Janifer Gatenby and Rein van Charldorp,
EUCAT: A Pan-European Index of Union Catalogs

from / *aus*:

Union Catalogs at the Crossroad
Edited by
Andrew Lass and Richard E. Quandt
pp. */ S*. 31-50

Erstellt am 31. März 2005

Impressum

Table of Contents

Part 4 Hungarian Union Catalogs

Part 5 Baltic Union Catalogs

Part 6 South African Union Catalogs

Chapter 1
EUCAT: A Pan-European Index of Union Catalogs

Why a Pan-European Index?

Janifer Gatenby and Rein van Charldorp

End-users want a single, comprehensive, online source as exemplified by the success of Internet search engines, of which Google is a notable example. In just 4 years, Google has indexed more than 2 billion URLs and has grown to be the most-used search engine.[1] As a result of experience with such search engines, users are increasingly expressing a desire for a single point of access to library resources.

A single point of access to European library resources would offer end-users comprehensive, high quality, verified materials, with access to related materials, online text, and delivery services for offline materials.

From the user perspective, it is the content and comprehensive coverage that are important, far more important than the software, techniques and protocols used to achieve the interface and service. Moreover, users want access to content without having to learn the names and coverage of all the databases that would potentially house what they need. They want access with the minimum of training. Users do not have the same needs, nor does any one user always want the same type of information, at times requiring exhaustiveness and at others just what is readily available. What people require is a system that is flexible in the views that it can present.

---

[1] Google Inc. Fact Sheet, http://www.google.com/press/facts.html.

From a library perspective, a single point of access to European resources would offer the ability to provide a comprehensive view of available library materials Europe-wide and worldwide, backed by inter-library loans and document delivery services. This would ensure maximum exposure to their collections. The index should serve as the pivotal point for document delivery services with the library in a central role, leading to both digital and non-digital materials and the necessary requisites for access where appropriate. The index should also provide a tool for cooperative collection building and allow the library to be a member of more than one contributing union catalog.

It is important to create an environment and architecture that enables union catalogs to flourish, since they would be the main contributors to a pan-European index. The diversity of languages, cultures, cataloging rules, subject and name authorities and classifications, and other national and regional conventions make the task of a fully centralized union catalog almost impossible and inoperative. A federated approach is therefore necessary to ensure comprehensive coverage. Additional potential benefits from participation in a large central index include maximized ILL and document delivery services, efficiencies in creating quality shared linking services and enriched data mining services. The contents of the union catalog can be analyzed statistically in relation with other union catalogs providing, for example, information for EU projects and cooperative efforts in general.

To date, there have been numerous attempts to create virtual union catalogs, with Z39.50 as the key protocol in achieving this via broadcast searching. Examples of such virtual catalogs include the ONE project, the TEL project, the Canadian Virtual Union Catalog (VCUC) and the Texan Union catalog (ZLOT). These projects have all achieved moderate success, but the more individual catalogs that are searched simultaneously, generally the more slowly the results are presented, and retrieval from those catalogs is inconsistent due to differences in data indexing. Indexing differences also mean that searches that are common to all databases are few and basic, and consequently, precise broadcast searches are often not possible. Duplicates are retrieved in the results, and duplicate detection and grouping fail to work in a timely fashion over large result sets—just when they are most needed. These principal drawbacks to virtual catalogs have resulted in interest in the

harvesting model used by Internet search engines. If the data can be gathered, loaded and indexed centrally with duplicates removed or grouped as part of the update process, then the major drawbacks of virtual catalogs could be overcome. The OAI protocol, originally conceived for the harvesting of documents, was recently being considered for the creation of physical, non-virtual centralized indexes and union catalogs. Instead of the virtual catalog, a better model would be a centralized index for searching links to full text, to specific catalogs for services such as loans and photocopies, and to suppliers of other services, e.g. online book suppliers, databases of reviews, biographies, encyclopedic articles and so on.

Building such a centralized index could only be possible if done in a cooperative manner. Some tasks are so great they can only be achieved with the cooperation of a large number of parties, some of whom are otherwise competitors. Realizing a single comprehensive user point of access is critical to the continued central role of libraries in information provision. If libraries were to drop out of the limelight and their funding consequently reduced, it is likely that their role in cultural preservation would be difficult to fulfil. The world would enter a period of information chaos due to the concurrent upheaval in publishing, where it is easy to publish directly on the Web without peer review, control of document authenticity, or preservation and archiving. The result would be a permanent loss to the cultural heritage.

