Portal metadata

Juha Hakala

Metadata is essential to all major applications used by libraries. E-resource management (ERM) systems, digital asset management systems (DAMS), metasearch portals and traditional integrated library systems (ILMS) and future tools still to be released all require metadata. Unfortunately we do not know yet what kind of metadata some of these applications need. This problem is tied to another, a more fundamental one: there is no collective agreement on how some of the above mentioned tools should "work". This article discusses the metadata needed in portals and standardization work carried out to develop metadata in the NISO Metadata Initiative.

Integrated Library Systems (ILMS) have been around for about two decades. Due to this long experience, the library community has quite a fixed definition of an ILMS. To function well, these applications require, for instance, bibliographic, holdings and authority data. There are cataloguing formats for all these data sets, and with the exception of holdings, also cataloguing rules have been in existence for quite a long time. Cataloguing practices of libraries have been aligned so well that national and international copying of bibliographic records is not only feasible but an essential part of modern cataloguing procedure.

Metadata requirements of DAMS and ERM systems overlap to some extent with those of ILMS. All these applications deal with publications, and reliable bibliographic data is essential to their operations. There is no way a digital asset could be managed properly, if our basic description of it is faulty.

The interesting question is: what do we need beyond the bibliographic kernel in order to build, for instance, an efficient DAMS? Preservation metadata is an obvious but also an obscure candidate: we need this kind of metadata to be able to preserve digital content to future users, but we do not have a standard definition of preservation metadata yet.

This article will concentrate on metadata needed in portals. It will describe the standardization work carried out in the NISO Metadata Initiative, which has produced such essential documents as NISO Z39.91 (Collection description specification) and NISO Z39.92 (Information retrieval service specification). As of this writing, these draft standards are under review; it is quite likely that they will be approved in autumn 2006.

Service descriptions

The purpose of metasearch portals is to provide an easy access to a set of remote databases. Contents and access mechanisms of these resources may vary a lot, but the portal should be able to hide all these complexities from the patrons. Its success in this depends on the accuracy and completeness of the metadata about the target systems contained in the portal.

A portal which does not have reliable and up-to-date information about targets may easily become a bottleneck, something that prevents people from using the target systems, instead of helping them in this task. We already have examples of commercial portal applications which have failed to deliver useful services due to persistent problems in the quality of service descriptions.

There are many reasons why service descriptions have been such a nuisance. First, lack of metadata format and exchange syntax have prevented portal hosts from sharing this data, at least across application barriers.

This is a pity, since creating a service description of - for instance - a library OPAC or union catalogue is a time consuming task, given the richness of search options: there are plenty of search terms that can be used, and often they can be truncated and/or combined with the Boolean logic. Describing formally the semantic richness of an OPAC is not a simple task, not even for systems librarians.

Second, potentially, a more hazardous issue is a fact that the vendors have regarded service descriptions as a strategic asset of their portals. Unlike bibliographic data, service descriptions have not been created and shared by libraries but built, maintained and sold by vendors.

In fact, some portal providers have bought service descriptions from the third party. Such a business model makes it rather difficult to guarantee that service descriptions are up to date, and may make it difficult for the libraries to fix the problems they encounter.

Third, and perhaps the most frustrating problem, is that service descriptions have been built manually. Usually high quality structured metadata is human made but service descriptions are an exception. It is possible to build an application which queries the networked databases in order to find out their access parameters.

Unfortunately this approach is only viable when the target system is accessible via a commonly known search protocol such as Z39.50 or SRU. Proprietary systems are unpredictable; there is no way to automate the screen scraping process via which descriptions of these targets are constructed, usually by programmers.

To make things worse, if a proprietary interface is changed, even smallest changes may cause the service description to break down. And only a programmer can fix the service description.

Portal hosts must have better control over service descriptions in their systems, and there has to be an efficient way of creating and sharing this metadata. Following steps can and should be taken:

1. Implement the NISO Z39.92 standard in metasearch portals and other applications which require service descriptions (such as Web OPACs)
2. Build an open source tool which can create NISO Z39.92 service descriptions
3. Organize international exchange of service descriptions between portal hosts

The first step should not be too complicated. Z39.92 defines an XML-based exchange syntax for service descriptions. A compliant portal must be able to read and write service descriptions in this syntax, just like ILMS must be able to read and write ISO 2709.

There are of course also differences between ILMS and AACR2/MARC/ISO 2709 and portals & NISO Z39.92. Service descriptions of non-standard databases - which are actually small programs, for instance Perl scripts - can not be expressed in Z39.92 syntax and therefore there is no way of exchanging them. And there are no cataloguing rules for service descriptions, so their completeness may vary a lot. But any exchange of (correct) service descriptions is better than none.

Helsinki University Library and Index Data launched in spring 2006 a project which will build an open source Z39.92 tool called Keystone. Once completed, any portal host or vendor can utilize this tool as deemed fit. From the host point of view, its usefulness will depend on whether the portal can ingest the harvested service descriptions.

As of this writing, no portal supports Z39.92. But this lamentable situation will change. In the meanwhile portal hosts can investigate if they could build themselves a conversion from the standard exchange syntax to the internal data representation format (and possibly vice versa).

