METALIS, an OAI Service Provider

Zeno Tajoli

Cilea, Sezione Biblioteche, via Raffaello Sanzio 4, 20090 Segrate (MI), Italy
tajoli@cilea.it

METALIS is an OAI Service Provider for the Library and Information Science field. This paper describes the metadata harvesting process and the crosswalks designed to homogenize metadata, the web interface of the Service Provider METALIS and the OpenUrl usage. To homogenize metadata it is necessary to analyse the OAI-PMH output of Data Providers and write ad hoc crosswalks. In particular METALIS homogenizes the fields with subjects, source archives, languages, types of material. As for the web interface, this paper shows how it is structured. As far as the OpenUrl usage is concerned, METALIS offers an innovative service, reversing the standard use of finding full-text versions (that are already available at the connected repositories) and providing a tool to find online resources that have a relation with the results found during the search

1 Introduction

METALIS[1], 'METAresearch in Library and Information Science', is a service accessible through the Internet that allows to search scholarly works in the field of Library and Information Science, available as full-text documents in open archives[2] . METALIS collects (harvests) metadata via the OAI-PMH[3] protocol and it can be defined a 'Service Provider' according to the OAI architecture.

In this paper the harvesting process is described. Then a list of metadata fields is provided, with the main conversion rules adopted to homogenize metadata from different archives. Indexing rules are detailed, followed by a section about the web user interface and an innovative service based on OpenUrl. METALIS is born at CILEA[4] , a consortium of Italian universities providing ICT facilities and services, within the AePIC[5] project, that actively promotes Open Access and Open Archives in Italy.

---

[1] METALIS is available at: http://metalis.cilea.it/. Service developed by Zeno Tajoli. Credits of S. Warner, A. Tugnoli, UKOLN and RDN.

[2] Open archives or repositories are compliant with the standards of the Open Archives Initiative (OAI): http://www.openarchives.org/ (see next note).

[3] OAI-PMH (2004), The Open Archives Initiative Protocol for Metadata Harvesting, Protocol Version 2.0 of 2002-06-14, Document Version 2004/10/12T15:31:00Z: http://www.openarchives.org/OAI/openarchivesprotocol.html.

[4] CILEA web site: http://www.cilea.it.

[5] AePIC Project web site (English pages): http://www.aepic.it/index.php?lang=en.

2 Harvesting.

Metadata are harvested from disciplinary and institutional repositories:

- @rchiveSIC : Sciences de l'Information et de la Communication, <http://archivesic.ccsd.cnrs.fr> [archiveSIC]
- arXiv, <http://arXiv.org/oai2> [ArXiv]
- Caltech Library System Papers and Publications, <http://caltechlib.library.caltech.edu/> [caltechLIB]
- CNR Bologna Research Library, <http://biblio-eprints.bo.cnr.it/> [CNRBologna]
- Digital Library for Information Science and Technology, <http://dlist.sir.arizona.edu/> [DLIST]
- E-LIS, <http://eprints.rclis.org/> [Elis]
- Librarians' Digital Library (LDL), <https://drtc.isibang.ac.in/index.jsp> [LDL]
- memSIC : Memoires en Sciences de l'Information et de la Communication, <http://memsic.ccsd.cnrs.fr/> [memSIC]
- Thèse-EN-ligne., <http://tel.ccsd.cnrs.fr/perl/oai> [telSIC]

METALIS harvests all metadata from @rchiveSIC, Caltech Library, CNR Bologna, DLIST, E-LIS, memSIC. It harvests only the 'Computer Science' section from arXiv, keeping only works with subject 'Digital Libraries' or 'Information Retrieval'. It keeps only the dissertation about Library and Information Science from CCSD. From LDL it harvests only two sections: Publications / Articles and Theses / Dissertations. The metadata are harvested in the OAI Dublin Core[6] format. This schema has the same semantics of 'unqualified Dublin Core'.[7] METALIS also uses metadata present in the 'header' and in the 'about' sections of the harvested XML format. Sections 'header' and 'about' are described in the OAI-PMH. In order to harvest metadata, METALIS uses a specific tool, a harvester. The harvester chosen to perform this task is Celestial[8] which has been partially modified[9]. The repositories selected are all the OAI-PMH compliant archives including papers in LIS field. The survey was done in October 2004, by reading through the available lists of OAI-PMH archives and some papers on the topic and by searching through various Internet search engines.