1   Description of EUCAT

EUCAT was conceived by OCLC PICA. As a not-for-profit company that is fully dedicated to libraries, it has served since the 1970s and supports major library installations in the Netherlands, Germany, and France. It is a company that has the business infrastructure, experience, software, and human resources to realize a pan-European index.

EUCAT is a pan-European index of union catalogs. It provides a quality catalog based on metadata, with duplicates identified and grouped and with authority control of authors, subjects and other headings ensuring consistent indexing and recall.

The main focus of EUCAT is as a discovery tool, linking to the contributing union catalogs or individual catalogs for services, in particular

inter-library loans and linking to licensed full text services. Document supply services, enriched contents, reviews, abstracts, and e-books may be accessed directly from EUCAT or from the participating union catalogs. EUCAT is the entry point for discovering and locating the riches of European libraries, both physical and digital.

EUCAT will also be used to support cooperative collection development by providing statistical analysis of coverage and by allowing libraries to record areas of proposed intensive development, e.g. digital projects. An additional benefit achieved by a centralized index is that in itself it becomes an authoritative source by its size alone and by being based on the resources of libraries, i.e. professionally created collections. This is particularly important at a time when the controls of traditional publication with editorial and peer review are being severely challenged by easy online publication and distribution. National libraries are among the major contributors to EUCAT that gives a combined index of national bibliographies and legal deposit indexes. EUCAT is thus a resource for establishing the authenticity of published and publicly available works.

Initially, EUCAT is not a source of copy cataloging. The index can direct to union catalogs from which copy cataloging may be made available depending on local arrangements. The index's main purpose is discovery; to add copy cataloging would entail complex arrangements to ensure the participation of some, and may deter some important libraries from participating.

OCLC PICA makes EUCAT available through different services, principally PiCarta and Publiekwijzer. It will also be possible to access EUCAT via external interfaces and portals using a search protocol, in particular ZING/SRU or SRW, Z39.50 and OpenURL.

2   Current Composition of EUCAT and Current Services

EUCAT currently consists of the holdings of the Dutch Union Catalog (*Nederlandse Centrale Catalogus*, NCC) and the libraries of the North German States (*Gemeinsamer Bibliotheksverbund*, GBV).
The central union catalog of the Netherlands represents the holdings of 642 libraries associated with 14.5 million bibliographic records. Text-based

materials, books, articles and serials represent nearly 95% of resources, with the remainder being printed music, sound, audio-visual and online resources (see Figure 1).
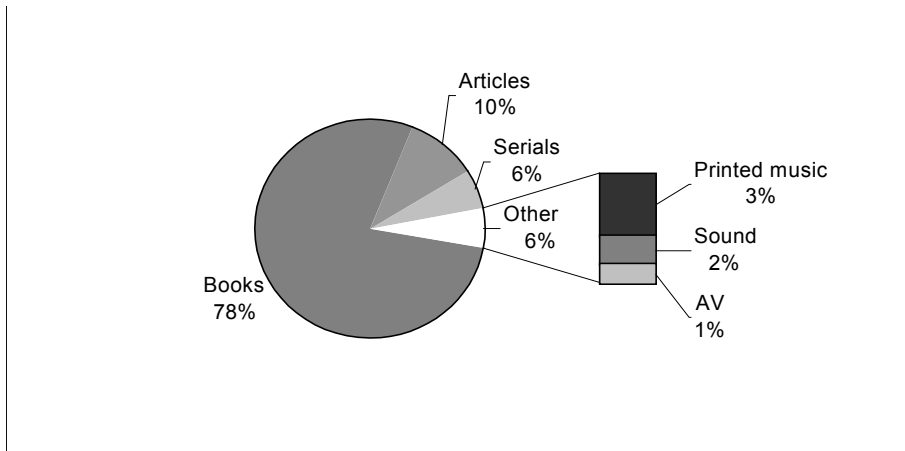

