

# **The European Library (TEL) – The Gate to Europe’s Knowledge: Milestone Conference**

29-30 April 2002 Die Deutsche Bibliothek, Frankfurt am Main, Germany

**Janifer Gatenby**

*Aiming at Quality and Coverage Combined - Blending Physical and Virtual Union Catalogues.*

## **Why Library Catalogues and Union Catalogues?**

A common question posed these days is the relevance of library catalogues when there is so much available on the web that is directly searchable through Internet Search Engines such as Google, AltaVista, Lycos, Northern Light, etc.

To assure ourselves, we need to do a comparison. Internet search engines:

- Lack precision
  - ☞ They retrieve too much because they are retrieving mostly from full text. To compensate they rely on relevance ranking. Library catalogues by contrast index mostly cataloguing (metadata) that ensure greater precision.
  - ☞ They retrieve irrelevancies.
  - ☞ Indexing full text is not precise because of lack of metadata; e.g. it is not possible to search a name specifically as an author and there is usually no controlled subject searching. At best, there is broad category searching.
- Retrieve Poor Quality Material
  - ☞ Relevance ranking according to number of times a site has been visited does not effectively filter out all poor quality material
- Recall –
  - ☞ Relevant material is often buried if it is not overly descriptive and lacks metadata
  - ☞ Broken links result in failure to recall relevant material
  - ☞ Web services and contents of databases are not touched; therefore the databases behind the web pages are not covered
  - ☞ Resources that are not available on the web are not covered. Libraries hold the key to vast amounts of information, whether available electronically or not.

## **The Importance of Metadata**

Library catalogues are based on the tenet that accurate descriptive and subject cataloguing of materials selected for their quality ensures search results that are relevant, largely complete and include for the most part, works of value and quality.

Authority control of names, titles and subjects aims at collocation to assist recall and ordered sorting of result sets.

A third and no less important function of metadata lies in its descriptive content that serves to accurately identify materials. Uncatalogued web documents, for example, are identified by their URL or location on the web and can disappear when they are re-located on a web server.

Publishers recognized the importance of metadata to describe their materials long before the arrival of the World Wide Web, illustrated by their cooperation in cataloguing in publication schemes. With the exponential growth of resources available on the World Wide Web, there has been a drive for publishers and authors to create their own metadata to assist in the retrieval and identification and hence preservation of their materials. Publishers have embraced metadata to a much greater extent than authors, even so, estimates currently indicate that less than 2% of web resources and pages contain metadata.

Libraries are also selecting web resources and creating metadata. For example, there are approximately 700,000 records added via OCLC's CORC system. In reality, there is an enormous task for the library profession world-wide. Much quality material is now bypassing more traditional publishing avenues and is only identified by a URL and available from a potentially ephemeral location. The library profession must accept the challenge to select out quality uncontrolled materials and to catalogue them. There is so much to cover that the only way forward is by co-operation on a large scale.

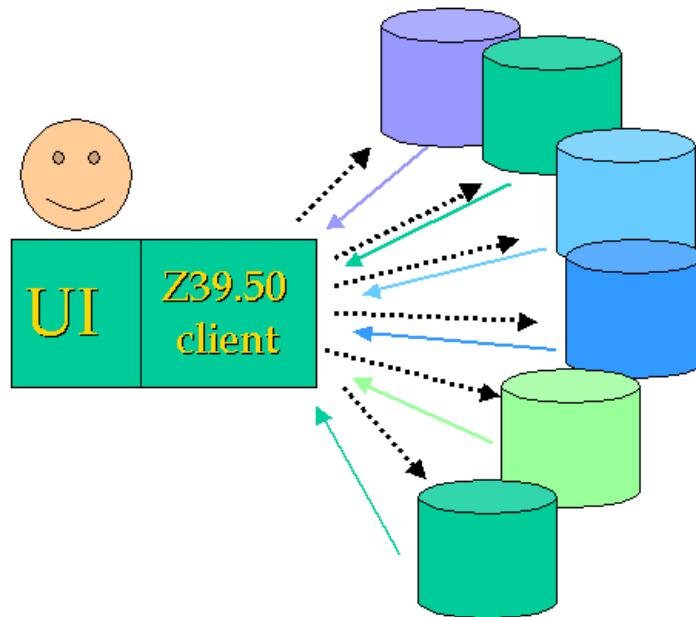
### **Metadata Retrieval Models**

What is the optimal model for storing and accessing cataloguing and metadata? There are 3 possible models that will be examined; distributed, distributed data with a centralised index and centralised union catalogue.