3 Crosswalks and indexing

In METALIS, all metadata harvested from different repositories are converted in an internal schema. Several fields have the same semantic as Dublin Core unqualified or a similar one. The elements of this internal schema are:

---

[6] OAI Dublin Core: http://www.openarchives.org/OAI/2.0/oai_dc.xsd.

[7] DCMI (2004), *Dublin Core Metadata Element Set, Version 1.1: Reference Description*: http://dublincore.org/documents/dces/ .

[8] Celestial: http://celestial.eprints.org/ .

[9] The modifications are available from http://www.aepic.it/docs/celestial/patch_celestial.tar.gz

- All the elements of the 'header' section of the XML record described in the OAI-PMH.
- All the elements of the 'about' section of the XML record described in the OAI-PMH.
- dccreator; union of dc:creator and dc:contributor.
- dcdate, the date of publication. It is one of the dc:date available.
- dcdescription, the same as dc:description.
- dcformat, the same as dc:format.
- dcidentifier, the same as dc:identifier.
- dcsubject, it contains classes of the JITA[10] classification. For every archive there is a specific conversion tool from their original data in dc:subject
- dctitle, the same of dc:title
- dcrights, the rights about the metadata. The policy is different for every archive.
- dcarchive, a string that identifies the archive. Archives codes are written above in the list of archives inside square brackets.
- dclanguage, codes of the languages used in the full-text document; the codes used are ISO 639 – 2 compliant. In most archives the field dc:language is present, with data in ISO 639-1. For other archives METALIS inserts a default language code.
- dctype, types of METALIS. For every archive there is a specific conversion tool from their original data in dc:type.
- dcbrief, this field contains the brief description view. The view is inserted with HTML code
- dcfulleng, this field contains the full description view. The view is inserted with HTML code
- sortauth, this field contains the string used for sorting by author. The string is composed by all authors' name in full. The order of the names is the same as in the original metadata.
- sortsubj, this field contains the string used for sorting by JITA classification. The string is composed by all alphabetic symbols of JITA classes. The order of the names is the same as in the converted metadata.

The field dctype has a fixed list of values.[11] These values are:

- Article = any scientific article or journal.
- Book = a whole book or a single chapter from a book
- ConferencePaper = any work presented at a conference or seminar
- Thesis = any dissertation or similar works
- ·GreyPaper = any other material (mostly grey literature)

---

[10] The JITA Classification Schema of Library and Information Science as described at: http://eprints.rclis.org/jita.html: the JITA Classification Schema has been developed starting from a merger of NewsAgentTopic Classification Scheme (maintained by Mike Keen at Aberystwyth, UK, until 31st March 1998) and the RIS classification scheme of the (now defunct) Review of Information Science originally conceived by Donald Soergel (University of Maryland).

[11] All these fixed values could be viewed as a specification of the type "Text" defined in DCMI (2005), Dublin Core Metadata Terms : http://dublincore.org/documents/dcmi-terms/ .

Below an example of the conversions performed on metadata is shown. In the table showing a subject conversion, JITA classes are represented with their alphabetic symbols for convenience. On the left hand side the original value from the archive, on the right hand side the converted one in METALIS.