Figure 1. Form of Material Analysis of NCC
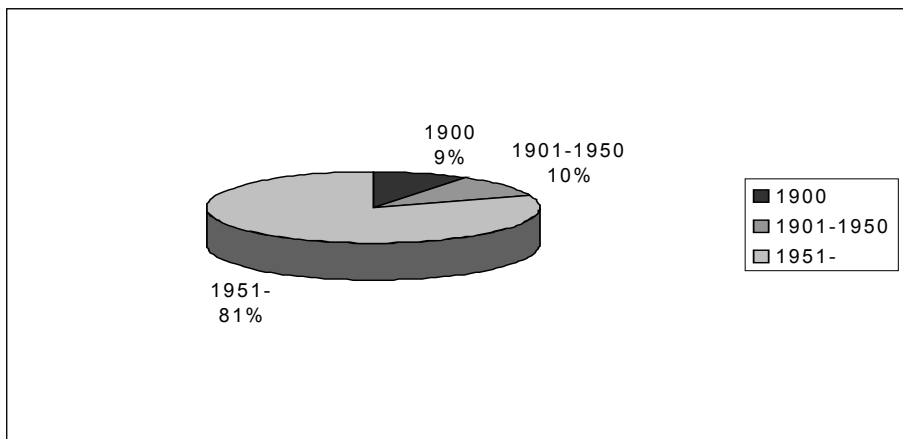
Figure 2. Date of Publication Analysis of NCC
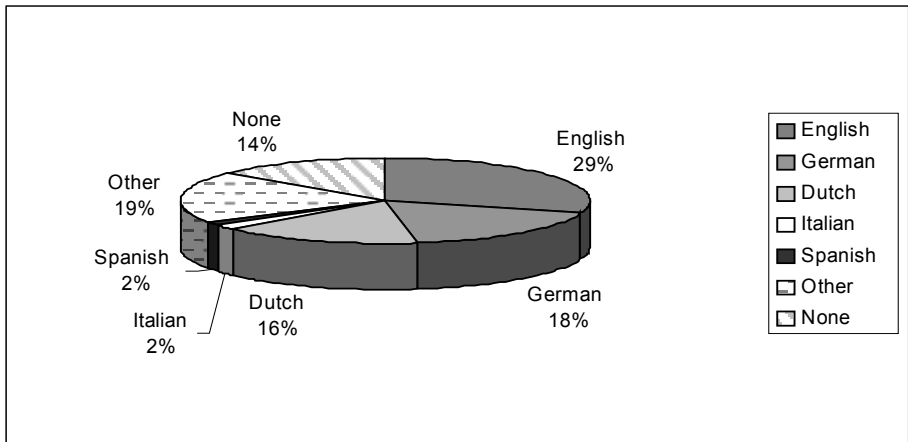
Figure 3. Language Analysis of NCC

Approximately 9% of titles were published before 1900, and over 75% since 1951 (see Figure 2).

The NCC catalog is available to Dutch end-users through the PiCarta service and to external systems via the Z39.50 protocol. As well as including EUCAT, the PiCarta service in the Netherlands also includes online contents data consisting of metadata and abstracts that are linked to full text services (depending on the license of the library). Users may view either the entire service or a specific catalog. Currently there are on average 150,000 inter-library loan requests and 40,000 copy requests per annum. There are just under 16,000 Dutch end-users who directly (i.e. unmediated) generate 40% of the loan requests and 20% of the copy requests. Direct access to full text (document delivery) is growing steadily, from just 5,000 in 2000 to an estimated 55,000 in 2002 (January to August showed 36,000). This is expected to continue growing, replacing inter-library loan requests. Figure 4 indicates end-user requests.

GBV records comprise the holdings of over 400 libraries, or 37 million holdings associated with 20 million bibliographic records. These have been loaded to EUCAT and matched with the Dutch bibliographic records. Where a match has occurred, a link is made between the records such that

the holdings of both records can be viewed no matter which record is retrieved and displayed.
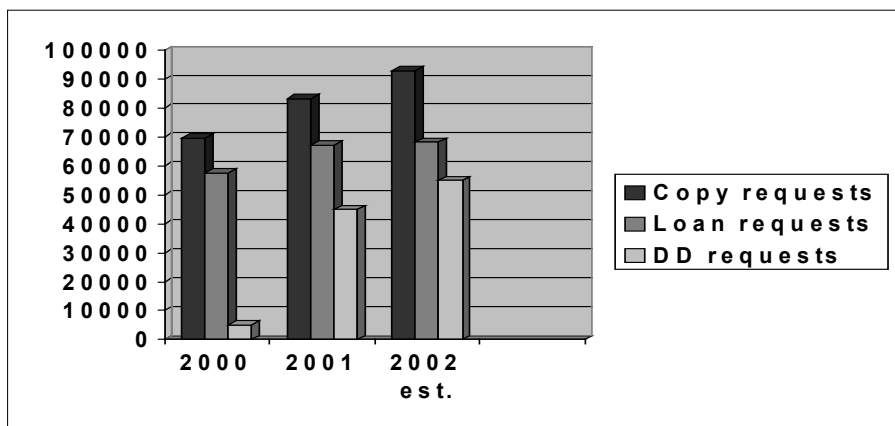
Figure 4. End-User Requests