#### ▀ Metadata Retrieval – Distributed model

The data is located on multiple databases and is accessed by clients using standard protocols such as Z39.50, its emerging off shoot ZING/SRW or the emerging XML Query. There is no centralization of the data. The searches are usually sent in broadcast mode to multiple databases and the results received are combined into a common result set.

## Distributed Model



### ➤ Advantages and Disadvantages of the Distributed Model

#### *Compared with a search engine*

Searching by a protocol can offer more function than that provided by a search engine, because the client is doing the requesting and is intelligent enough to know how to handle the results. For example, with Z39.50, the following is available:

- **Precise searching**  
For example it is possible to search the title area only. Also the full range of boolean operations and nested queries are possible.
- **Result set handling**  
Large result sets can be delivered by breaking them into segments and it is possible with some servers to request that the results should be sorted before delivery, e.g. by the latest date and to request the removal of duplicates. It is also possible to refine a result set by adding limiters, for example. Result sets can be referred to without the need for cookies.
- **Browse**  
It is possible to browse indexes of titles, authors, subjects, etc. Browsing assists in retrieval when a need is either imprecise or ill defined. Browsing also assists to confirm negative search results.
- **Retrieval options**

The client can request brief or full records or specify the content that is most suited to the display being built. It is also possible from some servers to choose the format of the results, e.g. displayable text or a MARC record.

- **Extended services**

It is possible to request that a set of results be saved on the server, or that a search be saved together with a schedule for periodic execution. It is also possible to order an item or to add, modify or delete database records.

- **Links to physical and electronic documents and resources**

Metadata retrieved from Z39.50 servers can point physically or can contain elements for the construction of dynamic links to electronic resources. It may also contain pointers to locations of physical documents not available electronically.

#### *Advantages of Z39.50*

- **Multi-target searching**

From a single user interface, an identical search can search multiple targets even where they are dissimilar in platform (UNIX, NT, IBM, etc.) or database system (relational, network) or database model.

- **Searching based on abstract concepts**

The use of abstract concepts, e.g. such as “title” enables each server to map to its actual structure. The client does not need to know the database columns.

- **Can combine results from diverse databases**

Because common standard record formats are exchanged, e.g. MARC or XML/DC, it is possible to combine the results from diverse databases

#### *Disadvantages of Z39.50*

- **Complex and difficult to implement**

Generally speaking, the standard is not quickly learnt and understood by programmers and the expertise is consequentially not widespread. Therefore it is costly to implement and for some institutions programmes may be very difficult to maintain if the staff that originally wrote them are no longer in their employment.

- **Too many options**

Interoperability problems have developed due to the large number of options available in the standard. Profiles such as the Bath profile have attempted to limit options and define precise semantic meaning to those options, but even with this profile, the number of defined searches has been successively reduced due to lack of common agreement on semantics and difficulty in encouraging widespread adoption. Publication of a profile is not automatically sufficient argument for existing systems to make the necessary programme and database model changes and the consequent re-indexing. The result of this is that in spite of the number of options in the standard, the actual searches that can be

achieved in broadcast searches of diverse targets is quite small and generally considerably less powerful than native search interfaces.

- **Interoperates with the web but .....**

Web Z39.50 interfaces exist but Z39.50 is an older standard and does not fit comfortably with web programs. Z39.50 requires its own TCP/IP port, more often than not port 210, and firewall restrictions in many institutions make implementation of Z39.50 politically and administratively difficult. The fact that Z39.50 maintains state whereas the web does not is a typical example of the awkwardness of the combination of the two in a common implementation.

- **Z39.50 has not achieved widespread acceptance**

As a result of these disadvantages and also the perception that Z39.50 is restricted to library type applications, the protocol has not achieved wide acceptance.

