories since servers and connections are slow, and the protocol itself is of high level.

To this end, in order to evaluate the typical Open Archive implementation with respect to the level of spread, metadata quality and service reliability, it has been necessary to perform a general harvesting and processing of the metadata collected from the data providers.

2 GRID Approach for Harvesting

The solution approach is based on OAI-PMH protocol, a REST-based full Web Service that exploits the HTTP protocol to communicate among computers, using either the GET or the POST methods for sending requests. It is well-known that web services are also a computing technique for systematically disseminating XML contents, but when the global amount of data increases, some problems come out. In the following paragraph, we outlined these issues requiring a parallelization approach to be tackled.

2.1 Harvesting from a huge number of repositories issues

According to OAI-PMH protocol, *Guidelines for Harvesting Implements* [24] and OA implementation tutorial [23], a client may put a request to OAI server to ask for the stored content descriptors. Answers are related to the accessible records, and adopted formats. The OAI-PMH protocol provides a list of discrete entities (metadata records) by XML stream. In many cases, these lists may be large and it may be practical to partition them among a series of requests and responses. In fact, the repository replies to a *list* request with an incomplete list and a resumption Token. In order to get responses as much as possible from the list of the OAs considered, the harvester has been performed more requests with resumption Token as arguments. The complete list then consists of the concatenation of the incomplete lists from the sequence of requests, known as a *list request sequence* [1].
Moreover, in the current version of the OAI-PMH protocol a 'verb' to obtain the number of the records that we are going to harvest is not defined. Thus it is impossible to estimate a priori the duration of the process in terms of counted metadata sets. It is clear that the number of records included in a incomplete list (or page) affects the harvesting performance. In some cases this number was only one, and yet the harvester had to perform requests as many as the records in the archive. The harvesting performance also depends on response delay that is related to the network bandwidth and machine performance used by the connected Open Archive. In some cases, this time was greater than 15s for each request. In order to cope with the complexity, a parallel solution has been set up and used as described in the next subsection.

2.2 GRID based metadata harvesting architecture

As it occurs with a web crawler, the harvester contacts and inspects the OA data providers automatically and it extracts metadata sets associated with digital objects via OAI-PMH protocol. Because of the computational weight of these processes, the harvester has been implemented by using the grid based parallel processing on DISIT cloud computing infrastructure. The grid solution has been realized by using AXMEDIS Content Processing (AXCP GRID) [3]. The computational solution has been implemented by realizing a parallel processing algorithm written in AXCP Extended JavaScript [17]. The algorithm has been allocated as a set of periodic processes replicated on a number of grid nodes, typically from 1 to 15 max. The process is managed by the AXCP Scheduler. It is possible to put in execution a number of rules that are distributed to the available grid nodes. Each rule can be periodically (or on demand) scheduled with an interval, for instance, of 1 minute from each running on a single node and the successive one. Each rule is a 'harvester' executor of an OAI-PMH request to obtain the metadata records, parsing the XML response and storing information in our local database called CHonline. In Figure 1, a schema of the architecture is shown.

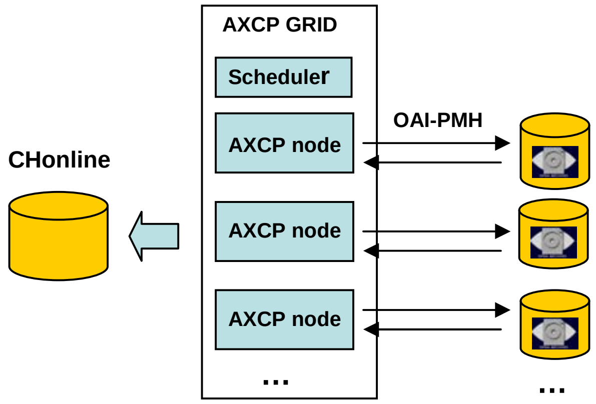

Figure 1 – Grid architecture for massive OA harvesting via OAI-PMH protocol on grid infrastructure of DISIT.