The most appropriate record is displayed, depending on the user's login and reflecting the language of cataloging, subject headings and classification. This grouping and merging has the effect of virtual enhancement. For example, one record may contain a classification number not present in the other, but both records are accessible from the single point of access.

Approximately 20% of GBV records have been clustered with NCC records. The actual number of duplicates could be as high as 30% if algorithms as well as standard identifiers were used in matching.

Negotiations are underway for the users of GBV to have access to EUCAT through the PiCarta service. The logistics of international inter-library loans have not yet fully evolved because libraries are reluctant to give end-users direct access to this facility. Therefore, all requests are directed first to the user's union catalog for inter-library loans, and libraries will be able to restrict placement of international requests.
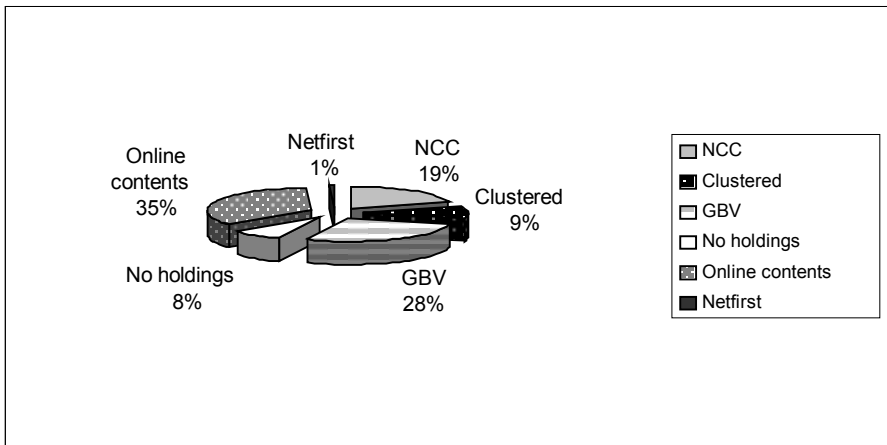
Figure 5. Database Composition of PiCarta

3   Expansion

The current contributors to EUCAT use the OCLC PICA CBS system for catalog maintenance and inter-library loans. This is not envisaged as a constraint on the system. Contributions from all major European union catalogs and libraries are necessary to provide the ideal index and single access point. As a first step in broadening the coverage of EUCAT, the European holdings from WorldCat will be loaded into EUCAT as a mirror copy.

OCLC WorldCat currently contains 21 million holdings from 430 European libraries. Actual figures on language, date coverage, and material type will be determined after loading. Arrangements for connecting to the various inter-library loan systems are currently being investigated, as is the determination of the business models required to ensure the widest possible cooperative participation in EUCAT. European WorldCat holdings represent full holdings for some libraries, and only the results of retrospective conversion projects for other libraries. The ideal situation would thus be to load from both union catalogs when possible, as well as from local catalogs if necessary.
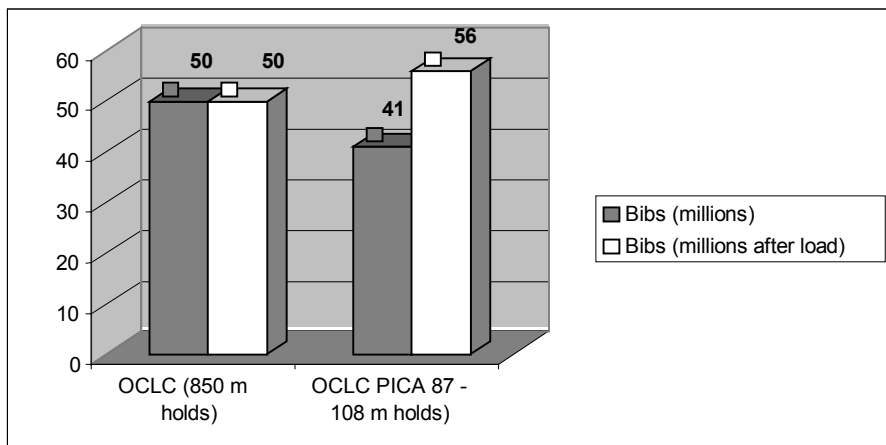
Figure 6. Estimated Unique Bibliographic Records