### *Other Search Protocols*

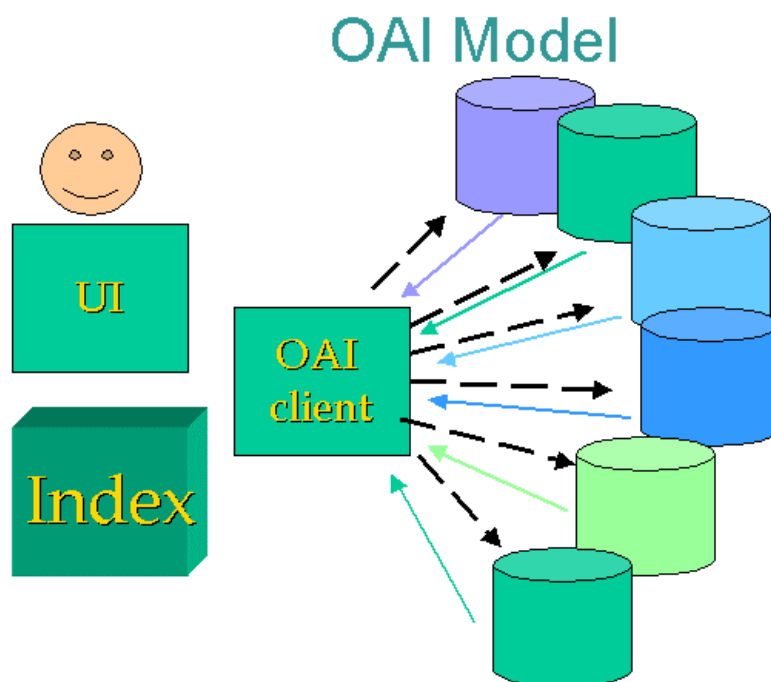
- **ZING SRW / SRU**

This initiative promises to retain the advantages of Z39.50 whilst at the same time modernising it to be more compatible with the web, to simplify it and thereby overcome many of its disadvantages. It can be implemented programme to programme in the form of SRW (Search and retrieve for the Web) using remote procedure calls (HTTP POST with SOAP). It can also be implemented for a thin client that is web browser based in the form of SRU (Search and retrieval URL).

The maintenance of state has been removed and there is only one record syntax. Both SRW and SRU retrieve text XML documents. The search has been simplified to a string search called CQL which is based on the Common Command Language, that inherits the search modelling of Z39.50 and the Bath profile.

## Metadata Retrieval - Distributed data; centralised index

The data is located on distributed databases but there is a centralized index. Internet search engines follow this model. They retrieve data from servers, index the data then discard it, retaining URL pointers to the data. The OAI harvest protocol has been developed as a standard way of harvesting data for purposes such as creation of a centralized index.



## Advantages and Disadvantages of the OAI Model

### *Advantages*

- **Common index more powerful than distributed searching**  
Broadcast searching of metadata distributed over several databases starts to develop inefficiencies beyond a certain number of targets, commonly believed to be between 5 and 10. If the databases are diverse, then the searching capabilities are reduced to basic level, there is no authority control that potentially could affect recall, and duplicate grouping or duplicate removal is not effective with large result sets just when it is most desired.
- **Common index is faster than distributed searching**  
Non-responding or slow responding targets make the consolidation of combined result sets a slow process.
- **Simple model for update**