Each GRID node executes an identical autonomous harvesting rule that collects metadata from an Open Archive and populates the database according to the general status also collected into the database. This solution reduces the computational time up to a factor equal to the number of nodes used for completing the harvesting of repositories. In effect, the parallel solution is not only an advantage for the speed up, but also for the reduction of the time needed to get a new global version of the metadata collected in the OI repositories.

2.3 GRID-based harvesting workflow

This paragraph describes the grid base harvesting algorithm and workflow. Figure 2 shows a schema representing the consecutive steps performed by the harvesting rules on the grid.
Before performing the effective harvesting of the single records, two preparatory steps are needed: (i) to get the repositories information; (ii)_to get the metadata sets available for each repository. These two steps are performed into the grid with specific aperiodic/on-demand rules.
During the *first step* a rule for getting the repository list from http://www.openarchives.org/Register/ListFriends website is launched. This rule parses the XML list of OA repositories baseURLs and populates the repository table

of CHonline database. For example, a segment of the repository list is as follows:

[…]

<baseURL id="UOV.es">http://www.tdr.cesca.es/TDR_UOV/NDLTD-OAI/oai.pl</baseURL>

<baseURL>http://diglib.cib.unibo.it/oai/oai2.php</baseURL>

<baseURL>http://docinsa.insa-lyon.fr/oai/oai2.php</baseURL>

[…]

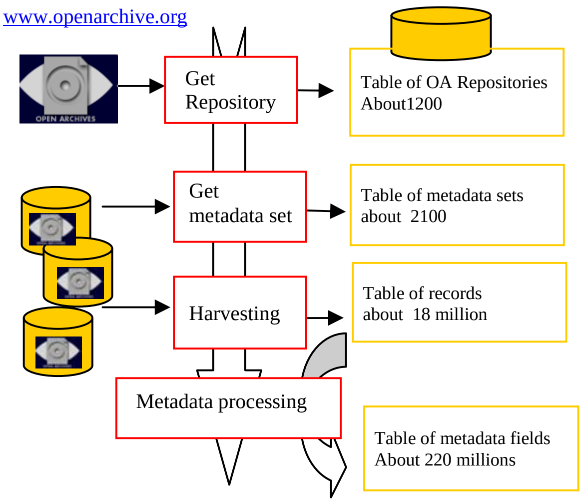

Figure 2 – Algorithm for harvesting

The repositories are identified with a <baseURL> field filled in with the URL of repository OAI-PMH interface and the repository ID. This rule may be also periodically scheduled for checking the availability of new repositories added to the list that have to be harvested by the system.

Once the repository list is obtained, the *second step* has to determine if the OA is active and which service may provide. To this end, a dedicated second rule is activated to both verify the activity of the OA and retrieve the metadata formats available by using the *ListMetadataFormats* verb of OAI-PMH. A repository is set 'not available' if it does not provide any response, so that at the next round it can be tried again. In fact, it may happen that a repository may be offline for some reasons. Therefore, each single OA provides the list of metadata according to the following example.

```
http://baseURL/request?verb=ListMetadataFormat
[
<ListMetadataFormats>
<metadataFormat>
<metadataPrefix>oai_dc</metadataPrefix>
<schema> http://www.openarchives.org/OAI/2.0/oai_dc.xsd
</schema>
<metadataNamespace>http://www.openarchives.org/OAI/2.0/o
ai_dc/</metadataNamespace>
</metadataFormat>
. . . . . . .
</ListMetadataFormats>
```

The list of metadata sets of each repository is stored in the metadata formats table of CHonline database. It should be noted that if a metadata format is declared as being supported by an OA, this does not mean that it is available for all the items in the repository.

OA Harvesting. The harvesting rule gets access to the status table in CHonline database to obtain the first not processed archive/metadata-set and it starts with its crawling. Moreover, the harvesting rule parses the XML response, it extracts only the metadata information and it saves it in a single database field/chunk as a string. The harvesting rule is designed to harvest the records only from one repository managing the resumption Token. This approach is meant to reduce the rule time activity, but there are some cases where a rule could stay alive for hours (for instance if there are a lot of records to harvest and the OAI request has provided a short number of records)

Metadata processing. The metadata harvesting is the first step to collect data and per se it not sufficient to evaluate the quality of metadata implementation. In fact, it is not possible to extract specific metadata values that are related to a specific argument. Moreover the high number of implemented different metadata sets requires a tool for processing them in order to get the single metadata element. In the preliminary analysis, the harvester was tuned to collect at most 100 records for each metadata format managed by each OA. The preliminary process has been used to tune the processing according to the large number of different metadata sets.