Figure 6 indicates the total bibliographic records of OCLC and compares them with the estimated total unique bibliographic records of all the OCLC PICA installations. It is estimated that the EU bibliographic records, with their holdings from OCLC's WorldCat, could add an additional 15 million unique bibliographic records to OCLC PICA's pool, and that the holdings would grow by 21 million, from 87 million to 108 million. Interestingly, it is estimated that fewer than 20% of Dutch titles and fewer than 40% of German titles in the Dutch national union catalog are also represented in OCLC's WorldCat. Similar overlap figures could be expected for other European national union catalogs, so it is clear that the bibliographic pool would be significantly increased with the contributions of such catalogs.

4   EUCAT Architecture and Standards

Architecture
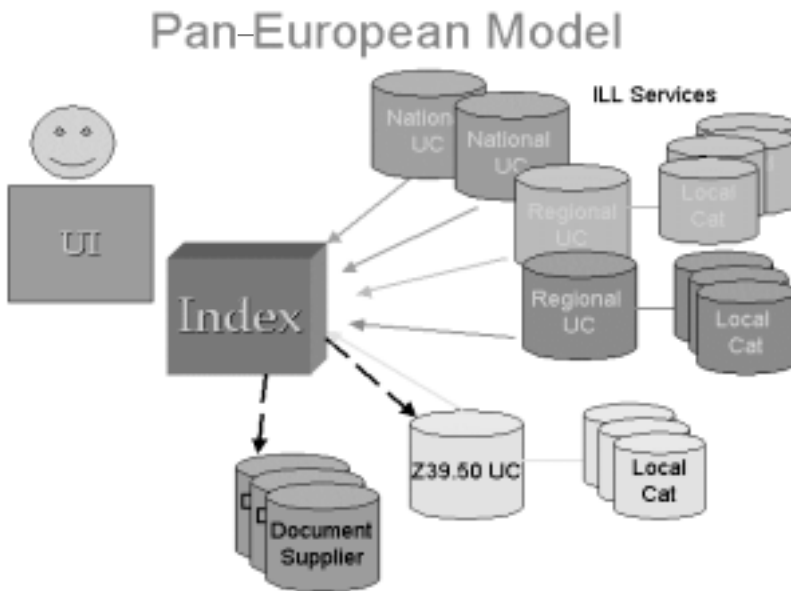
Figure 7. Pan-European Catalog

As new records are added and existing records are changed and deleted in the NCC and GBV union catalogs, they are also pushed directly (in the background) to EUCAT. All systems currently use the same OCLC PICA system, CBS/PSI, so that there is no need for an intermediary protocol. As others participate in EUCAT, alternative update mechanisms will be provided, most probably with a Z39.50 update (UCP profile) where the union catalogs push data to EUCAT, or with the OAI harvesting protocol, where EUCAT would poll the external systems, thus pulling the data. Consideration is being given to including the data structures of the UCP

with OAI, so that it will correctly handle modification and deletions as well as additions. Batch loading via FTP will also probably be an option.

From the user's viewpoint, the index can show everything or be filtered regionally, by format, by language or by other criteria such as date.By default, the records are filtered to show local records first, together with local holdings, with access to other regional holdings.
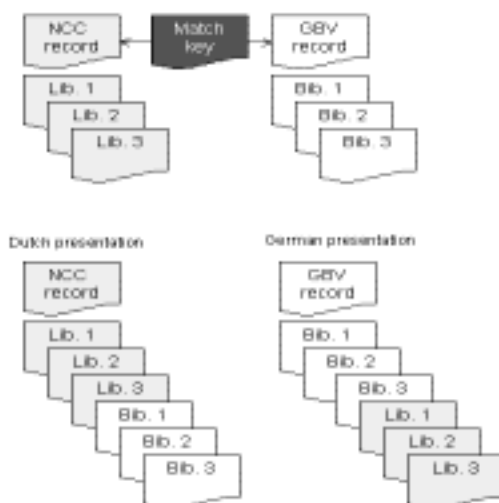
Figure 8. Record Filtering

End-user authentication is important both to determine the default views and confirm access to document delivery and other services. External authentication servers can be accessed using standards protocols such as LDAP and Athens.