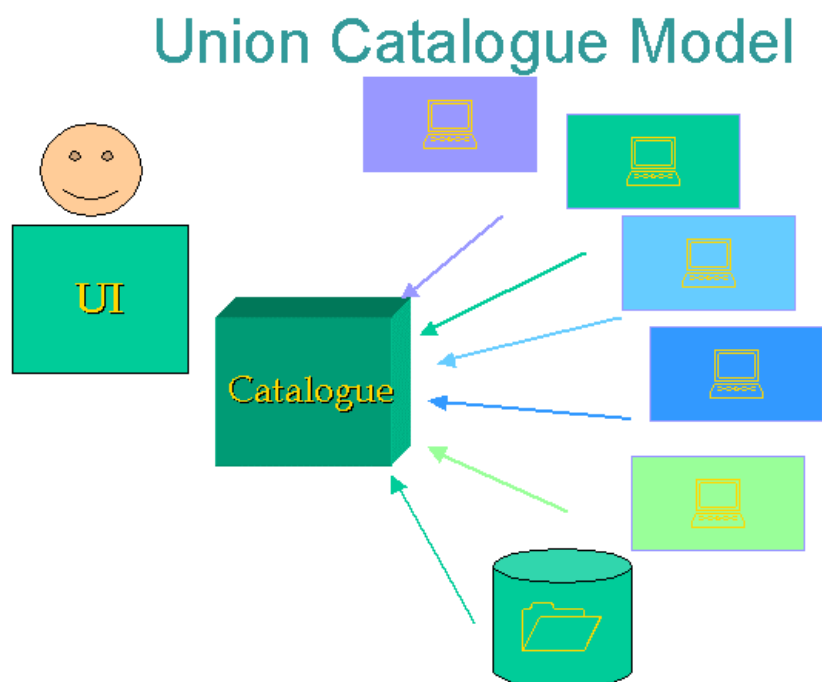
The OAI model facilitates distributed creation of data with centralised indexing.

### *Disadvantages*

- **Designed for documents not metadata**  
The OAI harvesting protocol was primarily designed to facilitate the indexing of online documents. Its primary purpose was not to gather metadata for the creation of a large metadata index.
- **Metadata can change when updated**  
The OAI harvesting protocol has no feedback mechanism in the way of diagnostics to indicate metamorphosis of the data that may have occurred once updated into a centralized index such as authority control changes and duplicate detection.
- **Incomplete update model**  
The OAI harvesting protocol includes additions but does not allow correctly for data modification, or deletion and hence the possibility of keeping databases aligned with the OAI protocol are limited.

## Metadata Retrieval - Centralised; union catalogue

A centralized catalogue is created by a mixture of methods, online addition and batch load. The catalogue may contain pointers to full text and it may contain other references to locations from which the full text may be accessed, e.g. library holdings symbols.



## Advantages and Disadvantages of the Union Catalogue Model

### Advantages

The union catalogue model has all the same advantages over the distributed model as the common index model.

- **Common index more powerful than distributed searching**
- **Common index is faster than distributed searching**
- **Simple model for update**
  - One of the main aims of union catalogues is to share the cataloguing effort, reducing duplication and saving costs.

In addition it overcomes some of the disadvantages of the OAI model

- **Designed for metadata not documents**
- **Complete update model**
  - Can include duplicate grouping or duplicate removal
  - Can include authority control
- **Discovery tool**

Union catalogues are authoritative sources. They identify the existence of works and facilitate their retrieval through links to online texts, abstracts and reviews, etc. and pointers to locations of physical copies. As pointers to location of physical copies, union catalogues are the backbone of Inter-library loan services.

### *Disadvantages*

- **Cost of maintenance**

Centralised union catalogues entail a maintenance cost, albeit often offset by the savings of cataloguing sharing. The virtual union catalogue has been proposed as an alternative.

### *Real versus Virtual Union Catalogues*

#### REAL

- Many accesses
- Generally more speedy because using a native search engine
- A greater precision in the definition of a search is possible
- Duplicate grouping or removal can be done in advance
- Consistent indexing and authority control results in better recall
- Delivery via ILL and text links
- Better and richer presentation is possible together with the possibility of customised multiple views based on user sign on to reflect language and region
- Resource for data mining, collection development, authentication