@rchiveSIC [archiveSIC]:

From dc:type to dctype:

'Text'	Article

From dc:subject to dcsubject:

'History of information/communication', 'Knowledge management', 'Theory of information/communication', 'Others'	A
'Bibliometry, scientometry', 'Cinema, art, esthetics', 'Conflicts, information strategy, intelligence', 'Geopolitics', 'Local authorities', 'Organisation and communication' 'Public Sphere', 'Sociology of information and communication'	B
'Mass media', 'Scientific communication and information'	C
'Museology'	D
'Information/communication law',	E
'Economy, cultural industry', 'Education, e-learning, training'	G
'Electronic publishing'	H
'Hypertext, hypermedia', 'Information retrieval'	I
'Information system engineering',	L

The indexes are built from metadata in an XML format, using Cheshire2[12] as indexing software. Cheshire2 allows to install a server that works with two protocols, http and Z39.50. This was the reason why Cheshire2 was selected as indexer. METALIS now operates only on the http protocol but it can be quickly made ready to use the Z39.50 protocol.

4 Web interface

The structure of the interface is composed by single web pages with reciprocal connections. These web pages have a header and a tail in common, used to provide a common set of links in every page. All web pages are static and have a '.html' extension. They can be divided into four categories:

Presentation pages	OpenUrl setup	Forms	Dynamic pages
index.html credits.html (with source section)	setopenurl.html	simple.html advanced.html	results brief results full OpenUrl popup error

[12] Cheshire II Project Home Page: http://cheshire.berkeley.edu/

The "presentation" pages provide general information about METALIS.
The simple search form provides one box to insert keywords. Keywords can be connected with the Boolean operators AND, OR, NOT.

The advanced search form provides three boxes. The first box allows to search among titles, the second one among authors, the third box, labelled as 'abstract', operates on the Dublin Core field dc:description.[13] Any string can be evaluated as a list of keywords (single words are searched one by one) or as 'exact phrase' (searched exactly as introduced). As default the system inserts the operator AND between words. Other boolean operators can be used to connect the three boxes.

Both search forms provide the same filters and sorting options. Filters are: for year, for type, for the source of data, for language, for JITA class down to the second level. Filters are connected with the implicit boolean operator AND. The sorting options are: for author, for title, for year (ascending and descending) and for classification.

Both search forms provide help on how to write queries. If the search does not produce results or is written in a wrong way, the system intercepts the error and generates a dynamic web page to help the user to rewrite the query.

Search results are shown in a brief view (with author/s, date, title), sorted according the selected option, and arranged in groups of 15. Every result can be clicked to retrieve the record full view, where all its available metadata are shown, including the link to the full-text document.

At the end of the full view an OpenUrl service is provided. Every OpenUrl is specifically produced for the retrieved record.

5 OpenURL

In the full view METALIS offers a service of dynamic linking with the standard OpenUrl. The link sends the metadata about the displayed record to an OpenUrl resolver, which connects 'on the fly' the record with the on-line resources available for the user. How it is shown in Van de Sompel and Beit-Aire [1], the OpenUrl standard is normally used to retrieve an available full-text document from the available metadata. In the case of METALIS the full-text is already available through the specific link; therefore METALIS reverses the usual use of the OpenUrl. It does not do a search to find an object, but it creates a set of links towards wider searches. METALIS aims to provide links to resources that are related to the retrieved record. In fact this operation can be more or less successful, depending on the quality of the OpenUrl resolver and the amount of online resources available to its user. METALIS provides a free OpenUrl resolver, that can only manage free general resources. However the system allows to use the OpenUrl resolver available at the user's institution. The management of the OpenUrl resolver is done with the web page 'setopenurl.html'.