EUCAT is also a part of WorldCat. EUCAT in Leiden and WorldCat in Dublin, Ohio will become the first two nodes of an extended WorldCat. Bibliographic and authority data from local nodes will be pooled together with a centralized and replicated international library directory. Holdings are held in the nodes, and hence the services are decentralized. A European and a world view will thus become available.
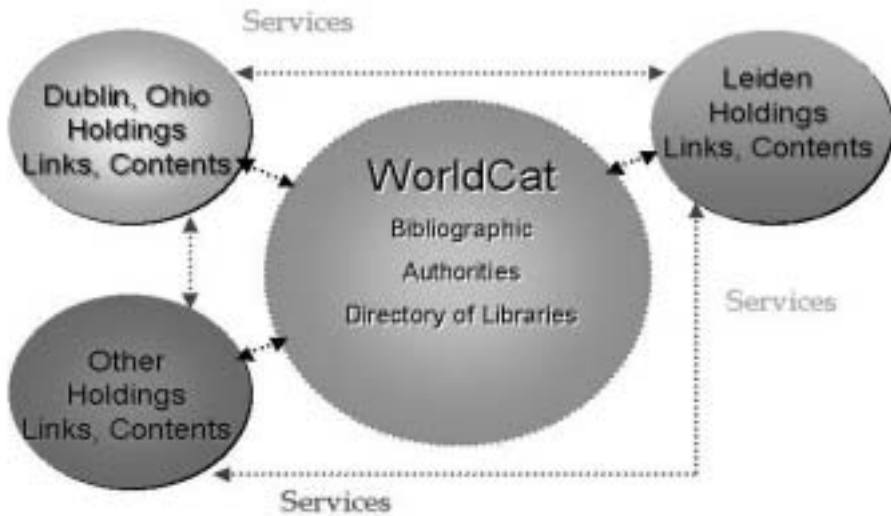
Figure 9. The Extended WorldCat Network

Importance of Standards

To build a resource cooperatively and to ensure its general usability, standards are essential. EUCAT will be a resource shared by many different and disparate systems. It will have many different interfaces.

Searching

So that external systems can access EUCAT for searching, the index will be available via standardized search protocols including:

Z39.50
(http://www.loc.gov/z3950/agency/) and the emerging

ZING SRW/ SRU
(http://lcweb.loc.gov/z3950/agency/zing/srw/specifications.html)

Bath Profile
(http://www.nlc-bnc.ca/bath/)

- Bibliographic
- Holdings
- Authorities
- Cross-domain

Other search standards may emerge.

These same protocol standards will be used to access those catalogs where the institutions have opted for partial participation in EUCAT.

Updating

For updating the protocols, Z39.50 update (UCP), OAI and FTP have already been mentioned. One very important advantage of the EUCAT architecture is that it can accept and deliver multiple formats. The following syntaxes and schemas are possible:

MARC
- ISO 2709: all variations including MARC21, UNIMARC
- MAB
- OAI XML encoding
  (http://www.dlib.vt.edu/projects/OAi/marcxml/marcxml.html)
- MODS (http://www.loc.gov/standards/mods/) .

Metadata
- Dublin Core (http://dublincore.org/)
- ONIX (http://www.editeur.org/onix.html)
- METS (http://www.loc.gov/standards/mets/)

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 and update protocols and to develop schemas for the structuring of records for exchange. FRBR also promises to permit copy cataloging at levels, and hence greater cataloging efficiency. (http://www.ifla.org/VII/s13/frbr/frbr.pdf)

Other Standards

OpenURL Standard

For linking, OpenURL is the emerging standard (http://www.niso.org/ committees/committee_ax.html). On one level, this standard enables a simple identification number search to be run on a foreign server. It is used to discover full text, reviews, and related materials such as citation index materials, but 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 minor revision.