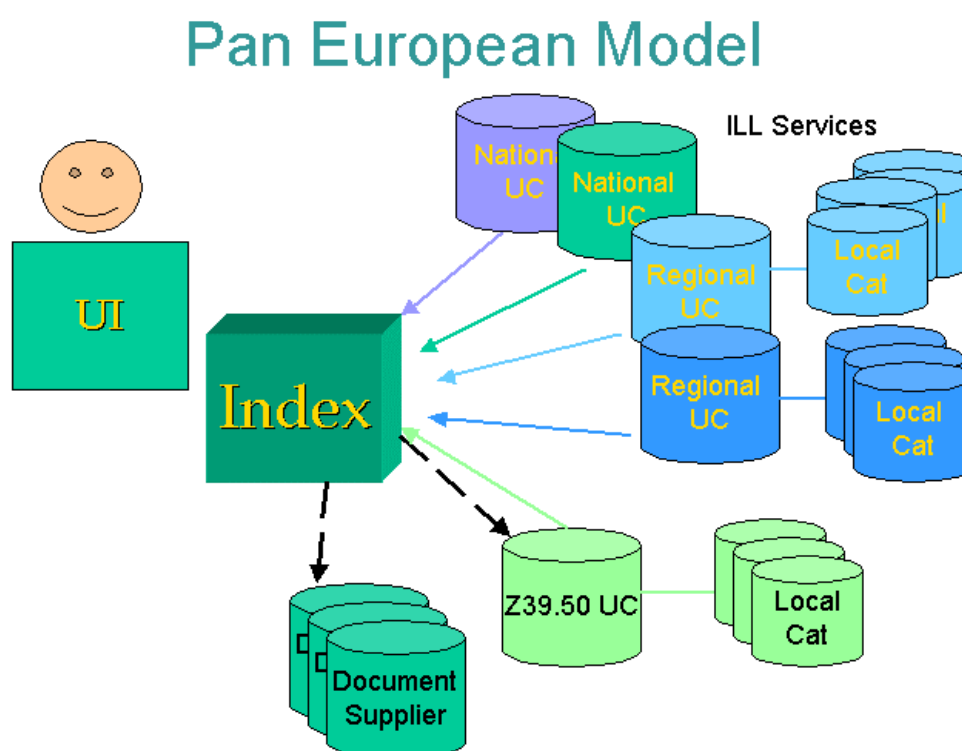
#### VIRTUAL

- Lower establishment cost
- Less visible maintenance cost
- Less politically difficult to establish
- Less politically difficult to ensure full participation

The arguments for and against virtual union catalogues mirror those for and against distributed searching. These are covered well in a paper by Clifford Lynch in 1997. This paper is still largely relevant despite its relative age. Building the Infrastructure of resource sharing: union catalogs, distributed search, and cross-database linking by Clifford Lynch <http://www.caslin.c2:7777/caslin99/a3.htm>

## ► Pan European Catalogue

A workable infrastructure for a Pan European Catalogue is a combination of the centralised and distributed models with a centralised index pointing to multiple large scale union catalogues from which services such as ILL, links to local systems, authentication, collection analysis and data mining can be provided. Links to document suppliers can be made from either the centralised index or from participating union catalogues. Where it is not practical for a union catalogue to contribute to the centralised index, search via protocol can be used.



## ➤ EUCat as a Basis for a Pan European Catalogue

EUCat is conceived as an index created from various large union catalogues in Europe. The index is created currently by proprietary protocol linking union catalogues that share the same database search and index engine, PSI. Other protocols such as Z39.50 update and the OAI harvesting protocol could also be introduced for the creation of the centralised index.

EUCat currently has several features that make it an attractive European resource

- **Multi-lingual**  
The catalogue is currently in English, Dutch and German and is being extended to French. Addition of other languages is a minor development and will flow on from new agreements as they are made. UNICODE has been implemented.
- **Duplicate clustering (not removal)**

Records describing the same materials are identified and linked. They are not removed because they usually contain different language of cataloguing, different subject headings and classification and links to local holdings.

- **Local views**

Depending on the affiliation of the user, the view of the system and of the records will vary. For example a German user will see a tailored home page in German and will see the German record with German language of cataloguing, German subject headings and classifications. The user will also see at first all holding locations in his region, followed by other locations in other regions and countries, e.g. the Netherlands.

- **Multi-cultural retrieval**

Because of the clustering of duplicates, any access point in any record will result in the retrieval of the whole group of records. Thus a more complete record in one language will compensate for an incomplete one in another.

