

WORKSHOP 1

Harvesting metadata and the use of application profiles

Author: Theo van Veen, Koninklijke Bibliotheek

Objective

Since the introduction of the World Wide Web, physical distances hardly play a role anymore in searching for information. But now new barriers become visible in our search for information. The challenge we are facing now is that we want to find information hidden in databases with different record formats in different languages and accessible through different search and retrieve protocols and make this information usable by a wide range of users.

In this workshop I present an approach for meeting this challenge as we have been using for The European Library. The choices being made are based on making retrieved information from different sources usable and understandable for systems and users in order to increase the functionality offered to the user. We aim at lowering barriers for the user to access information and for data providers to provide that information. The final goal is to create a concept for a common information retrieval infrastructure that will also be adopted outside The European Library.

The objective of this workshop is to find out whether this approach is expected to be successful, to what extent we need to adjust this approach and what the next barriers will be. Although this workshop is not about The European Library, The European Library approach is taken as an existing example of how to deal with these aspects.

Analysis

To analyse the problem in more detail we need be clear about the actual objective. The main objective is to make more information easier accessible. Whatever information is there, we should be able to find it, retrieve it, understand it and use it. It should be possible to improve navigation between resources in a more or less automated way based on previous results. And the obtained metadata should enable functionality based on the machine readability of this metadata. How to deal with these requirements will be the main subject of this workshop.

We will divide the problem into two areas: 1) search and retrieval of metadata and 2) understanding and using metadata.

Harvesting versus distributed searching

The first aspect is search and retrieval. How do we search in all those different databases describing the interesting objects? Do we harvest all relevant sources and index the metadata centrally and let users query this central index? Or do we broadcast each query to multiple distributed targets?

Workshop Discussion Papers

To answer this question a comparison between central indexing and distributed searching is shown in the table below.

Central indexing	Distributed searching
Search options apply to index as whole	Search options may differ per target
Predictable performance	Performance determined by slowest target
Easy integration of results	Integration of results more difficult
Central controlled index	No control on distributed indexes
Less system load (one search one request)	More overall system load (multiplied by number of targets)
More resources required for indexing	Less resources required for indexing

Distributed searching is mainly implemented via Z39.50 but since the end of 2002 a new protocol is available Search and Retrieve via the Web (SRW). There are two versions having a different access mechanism (SRW based on the use of SOAP) and SRU (based on the use of URLs).

The current protocol for metadata harvesting is OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting). This protocol offers the possibility of harvesting records that are updated within a time span and belonging to a set, both specified by the harvester. The protocol is based on http and therefore simple to implement. It is not a search protocol: the use of sets is the only way to restrict harvesting to a specific area.

A fundamental aspect of cross searching different targets has to do with differences in search options like index names, operators, truncation etc. Records may contain different fields, different records types and different record formats. Some aspects of distributed searching are just related to actual Z39.50 implementations and not fundamental for distributed searching. For example in conventional Z39.50 systems, it is considered an error if a target gets a search request for a not-supported index and sometimes a portal would not even perform the query, when not all targets support it.

Question: How do we deal with a search for fields that are not present in all individual databases? Servers best choice? An error message? And how should searching in a central index deal with that?

The aspect of the system load can be best discussed by considering two extreme cases: indexing the whole world or broadcasting a query to the whole world. It is quite obvious that indexing the whole world is to some extent what Google does and obviously it is feasible. It may take some time for indexing but not for searching in the index. Broadcasting a query to the whole world and collecting the answers real time is however quite unrealistic. Both approaches can however be combined by a system of “distributed central indexes”, each of which is the result from harvesting distributed databases.

Question: would a hybrid solution of “distributed (subject or geographical oriented) central indexes” be a useful alternative for pure distributed searching or pure central indexing.

The distinction between distributed searching and harvesting is quite small especially since there is a new search and retrieve protocol (SRU/SRW), which is based on http also and could almost do what OAI-PMH does.

Workshop Discussion Papers

Question: Should AOI-PMH become merged with SRW to become a generic search, retrieve and harvest protocol (allowing a more sophisticated refinement than the use of sets in harvesting e.g. restrict a set by subject or classification)?

Currently the most popular way for searching is the Google type of searching but that doesn't show us information hidden in databases. In order to make the information in these databases available – either for searching or for harvesting – we first need to find the relevant databases.

Question: how do we find the right targets for harvesting and or searching? Is the central OAI registry sufficient or do we need more sophisticated mechanisms of automated use of databases with collection level descriptions?

And finally – assumed that we actually have a choice - we have to make a decision for each database to have it harvested or have it put on the list of targets for broadcasting our queries to.

Question: What is the best balance between harvesting and distributed searching?

Application profiles

The second area of interest of this workshop is understanding and using retrieved metadata. MARC was introduced as a machine-readable format for bibliographic records. The use of MARC helped a lot in converting to and exchanging metadata in a more or less common format regardless of the many MARC dialects.