Moreover, an additional grid rule got the XML of each non processed record stored in the database and it extracted the single fields. Therefore, each field of each specific record has been stored with its value, type, and additional information in the CHonline database. This poses the basis to perform a deeper analysis, as described in the following. This process led to a sort of an extended RDF [23] model and thus to a metadata normalization allowing queries on the single fields. This table turned out to be very huge (for each field of each metadata record a detailed field record is generated. For instance 15 new records are generated from a single DC based metadata record). The resulting table of single fields (of about 220 millions of records) has been mainly used as a qualitative metadata evaluation for the purpose of this paper. In particular, in this paper we are presenting a metric analysis on OA with some comments about the usage of some metadata fields, always aiming at improving the general understanding and quality of the OA.

3 OA Analysis via OAI-PMH

According to the above described solution for massive OA inspection and metadata harvesting, a set of metrics and considerations has been performed. They may be used to evaluate the implementation of OA as an effective tool for

disseminating scientific works via OAI-PMH service protocol. The performed analysis has been focused on understanding problems, practise and weakness at both quantitative and qualitative levels.
The proposed evaluation model has been developed and validated at three levels:

- **General Metrics**, a quantitative evaluation outlining a global view of the diffusion and effective available OA PMH implementation services;
- **Archive Metrics**, quality evaluation related to the usage and trustability of adoption of metadata sets;
- **Metadata Record Metrics** modeled and provided according to the adopted standard, assessing the usage of single metadata fields and their potential weakness.

In the following paragraphs, a selection of the adopted metrics for each level of analysis is provided to evaluate the current worldwide OA PMH services implementation. The proposed metrics are presented together with the values obtained on previously described tests and configurations.

3.1 General Metrics

At the general level a set of considerations has to be performed, which has been identified by the following proposed metrics.

OA available for harvesting. Sometimes the OA are not available for many different reasons but still remain registered on official/unofficial registers with their baseURL. This index is calculated on the basis of the number of answers by repository performing an OAI-PMH *listmetadataformat* request. In our experiment which was based on the general list of OA, only the 74% of them have an active OAI-PMH service. Generally, many institutions begin with a trial OA project within library or archive institutional department with no cost evaluation, usage training, institutional awareness, etc., and most of them run the risk of giving up soon, since the effective commitment was unexpected or underestimated.
OA with working metadata sets. This criterion has allowed evaluating the effective availability of a repository in providing metadata sets. In the general analysis we registered that about the 10% of the OAs have almost an error on a metadata set. The error can be caused by an XML not well-formed or by content not found. A more detailed statistical evaluation about the identified errors and problem needs should take more space to the provided in this short paper.
Metadata sets in the OA community
This criterion highlights the level of fragmentation of the present OA metadata implementations. There are several standards promoted by different communities to describe resources managed by OA, i.e., different metadata standards. The most common are: Dublin Core [6] (generally supported by default), METS [5], MPEG21 DIDL [4] (as a wrapper of other metadata models), etc. It is also well known that metadata sets may be different for different domains, cultural background, etc. For instance, metadata required to catalogue physics resources can be different with respect to those used for media or ICT works, and again different from those adopted for administrative institutional documents, etc. This lack of uniformity has generated several different standards and again for each of them, several different implementations and/or personalization of metadata sets. In our analysis, 153 different metadata schemas have been identified, over only 853 repositories; thus a high percentage. This count aggregates the records on the metadata Schema field (that is mandatory) from metadata format table.
Metadata sets managed by single OA
Typically, an OA provides its records via OAI-PMH service with more than one metadata set. The possibility for the user to select and harvest records with metadata set more structured and richer with respect to simple DC allows an easier metadata processing. For this reason, a systematic analysis to understand better how this possibility has been perceived and exploited, has been performed. In Figure 3, the distribution of the number of metadata set is reported. It is self-evident that most of the OAs prefer providing only one metadata sets, while the usage of two metadata sets is not unusual.