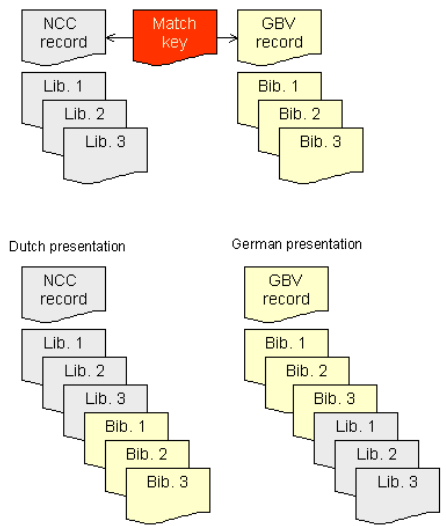
- **Multi-authority view**

Just as the bibliographic records are clustered, so also are the authority records. This architecture is in harmony with approaches taken by current experimental project work such as MACS <http://infolab.kub.nl/prj/macs/>, VIAF (a project to link name authorities involving the Library of Congress, The Deutsche Bibliothek and OCLC) and LEAF <http://www.crxnet.com/leaf/>. Both OCLC PICA and the Office of Research at OCLC are also experimenting with multi-lingual subject and classification enrichment.

- **Part of an extended Word Cat**

EUCat has a sister across the Atlantic that lives in Dublin Ohio, called WorldCat. The concept of extending WorldCat was developed in 2000 as a modern infrastructure that would address the challenges of bibliographic control and discovery in the current and possible future environments (<http://www.infotoday.com/newsbreaks/nb001030-1.htm>). EUCat in Leiden and WorldCat in Dublin, Ohio will become the first two nodes of an extended WorldCat. WorldCat currently comprises 49 million bibliographic records and EUCat more than 30 million although in associated services, namely PiCarta and FirstSearch, each contains many additional analytic records pointing to full text. PiCarta includes 70 million records.

WorldCat will comprise pooled bibliographic and authority data with local nodes responsible for holdings and services. It will also comprise a pooled international library directory. Some nodes may choose to access the services of a different node and there will be cross node services for holdings discovery and Inter Library Loan. Detailed statistics combined with authentication details to support multiple financial arrangements, e.g. concerning re-use of records from the common pool.

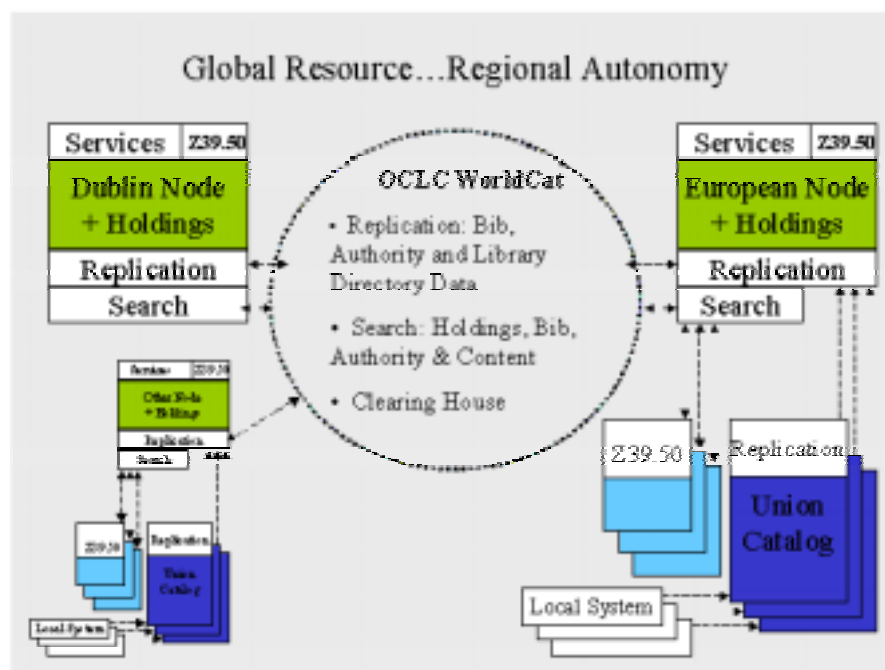


*Customised local views from EUCat*

## EUCat

Three levels of partnership are possible with EUCat. A full member will contribute metadata and holdings to the EUCat index. A second level member will contribute metadata but no holdings and a third level member will make protocol search access arrangements.