It remains to be seen how useful the automatically harvested service descriptions are. Optimism is not groundless: Many years ago Index Data built a similar tool called Z-Spy, which has harvested service descriptions about 1400 Z39.50 targets. This data has proved to be quite useful. Yet this can not be taken for granted.

One of the challenges facing the developers is that some Z39.50 servers can be mis-configured. For instance, instead of telling honestly which search attributes it really supports, a server may claim that it supports any Z39.59 use attribute.

This is not possible, since no OPAC contains every data element known by MARC21. When the server receives a query with non-supported attribute it may convert it into a much broader search using 'Any' attribute. From the searching point of view, the result would be a disaster. Diagnosing these situations and fixing them is a major challenge; it remains to be seen if it can be solved programmatically.

Standardization and technical development based on it is often relatively straightforward. But politics may be more complicated. Even if NISO Z39.92 is approved and implemented, international cooperation - which is a necessity, since Internet knows no national borders - may still not take off. Initiatives such as TEL (The European Library) are paving the way to a common understanding that Google is not a solution for every problem. There are a lot of valuable resources embedded in Internet databases, hidden (usually) from Google but accessible via portals.

Collection descriptions

The problem with Internet databases is that we do not know what they contain, and there are no efficient tools for finding this out. Google is of no

help, and service descriptions do not say anything much about the contents of the services.

The NISO Metasearch Initiative has solved this problem by complementing service descriptions by collection descriptions, and developing a data model which binds the two together. A service, for instance the Z39.50 server of the Helsinki University Libraries' OPAC, Helka - makes available all the (catalogued) collections of the Helka libraries. On the other hand, any such collection is available also via an SRU server and the Voyager proprietary access protocol.

From a technical point of view, a portal cannot operate without service descriptions, still it can do without collection descriptions. But from a patron point of view, trying to find a relevant database from a portal is like seeking a needle from a haystack if there are no collection descriptions at all, or if they are not good enough.

The scale is definitely an issue here. If a portal contains about 50 target databases, not much support is needed to select the most appropriate ones. When there are 500 targets, like in the Nelli portal in spring 2006, it is a challenge to find the relevant ones. And when there are 5000 targets, reliable collection descriptions are a must.

We could of course try to solve the problem by limiting the target proliferation. But in a national portal such as Nelli there is no way of limiting the growth of services configured into the system. Different libraries have different priorities; technical universities will add to the portal their international peers, just like the national library will add its own.

If and when international exchange of service (and collection) descriptions becomes popular, large amounts of this information can be harvested and utilized in such portals as.Nelli. From patrons' point of view this opens new possibilities: they will have an efficient means of finding relevant resources and making queries from them. And this service can not be challenged by Google, since creating collection descriptions is a manual and laborious process. It can not be automated; Google can not replace a human (in a foreseeable future) in making a collection out of a set of items.

The NISO Collection description specification (Z39.91) is a simple tool for describing collections. Just like Z39.92, the standard defines a set of metadata elements (28 in all) and XML-based exchange syntax. Both draft standards have deep roots: Z39.92 is based on the Z39.50 Explain service and subsequent attempts to simplify it, while Z39.91 is an extension of the Dublin Core Collection Description Application Profile, which in turn has its inspiration in the RSLP Collection Description Specification.

The 28 elements of Z39.91 are a core set of metadata elements needed for description of collections. It is certain that libraries, museums and archives using the standard will recognize omissions, and new elements will be added in the future. This is something to be expected: the first version of the MARC format had only 20 tags, while the MARC21 of today has about 200.

Whether the developers of Z39.91 identified the 20 core elements correctly remains to be seen. But by building upon earlier work the risk of making mistakes has been diminished, if not eliminated,

I am confident that for now the standard developers have done what is needed, and it is up to libraries and other portal hosts to start producing a sufficient amount of portal metadata. This is a challenge, since we must continue to produce metadata to ILMS, and start planning metadata production to DAMS and ERM systems as well. And this must be done in the face of the common trend of moving personnel resources from the back to the front office. But the work in the front office is difficult without well working tools, and with incorrect metadata no ILMS or portal will be efficient enough.

References for the standards:

NISO MetaSearch Initiative:
http://www.niso.org/committees/MS_initiative.html

NISO standards
http://www.niso.org/standards/index.html

NISO Z39.91-200x, Collection Description Specification
http://www.niso.org/standards/resources/Z39-91-DSFTU.pdf

NISO Z39.92-200x, Information Retrieval Service Description Specification
http://www.niso.org/standards/resources/Z39-92-DSFTU.pdf

ANSI/NISO Z39.50-2003 Information Retrieval : Application Service Definition & Protocol Specification http://www.niso.org/standards/std_info_retrieval.html#Z39.50

MARC standards
http://www.loc.gov/marc/

AACR2
http://www.aacr2.org/

Juha Hakala, Development Director, Helsinki University Library - The National Library of Finland - Chair, NISO Standards Committee BB (Task Group 2): Collection & Service Descriptions
email. juha.hakala@helsinki.fi