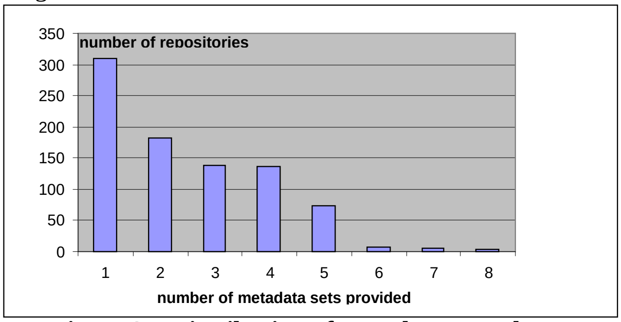

Figure 3 – Distribution of metadata set volume

Adoption of metadata sets
In Table 1, the percentages of the metadata sets' spread among the observed OAs are reported together with their schema and typical prefix. Table 1 reports the spread percentages of 16 most used metadata, with respect to the total number of different sets of 153. Noteworthy is that the DC is largely the most common, while after DC a number of metadata sets is in the range of 8-15% such as RDF, METS, MPEG-21, etc.

N	Prefix	Schema
100%	OAI_DC	http://www.openarchives.org/OAI/2.0/oai_dc.xsd
15%	MARCXML	http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd
15%	METS	http://www.loc.gov/standards/mets/mets.xsd
14%	Rfc1807	http://www.openarchives.org/OAI/1.1/rfc1807.xsd
14%	Oaimarc	http://www.openarchives.org/OAI/1.1/oai_marc.xsd
11%	MPEG21-DIDL	http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-21_schema_files/did/didl.xsd
8%	RDF	http://www.openarchives.org/OAI/2.0/rdf.xsd
6%	Uketd_dc	http://naca.central.cranfield.ac.uk/ethos-oai/2.0/uketd_dc.xsd

6%	Junii2	http://ju.nii.ac.jp/oai/junii2.xsd
5%	Context_ob	http://www.openurl.info/registry/docs/info:ofi/fmt:xml:xsd:ctx
4,5%	Oai_etdms	http://www.ndltd.org/standards/metadata/etdms/1.0/etdms.xsd
4,5%	Xepicur	http://www.persistent-identifier.de/xepicur/version1.0/xepicur.xsd
4%	junii	http://ju.nii.ac.jp/oai/junii.xsd
3%		http://epubs.cclrc.ac.uk/xsd/qdc.xsd
3%		http://www.proprint-service.de/xml/schemes/v1/PROPRINT_METADATA_SET.xsd
3%	xmetadiss	http://www.d-nb.de/standards/xmetadiss/xmetadiss.xsd

Table 1 – Diffusion of the most used metadata sets

4.2 Archive Metrics

At the level of a single archive, a number of issues has been addressed, namely: the mentioned schema, the distribution of metadata sets, the number of records, the presence of empty records, etc.

Reference Schema for the Metadata set. As to the schema definition, a clear problem has been identified for Dublin Core. In particular, the most common reference to DC schema is:

URI: http://www.openarchives.org/OAI/2.0/oai_dc.xsd

While, we found other four different aliases in the xml namespace definition that have problems:

- URI:http://www.openarchives.org/OAI/2.0/oai_dc/oai_dc.xsd (does not work)
- URI:http://www.openarchives.org/OAI/2.0/dc.xsd (does not work)
- URI: http://openarchives.org/OAI/2.0/oai_dc.xsd (work, may be not official or identical to the main)
- URI:http://www.openarchives.org/OAI/2.0/oai_dc/ (does not work)

Another example is for Xmetadiss metadata set schema. The common reference to Xmetadiss should be:

URI:http://www.d-nb.de/standards/xmetadiss/xmetadiss.xsd,

while some OAs use an alias which does not work:

http://www.ddb.de/standards/xmetadiss/xmetadiss.xsd

This scenario shows that an extended quality check is needed not only for the value of metadata field, but also for the schema reference to certify the metadata quality.