Circulation—NCIP

(http://www.niso.org/committees/committee_at.html). This new standard enables local systems to be accessed for the placement of loan requests and reservations and also to discover 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. The library directory will play an essential role in the extended WorldCat.

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). The main user interface to EUCAT will provide users with the ability to pose questions from any result page. Data from EUCAT can be used for creating a question or providing an answer.

Cooperative Development and Experimentation

There is plenty of scope for cooperative development and experimentation among union catalogs that would be facilitated by participation in a common project. Examples are:

- Enhancements in retrieval; extension of the concepts of views and filters (by user, region, language, interest etc.);
- Improvement in the efficiencies of creating metadata—author-applied, program generation and extraction using algorithms, application of new information data models to evolve simple copy cataloging,
- Evolution in authority control to facilitate multilingual and multi-classification retrieval. FRANAR, VIAF and OCLC PICA's Colibri are project examples;
- Digital preservation and digital vault facilities;
- Sharing Web resources, pathfinders, predefined pages and links;
- Data mining to identify high-quality works that can be used in relevance ranking;
- Systems and programs for regular testing of URLs and 'shingles testing' to detect substantial changes in Web resources;
- Remote access, authentication and rights management.

5   Conclusion

EUCAT, as a centralized index to European library resources, provides fast, relevant and comprehensive searching through consistent indexing. Already large, it is capable of growing on a much larger scale, and OCLC PICA has the infrastructure to realize it. The cooperatively built index accommodates the diversity of European cataloging practices (codes, subject headings and classifications), languages and formats. It links to union catalogs, local catalogs and online providers for services. It is also capable of being linked from other services, e.g. e-learning environments, local library Web pages etc. and from abstract and indexing databases for European holdings.

EUCAT, with its association with WorldCat, can provide a gateway to library resources that rivals Google as a gateway to Internet resources.

Unlike Google, the library resource will lead to an available copy somewhere, the quality of resources retrieved is more consistent and the search capability yields more precise and complete results.

Glossary and References

ABES	*L'Agence bibliographique de l'enseignement supérieur.* See http://www.sudoc.abes.fr
Bath Profile	The Bath Profile is an ISO Internationally Registered Profile (IRP) of the Z39.50 Information Retrieval Protocol, intended as a basis for effective interoperability between library and cross-domain applications. Conformance to this Profile's specifications will improve international or extra-national search and retrieval among library catalogs, union catalogs and other electronic resource discovery services worldwide. See http://www.nlc-bnc.ca/bath/ap-bath-e.htm.
DublinCore	Dublin Core Metadata Initiative. See http://dublincore.org/.
EUCAT	EUCAT is a Pan European index of union catalogs. It may be accessed by the PiCarta service. See http://www.oclcpica.org/? id=102&ln=uk.
FRANAR	Functional Requirements of Authority Numbering and Records (FRANAR). The group is working on a conceptual model for authority information and international numbering for authority entities.
FRBR	Functional Requirements for Bibliographic Records. See http://www.ifla.org/VII/s13/frbr/frbr.htm.
FTP	File Transport Protocol. See http://searchnetworking.techtarget.com/sDefinition/0,,sid7_gci213976,00.html.
GBV	*Gemeinsamer Bibliotheksverbund*, GBV. See http://www.gbv.de.

ISO 2146	Documentation—Directories of libraries, archives, information and documentation centre, and their databases. See http://www.nla.gov.au/nla/staffpaper/jpearce2.html.
ISO 2709	Information and documentation—Format for Information Exchange. See http://www.iso.org/iso/en/CatalogDetailPage.CatalogDetail?CSNUMBER=7675.
ISO ILL	Information and documentation—Open Systems Interconnection—Inter-library Loan Application Service Definition (ISO 10160 and ISO 10161) See http://www.nlc-bnc.ca/iso/ill/standard.htm.
MAB	*Maschinelles Austauschformat für Bibliotheken.* See http://www.ddb.de/professionell/mab.htm.
METS	The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language of the World Wide Web Consortium. See http://www.loc.gov/standards/mets/.
MODS	The Metadata Object Description Schema (MODS) is an XML schema intended to be able to carry selected data from existing MARC21 records as well as to permit the creation of original resource description records. It includes a subset of MARC fields and uses language-based tags rather than numeric ones. See http://www.loc.gov/standards/mods/.
NCC	*Nederlandse Centrale Catalogus*, NCC. See http://picarta.pica.nl/DB=2.4/LNG=EN/.
NCIP	NISO Circulation Interchange Protocol (NCIP) is designed to perform the functions necessary to lend items, to provide controlled access to electronic resources and to facilitate cooperative management of these functions. See http://www.niso.org/committees/committee_at.html.
OAI harvest protocol	The Open Archives Initiative Protocol for Metadata Harvesting (referred to as the OAI–PMH) provides an application-independent interoperability framework based on metadata harvesting. See http://www.openarchives.org/OAI/openarchivesprotocol.htm.