EUCat currently consists of the holdings of the Dutch union Catalogue (NCC), the libraries of the North German States (GBV) and the Newspaper Index of the Deutsche Bibliothek (ZDB). Within the next few months, EUCat will commence loading as a mirror, all the European holdings that currently exist on WorldCat. Arrangements for connecting to the various InterLibrary Loan systems are currently being negotiated as are arrangements for the incorporation of new full partners into EUCat.



- Which Standard are Needed to drive the Pan European Catalogue?

### *Which standards - Searching?*

For distributed searching, the following standards are necessary:

- Z39.50 (<http://www.loc.gov/z3950/agency/>) and the emerging ZING SRW / SRU (<http://www.loc.gov/z3950/agency/zing/zing.html>)
- Bath profile (<http://www.nlc-bnc.ca/bath/>)
  - Bibliographic
  - Holdings
  - Authorities
  - Cross domain
- Other searching standards may emerge
- Record syntaxes and schemas
  - MARC
    - ISO 2709
    - OAI XML encoding (<http://www.dlib.vt.edu/projects/OAi/marcxml/marcxml.html>)
    - MODS (<http://www.loc.gov/standards/mods/>) etc.
  - Metadata
    - Dublin Core (<http://dublincore.org/>)
    - ONIX (<http://www.editeur.org/onix.html>)
    - METS (<http://www.loc.gov/standards/mets/>)
    - Etc.
  - Z39.50 Holdings schema (<http://lcweb.loc.gov/z3950/agency/defns/holdings.html>)
- FRBR is under investigation as a standard for the provision of better presentation and navigation. If it proves successful, then work is needed to incorporate FRBR elements into existing search protocols and to develop schemas for structuring of records for exchange. (<http://www.ifla.org/VII/s13/frbr/frbr.pdf>)

### *Which standards – Update?*

To create the centralized index, the following standards are necessary

- Z39.50 update and the Union Catalogue Profile (UCP) have not been widely implemented.
- OAI. This could be extended with record structures from the Union Catalogue Profile (UCP) for record deletion, record maintenance and diagnostics to cope with authority control and reports on duplicate grouping.
- FRBR promises to enable copy cataloguing at levels and hence greater efficiency. It also promises to address the need for better management of rights. As with searching, update demands the creation of record structures for the exchange of records adherent to FRBR. This also entails the creation of identifiers for the various component records of FRBR.

- Record structures for simplified cataloguing – Dublin Core, Mods, Onix.

*Which other standards?*

- For linking OpenURL is the emerging standard ([http://www.niso.org/committees/committee\\_ax.html](http://www.niso.org/committees/committee_ax.html)) . On one level this standard enables a simple identification number search on a foreign server. It is used to discover full text, reviews and related materials such as citation index materials and also for order placement. What makes it different from other identifier standards like ISBN, ISSN etc. is that it is also a standard for the dynamic creation of an identifier for serial articles.
- ISO ILL (ISO 10160 / 10161) (<http://www.nlc-bnc.ca/iso/ill/standard.htm>) . This standard is currently undergoing a minor revision.
- Circulation – NCIP ([http://www.niso.org/committees/committee\\_at.html](http://www.niso.org/committees/committee_at.html)) . This new standard enables accessing local systems for the placement of loan requests and reservations and also for discovering the status of items and users. It also includes authentication and is used as an alternative to more mainline authentication standards such as LDAP.
- Directories – ISO 2146. The ILL implementers' group (IPIG) is currently creating a structured library directory. This will be used as the basis for a revision of the ISO directory standard ISO 2146. Sections on curriculum strengths and reference services will be added to inter-library loan descriptive elements. For extended WorldCat, the library directory will play an essential role.
- NISO is currently working on standards for an XML data schema and protocol for the exchange and forwarding of reference queries. ([http://www.niso.org/committees/committee\\_az.html](http://www.niso.org/committees/committee_az.html))