Use of Metadata set. As discussed in Section 4.1, the distribution of metadata set is quite spread. Moreover, there exists a 15% of institutions using metadata sets which are personal model (single instances in the distribution) or which do not have a significant number of institutions. The adoption of non-standard metadata set and schema affects the effectiveness of archive visibility and distribution. Here below some examples:

URI schema: http://libst1.nul.nagoya-u.ac.jp/akf/akf.xsd

URIschema: http://uhasselt.be/agris/1.0.xsd 1

URI schema: http://ubib.eur.nl/eur_qdc/1.0.xsd

Small OA Archives. Among the identified OA from the general list, a number of them are quite small. For example, 14% of the archives have less than 100 records. In some cases, they are experimental projects instead of well established institutional archives. The threshold at 100 records may be set to another value, while for this preliminary assessment we have considered smaller archives, namely those with less than 100 records.

Empty Metadata Records. In some cases, the OA presents some empty records. This problem is more common than expected. In particular, from the performed analysis we have identified that more than the 2% of archives have 80% empty records, the 84% of the archives have no empty records, while the 9% have less than the 5% of empty records. We can identify in 5% a threshold beyond which the quality of the archive falls over.

4.2 Metadata Record Metrics

At the level of the single metadata record some problems have been detected regarding the interpretation of the single metadata fields. In this case, a greater attention has been focussed on DC since it is the most common metadata model (as shown in the past). Moreover, it is well know that the use of simple Dublin Core [6] foresees a high level of flexibility for filling in the metadata field. The performed analysis has shown that a very few number of institutions did adopt a qualified DC model, as defined by standard recommended best practice with a controlled vocabulary such as RFC 4646[14] or ISO639-1 [15]. Moreover, the metadata multi-language system is managed in two modalities: using different instances of DC: language for each language or expressing different languages in the same field with a separator. The analysis has outlined that this separator can be arbitrary the sequent types: ' ,' ' ;' ' – ' '/'. Table 2 gives an overview of the different instances for coding languages as found in the harvested metadata , please note that some of them are even not correct.

language	instances	tot
English	en, eng, English, en_GB, en-GB, Englisch	6
Spanisch	es, spa, Espanol, Spanish; spa; , sp	6
French	fr,fre, French, French;, Francais, fra	6
Deutsch	ger ,de, German, Deutsch, ge	5
Greek	gr, gre, grc, ell	4
Italian	it,ita, Italian	3
Japan	jpn, ja, jp	3

Table 2 – Diffusion of most used metadata sets

DC:format. The DC:format field can be filled in with file format, or with the physical medium, or with the dimensions of the resource as described by standard definition [6]. The recommended best practice refers to using a controlled vocabulary such as the list of Internet Mime Types [13], [16]. Here below we provide an example of the use of this field for JPEG and PDF file format. According to the common standards of IANA, each mimetype [13] can be coded with a limited number of instances, on the contrary, the collected mimetype contained even wrong coding in more than 10% of cases (depending on the specific type coded).

4. Conclusions

The adoption of Open Access model by the implementation of freely accessible institutional archive is growing more and more, particularly in scientific domains. On the other hand, the quality and trustability of the implementation and services provided still remains a requirement not well addressed and hard to assess formally and automatically.

The OA implementation landscape is very fragmented. Up to now, each institution tends to adopt its own policies for workflow, publication, access, descriptive metadata adoption, type and format of contents, distribution, etc. The scenario of the OA works by the scientific community is based on the assumption of interoperability; that is de facto hard to be massively viable. During the experiments, we realized that it would be difficult for an aggregator or OAI-PMH client to trust archives and get access to high quality records in every case. This paper has shown some of the most evident difficulties to maintain the archive services available on internet, so as to be able to answer to OAI-PMH requests. Moreover, the institutions present a certain delay in adopting and accepting standards. The number of different metadata schemas used to describe and classify OA resources, suggests that an international and community oriented policy for repository and related resource evaluation is missing. This can cause a reduction of credibility for the open access model as a viable channel to disseminate scientific knowledge, thus impacting on scientific community.