OCLC	OCLC, Inc. is a non-profit membership organization serving 41,000 libraries in 82 countries and territories around the world. See http://www.oclc.org/about/.
OCLC PICA	OCLC PICA B.V. European organization of cooperating libraries. See http://www.oclcpica.org.
ONE2	OPAC network in Europe. See http://www.one-2.org/.
ONIX	ONIX is the international standard for representing and communicating book industry product information in electronic form, incorporating the core content. See http://www.editeur.org/onix.html.
OpenURL	The OpenURL is designed to enable the transfer of the metadata from an information service to a service component that can provide context-sensitive services for the transferred metadata. See http://www.niso.org/standards/resources/OpenURL-release.html.
PiCarta	PiCarta is an integrated, multi-material database which contains request facilities and which offers access to online resources and electronic documents. See http://www.oclcpica.org/?id=102&ln=uk.
Publiekwijzer	Publiekwijzer is an information service directed at public library users. See http://www.oclcpica.org/?id=103&ln=uk
SUDOC	*Système Universitaire de Documentation*. See http://www.sudoc.abes.fr.
TEL	The European Library, the gate to Europe's knowledge. See http://www.europeanlibrary.org/.
UCP profile	The Union Catalog Profile is a protocol over the Z39.50 update service. See http://www.nla.gov.au/ucp/.
UNICODE	Unicode is an entirely new idea in setting up binary codes for text or script characters. Officially called the Unicode Worldwide Character Standard, it is a system for "the interchange, processing, and display of the written texts of the diverse languages of the modern world." It also supports many classical and historical texts in a number of languages. See http://whatis.techtarget.com/definition/0,,sid9_gci213250,00.html.

URL	A URL (Uniform Resource Locator) is the address of a file (resource) accessible on the Internet.
VCUC	The Virtual Canadian Union Catalog. See http://www.nlc-bnc.ca/resource/vcuc/.
VIAF	VIAF is a joint project with the Library of Congress and *Die Deutsche Bibliothek*. VIAF explores the virtual combination of the name authority files of both institutions into a single name authority service. See http://www.oclc.org/research/projects/viaf/index.shtm.
WorldCat	WorldCat (the OCLC Online Union Catalog) is the world's most comprehensive bibliographic reference resource, with over 53 million bibliographic records representing 400 languages and covering information dating back to the 11th century and holdings information from libraries in 45 countries. See http://www2.oclc.org/worldcat/.
XML query	XML Query aims to provide flexible query facilities to extract data from real and virtual documents on the Web. See http://www.w3.org/XML/Query.
Z39.50	Z39.50 is an ANSI/NISO standard that specifies a client/server-based protocol for searching and retrieving information from remote databases. It is also an ISO standard ISO 23950. See http://www.loc.gov/z3950/agency/.
Z39.50 Holdings Schema	See http://lcweb.loc.gov/z3950/agency/defns/holdings-1-0.html.
ZDB	*Zeitschriftendatenbank*. See http://www.zeitschriftendatenbank.de/.
ZING SRU	The 'Search/Retrieve URL Service', SRU, is a proof-of-concept initiative to permit the development of value-added search and retrieve applications, such as the scholar's portal, that will integrate access to various networked resources. See http://lcweb.loc.gov/z3950/agency/zing/srw/sru.html.

ZING SRW	The 'Search/Retrieve Web Service', SRW, is a proof-of-concept initiative to permit the development of value-added search and retrieve applications, such as the scholar's portal, that will integrate access to various networked resources. See http://lcweb.loc.gov/z3950/agency/zing/srw/specifications.html.
ZLOT	Z Texas Implementation Component of the Library of Texas. See http://www.tsl.state.tx.us/lot/ZLOTwhitepaperlib.html.