CIC Virtual Electronic Library Z39.50 Project

Technical Issues Associated with the Creation of the CIC VEL Public Catalog

A Discussion Paper on some of the Technical Issues Associated with the Creation of the CIC VEL Public Catalog, including searching, merging, sorting, refining and broadcasting searches across the OPACs of the CIC (2/11/97)

The purpose of this paper is to encourage discussion within the CIC Libraries and between the CIC Libraries for the purpose of developing a common understanding and agreement of the functionality to be provided in the union catalog of the CIC virtual electronic library. This document is a revision of a draft discussed on November 1920, 1996 by the CIC Library Automation Directors.

BACKGROUND

For several years the CIC Automation Directors and others have been discussing the issues surrounding the creation of an effective seamless online public catalog for the CIC user community. All groups have recognized the importance of this goal, while recognizing the significant technological and intellectual problems that impede the attainment of this goal. Not the least of these impediments has been the lack of a clear, shared understanding of the goal and the problems. In late September, Nolan Pope charged Susan Logan (Ohio State) and Nancy John (UIC) to prepare a working paper to move the CIC library participants toward a more commonly shared and more clearly defined statement of the problems and possible solutions. The goals of this assignment were a statement of the technical issues that could be used to work with others (e.g. OCLC) and a growing (and ultimately comprehensive) list of the issues and their proposed solutions.

In his charge, Nolan wrote, "Some scenarios are much more difficult or require either larger desktop machines or multitier architecture to allow the set manipulation to occur on a machine probably at the end user site, other than on the desktop. But if we could determine that some are not feasible in today's environment, or that some are not worth doing because they are too simplistic, we could narrow down the options for everyone to consider." Further, CIC VEL development is being affected by our lack of shared understanding. For example, when one or more sites said that they must have support for textbased (e.g., lynx) clients, OCLC interpreted this to mean that lynx is the CIC standard and represents the highest level of common client functionality rather than the lowest. As a result, we understand that OCLC decided that pursuing HTML version 3 was not possible given the CIC reliance on lynx.

Whether this is the true reason or not, (or is it because of our local system vendors), we feel strongly that we must assert the following: functionality across a wide range of clients should not mean that all clients have the same functionality. Instead, we will define a range of acceptable functionality and then guarantee a minimum level of functionality that must be supported at the lowest level of client, while users of higher end clients or client workstations will have greater functionality.

SEARCHING ACROSS CIC LOCAL OPACs

We need a paper that outlines the impact of the various local systems on the overall compatibility of the CIC catalogs. Simply put, all catalogs are not created equal, even if their component bibliographic records are. What is needed is a paper exploring the role that the following (among other) features play in promoting or undermining compatibility:.

1. local definitions of Z39.50 attributes
2. local options in search defaults, such a Boolean operators
3. local indexing options, including stoplists in use and the maximums that can be set for results sets
4. local thesauri or other local authority standards
5. local library systems implementations of Z39.50, including versions and functions
6. local library system catalog software and its impact on functionality of the various Z39.50 servers.

Once gathered, this data can form the basis of determining the current lowest common denominator across all catalogs. Based on that understanding, we may choose to identify changes or strategies to raise that level. For example, we might choose to develop a set of VELwide options that each site would implement in support of the virtual union catalog.

The technical issues of searching via Z39.50 across 6 or 7 or more separate implementations also must be explored and documented. Ideally, such a paper would form the basis of a primer in the CIC implementation of Z39.50 related to attribute set options, levels of compliance with the versions of Z39.50, and the impact of the our different local systems. In the current environment, the failure of a search to work in a reliable or compatible or comprehensible way is too easily attributed to such excuses as "that's the way NOTIS works," "Innovative is different," "LIAS is a standalone system," which only adds to the CIC folklore, and ultimately undermines the VEL's potential. Worse, of course, is the assumption that a search actually works in a familiar way when it doesn't. If our libraries' staff are to trust the virtual catalog, they must understand when and why it works and when and why it doesn't.

THE ROLE OF THE ARCHITECTURE OF THE CIC VIRTUAL OPAC

The scalability of the virtual catalog raises important and difficult questions. Not the least of the problems stems from the wide range of possible architectures that could be employed to support the virtual catalog. We can argue that we must achieve a common agreement of the functions to be performed by the client software, the server software, and any middleware. Without such a consensus, it is too easy to let technical constraints at one level or another dictate functionality. For example, if we agree that the endusers workstation will do all sorting, the workstation must receive all the output before it can perform this computation unless we further agree that sorting is accomplished as a rolling activity, constantly being updated as new data are encountered. Without laboring this point too much. we need to develop a working understanding of the most desirable place to perform the Searching, Merging, Sorting, Refining and Broadcasting functions in the CIC VEL union catalog. Unfortunately functionality is directly affected by resource constraints, and resource constraints in turn depend upon the placement of certain functions within the architecture, as well as numbers of transactions, size of retrieval, statefulness etc.