Presently, we are harvesting from repositories all the available records without limitation and by exploiting the GRID infrastructure. At the end of this process, a more exhaustive and detailed evaluation will be provided. If you are interested in inserting your OAI-PMH service in the analysis, please send an email to the authors of the paper at DSI/DISIT University of Florence. In this perspective, this paper's authors can perform on demand a direct specific assessment of your OAI-PMH service site and provide a corresponding record with suggestions to improve your service.

References

[1] C. Lagoze and H. V. de Sompel. The Open Archives Initiative Protocol for Metadata Harvesting, version 2.0. Technical report, Open Archives Initiative, 2002. http://www.openarchives.org/OAI/openarchivesprotocol.html

[2] Herbert Van de Sompel et al. Resource Harvesting within the OAI-PMH Framework. D-Lib Magazine December 2004 Volume 10 Number 12 ISSN 1082-9873

[3] AXMEDIS: Automating Production of Cross Media Content for Multichannel Distribution http://www.axmedis.org

[4] MPEG-21, Information Technology, Multimedia Framework, "Part 2: Digital Item Declaration," ISO/IEC 21000-2:2003, March 2003.

[5] METS, http://www.loc.gov/standards/mets/

[6] Dublin Core Metadata Initiative. Dublin Core Metadata Element Set, Version 1.1. http://dublincore.org/documents/dces/

[7] MARC ISO 2709 http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=41319

[8] DSpace http://www.dspace.org/

[9] Fedora http://www.fedora.info/

[10] EPrints for Digital Repositories http://www.eprints.org/

[11] Emanuele Bellini and Paolo Nesi - A Trust P2P network for the Access to Open Archive resources - IFLA2009 Milan

[12] Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. http://oa.mpg.de/openaccess-berlin/berlindeclaration.html

[13] Freed, N. and N. Borenstein. 1996. "RFC 2045: Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies," November 1996. http://www.ietf.org/rfc/rfc2045.txt?number=2045

[14] RFC4646- http://www.ietf.org/rfc/rfc4646.txt

[15] ISO639-1 http://www.infoterm.info/standardization/iso_639_1_2002.php.

[16] MIME-IANA http://www.iana.org/assignments/media-types/

[17] P. Bellini, I. Bruno, P. Nesi, "Visual Programming of Content Processing Grid", The 15th International Conference on Distributed Multimedia Systems, DMS2009, Hotel Sofitel, Redwood City, San Francisco Bay, USA, September 10 to September 12, 2009;

[18] P. Bellini, I. Bruno, P. Nesi, M. Spighi, “Taxonomy and review of Complex Content Models”, in Proceedings of the 14th International Conference on Distributed Multimedia Systems (DMS’2008), Boston, USA, 4-6 September 2008, pp.71-76.

[19] Xiaoming Liu et al. « Arc - An OAI Service Provider for Digital Library Federation” D-Lib Magazine April 2001 Volume 7 Number 4 ISSN 1082-9873

[20] Pleiadi www.openarchives.it/pleiadi/

[21] Citeseer http://citeseer.ist.psu.edu/

[22] U. Muller, A. Powell, P. Cliff, H. Van de Sompel, C. Lagoze, Nelson M. and Warner S - OAI for Beginners – the OpneArchive Forum online tutorial http://www.oaforum.org/tutorial/english/intro.htm

[23] W3C- RDF/XML Syntax Specification http://www.w3.org/TR/REC-rdf-syntax/,

[24] Carl Lagoze, Herbert Van de Sompel, Michael Nelson, . 2002. *Implementing Guidelines for the Open Archives Initiative for Metadata Harvesting: Guidelines for Harvesting Implementes* http://www.openarchives.org/OAI/2.0/guidelines-harvester.2002-06-10.htm

[25] OCLC – OAIHarvester, http://www.oclc.org/research/software/oai/harvester.htm