The system works using cookies in the user's browser. The configuration is specific for the computer used. In 'setopenurl.html' there is a box where the user can

[13] According to DCMI (2004) this field contains a more or less wide description of the work, and it is widely used to past the article abstract

insert the web address of his own OpenUrl resolver.[14] Then METALIS sends a cookie that is used to build the OpenUrl link in the full view. Behind the scene, it is necessary that the OpenUrl resolver used is able to understand the metadata sent by METALIS, that is METALIS has to be included in the description of the resolver sources. METALIS sends to the OpenUrl resolver the identifier found in the metadata. Through it, the OpenUrl resolver can find the metadata the connected record. This is the syntax of the OpenUrl link as established in the stardard[15]:

<Base Url>?url_ver=Z39.88-2003
&rft_id=info:<identifier type>/<identifier>
&rfr_id=info:sid/metalis&sid=metalis&id=<identifier>

The OpenUrl resolver used as default by METALIS is based on the ERRoLS[16] service provider. It retrieves the metadata connected with the identifier inserted in the OpenUrl link. The result is similar to the fig. 1:

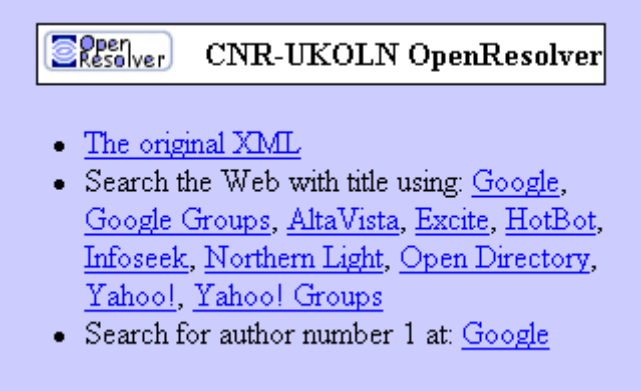

Fig. 1 (OpenUrl pop-up)

Graphics is different from other parts of the site to convey the information that another tool is brought in use. Title words are used as keywords to launch searches on Internet search motors as Google, AltaVista, Excite, Yahoo!, etc. Another option is the authors search on Google. More free on-line resources, such as OAI service providers, can be configured to provide extended searches. An informal survey on the usefulness of this OpenUrl service was done in January 2005: better results could be probably gained using the metadata 'keywords'. But 'keywords' are not available in OAI Dublin Core format. The best solution for this problem could be an agreement among Data Providers in LIS field for a specific metadata format. The experience of COLAP[17], as described in Simons and Bird [2], could be a useful guide.

[14] NISO Committee AX calls it "BaseUrl", in *The OpenURL Framework for Context-Sensitive Services*: http://library.caltech.edu/openurl/StandardDocuments/Z39_88_Pt1_ballot%20final.pdf
[15] see previous note.
[16] OCLC (2003), *ERRoLs for OAI Identifiers*: http://www.oclc.org/research/projects/oairesolver/default.htm
[17] Open Language Archives Community. http://www.language-archives.org/

6 Conclusions

This paper has described METALIS as a service provider, that allows to search through metadata about LIS papers harvested from open archives via the OAI-PMH protocol. METALIS also allows to find other resources related to the retrieved record via an innovative service based on the resolution of OpenUrls. The same architecture[18] and solutions may be proposed to build similar services for other disciplines or institutions.[19]

As a final consideration, you could say that, in order to improve the service in the near future, a more close interaction between Data Providers and METALIS is necessary. As said before the metadata 'keywords' can be quite useful, along with the bibliographic references of the work and the name of the journal or of the book where the paper is inserted.

The future plans for METALIS should be: to interact more closely with Data Providers, to insert new archives, to improve the user interface using the results of a survey amongst users.

Reference

[1] Van de Sompel, H. and Beit-Aire, O., Open Linking in the Scholarly Information Environment Using the OpenURL Framework, *D-Lib Magazine,* vol. 7, n. 3, March 2001. URL: http://www.dlib.org/dlib/march01/vandesompel/03vandesompel.html

[2] Simons, G. and Bird, S. The Open Language Archives Community: An infrastructure for distributed archiving of language resources *Literary and Linguistic Computing* 18, 2003, pp.117-128

[18] The code of METALIS is available from: http://metalis.cilea.it/credits.html#download

[19] Thanks to Marta Plebani for a revision of this paper and Susanna Mornati, AePIC Project Leader, for project support to METALIS and a revision of this paper.