The introduction of XML as record syntax and the introduction of Dublin Core as metadata standard brought a large change. With XML we now have a single record syntax that is not specific for an application area (like MARC was for bibliographic data). It came along with lots of tools to process XML encoded data and the number of XML tools is much larger than for MARC. XML is being used now for storage, conversion, as exchange-format and for separating contents from presentation. While MARC records are restricted to tags and subfields (only two dimensions) with restricted ranges, XML has hardly any limitations in that respect. And besides that it is also possible to encode MARC in XML in a reversible way without losing information. Although MARC is still being used in most library systems we see a trend towards using XML.

But what about Dublin Core? Being a simple pidgin language to allow for a high level of interoperability it is certainly not enough to provide the minimum level of functionality that is generally required in our applications. And there is certainly a lot of criticism on Dublin Core especially from librarians.

Question: What purpose is served by Dublin Core in the bibliographic world and for what reason should we use Dublin Core?

And additionally if Dublin Core in its current form does not serve this purpose well enough there is the question:

Workshop Discussion Papers

Question: What do we need to make Dublin Core better usable for exchanging and sharing bibliographic records?

Part of the answers on the questions comes from the DCMI organisation by the introduction of application profiles. ***The concept of application profiles (see [Application profiles: mixing and matching metadata schemas](#)) has emerged within the Dublin Core Metadata Initiative as a way to declare which elements from which namespaces are used in a particular application or project. Application profiles are defined as schemas, which consist of data elements drawn from one or more namespaces, combined together by implementers, and optimised for a particular local application.***

Application profiles offer a unique opportunity for sharing metadata and trying to find the right balance between convergence and diversity. The starting point for application profiles is simple Dublin Core. Where simple Dublin Core is not specific enough, there are predefined qualifiers and encoding schemes to refine the meaning and use of the Dublin Core terms. When this is still not enough one can "borrow" terms from other application areas and finally one can introduce new terms.

The qualifiers defined by DCMI are refinements of the 15 core elements and follow the dump down rule: the meaning of a term should, even without the qualifier, still cover the contents of a field. An example of a qualifier is the role of a creator like author or painter. Encoding schemes do generally not refine the meaning of terms but specify the use or format of a term. Examples are ISBN and ISSN as possible encoding schemes for the identifier element.

When we extend the 15 elements with new elements, new qualifiers or new encoding schemes, that is where the application profile starts to become useful. Different application areas may require different terms, elements, refinements and encoding schemes but will also share existing terms, as much as possible, especially the core elements. An example of an application profile is the library application profile.

Considering the many MARC fields and subfields that are in use, we might expect a large number of new terms waiting to be introduced in for example the library application profile. We can introduce new terms and when they do not fit in a common application profile one could start making special application profiles. However, that is the turning point where diversity possibly starts increasing divergence.

Question: How do we find the right balance between introducing new terms and keeping aligned with other applications and application areas.

What often happens, is that for new collections new terms are introduced because these terms fulfil a local function. Consequently new metadata models arise and different collections use different terms for the same entity. Some terms are only of local interest and there is no reason for sharing those terms. In a lot of cases however, other applications use similar terms. At this point we have to keep in mind what we are aiming at. When the metadata are presented through a specialised interface then there is no use in sharing, but when metadata are presented independent of the original user interface and external applications are expected to offer functionality based on this metadata, that is a different case.

Most metadata allow for some kind of functionality. Resource discovery and describing an object are probably the most important functions of metadata. Other functionality is translation of fields, linking, navigation etc. Making fields usable by external applications for these functions require a very precise agreement on the definition of these terms and

Workshop Discussion Papers

have a common understanding of these terms. We need controlled vocabularies and value ranges and these have to be defined in application profiles. And we need to agree on implementation details.

Question: how do we harmonise the use of terms in different application profiles?

Perhaps answering this question is the most important objective of this workshop. The way we deal with this problem in The European Library is by introducing a metadata registry. This metadata registry contains information about metadata in the application profile but also metadata that are proposed but not or not yet accepted for the application profile. This registry is open for everyone for inspection, but only partners in the project may submit proposals for new terms. It is important for The European Library to continue to develop and introduce new terms when needed. The same holds for other organisations and projects. And we need to find a way to exchange information in these local registries in order to deal with diversity and at the same time convergence.

Question: Do we need a central agency to store and expose a common application profile?

Related links:

Z39.50 - <http://www.loc.gov/z3950/agency/>
SRW/SRU - <http://www.loc.gov/z3950/agency/zing/>
Open archives Initiative- <http://www.openarchives.org/>
Dublin Core Metadata Initiative - <http://www.dublincore.org/>
The European Library – <http://www.europeanlibrary.org>