Activities like broadcast searching, sorting and merging may require resources beyond those generally expected to be available at the enduser workstation. However, placing activities such as set manipulation (e.g. limiting results by the results of another search) on the middleware or server machines may place a greater burden to maintain stateful contact with the enduser's machine.

BROADCAST SEARCHING IN THE CIC VIRTUAL OPAC

Surely no feature is more attractive and eagerly awaited by the users. On the other hand, no feature creates more of a concern for the technical experts. Broadcast is the ability to send a search simultaneously to two or more remote catalogs or databases. Broadcast searches are the most likely of all the types of searches to encounter the differences in the local implementation of Z39.50 and other system configurations. Some of these are: limits on the number of records that can be retrieved, different indexing policies, differences in the way punctuation is handled, etc…

The challenge we face is being able to manage the technical differences and the sheer increase in traffic that these searches can potentially create. At the very least, we should try to let the users know in which catalogs the best concentration of the resources they are seeking can be found. At the other extreme, where the search has produced potentially very large results, we should post the result only and lead the user to a meaningful restatement of the search problem. We should seek ways to encourage more specific searches when searches are so general they retrieve unwieldy sets of records. In the case where the search is specific enough to handle a full broadcast search of all databases, the software needs to be able to create a meaningful merge of those records and present them in a fashion that is user friendly.

Other issues for broadcasting searches include whether we might develop a preferred order of site to search, or preferred groups of sites, or whether we might broadcast searches against randomly selected catalogs. The ultimate in searching logic would be to develop a truly artificial intelligent interface which could route the searches to the most effective and useful resources without user intervention.

SORTING FINAL SETS IN THE CIC VIRTUAL OPAC

Sort is the logical arrangement of results by some specified variable or variables. Sorting does not imply deduplication. A specific sort may be set as a default and the others made optional. Comparison of dissimilar as well as nearly similar records to identify a preferred order or method of display is required. The provision of secondary sorts which might be under user control are also possible. A minimum level of sorting for basic bibliographic records might be:

1. alphabetically by author (using 1xx) or if lacking, using the 245 field
2. alphabetically by title (245 field) ignoring initial articles (2nd indicator)
3. chronologically (ascending or descending) by date of publication (using the 260 or the pubdate in the fixed field)
4. descending by relevance
5. institutional sort by holding library
6. subsort by series volume or serial volume.

MERGING SET RESULTS

Merge is the combining of results so that duplicates are combined so that information about a particular item is grouped together for display and/or the redundant information is removed, but certain specified local library information (such as location, call number, exact holdings, link to local full bibliographic record) is retained. Merge is accomplished by matching on standard numbers:

1. 035, OCLC number (does RLIN number or others make sense?)
2. 020, ISBN,
3. in some circumstances, 022, ISSN

Such a plan begs issues such as multiple versions (e.g. print and microformat, disk and cassette) on the same record, libraries that have failed to remove incorrect ISBNs, etc.

The first bibliographic record encountered provides the bibliographic record to which the local information of it and other matching records is connected for display. Exception: Ideally, when the user's local library owns the item, that local library's record is the preferred source for bibliographic information. Such display might be by dumping all the records in the same directory/folder, presenting them in a clump, or nesting local information under the bibliographic information.

REFINING SEARCHES IN THE CIC VIRTUAL OPAC

Finally, we need to explore how (and if) users will be able to perform iterative routines on search results. Users will want not only to perform simple tasks such as limiting a search set to only the current year's items. They will also want to combine two sets. In this case, if deduplication, merging and sorting has occurred to either one or both sets, further manipulation of the sets may be extremely difficult unless the system is able to reconstruct the raw search results and then perform the operations requested.

It is not enough to understand Searching, Merging, Sorting, Refining and Broadcasting but we must also explore the effect each has on the ability of the system we design to perform the other activities.

This paper only lays out some of these issues. It is hoped that by making a start, we can begin to explore the issues in greater depth and toward greater clarity of understanding.

Center for Library Initiatives Homepage
CIC Homepage

Revised:March 25, 1998