THE VIRTUAL UNION CATALOG: A COMPARATIVE STUDY

Karen Coyle

ABSTRACT
A Virtual union catalog is a possible alternative to the centralized database of distributed resources found in many library systems. Such a catalog would not be maintained in a single location but would be created in real time by searching each local campus or affiliate library's catalog through the Z39.50 protocol. This would eliminate the redundancy of record storage as well as the expense of loading and maintaining access to the central catalog. This article describes a test implementation of a virtual union catalog for the University of California system. It describes some of the differences between the virtual catalog and the existing, centralized union catalog (MELVYL). The research described in the paper suggests enhancements that must be made if the virtual union catalog is to become a reasonable service alternative to the MELVYL® catalog.

RESUMO
Um catálogo de união Virtual é uma alternativa possível para o banco de dados centralizado de recursos distribuídos encontrados em muitos sistemas de bibliotecas. Este catálogo não seria mantido em um único local, mas seria criada em tempo real, pesquisando cada campus local ou catálogo da biblioteca da filial através do protocolo Z39.50. Isto eliminaria a redundância de armazenamento registro, bem como à custa de carga e manter o acesso para o catálogo central. Este artigo descreve uma implementação de teste de um catálogo união virtual para o University of Califórnia sistema. Ela descreve algumas das diferenças entre o catálogo virtual e o existente, catálogo união centralizado (MELVYL). A pesquisa descrita no artigo sugere melhorias que devem ser feitas se o catálogo virtual é a união para se tornar uma alternativa de serviço razoável para o catálogo ® MELVYL.

THE UNIVERSITY OF CALIFORNIA UNION CATALOG

The University of California, with its nine campuses located throughout the state, adopted the goal of "One University, One Library" in 1977. Under this goal, the resources of these geographically distributed libraries would be treated as a single collection available to the entire scholarly community of the University. The first step toward this goal was the development of a union catalog for the libraries. After early attempts at a book catalog and a subsequent microfiche version, in 1982 the union catalog came into being as an online public access system known as MELVYL®. This centralized database is built from catalog records sent by all the cataloging departments of participating libraries to the California Digital Library (CDL) where the MELVYL catalog is housed. Participating libraries include the California State Library, the Center for Research Libraries, and a number of affiliated institute libraries. In all, there are twenty-nine separate (and diverse) input streams that feed into the union catalog on either weekly or monthly update schedules.

Parallel to the MELVYL database, which contains records for monographs and non-book format materials, is the state of California's union database of serials. This includes serials records for the University of California, the California State Universities, other public and private research libraries such as Stanford and University of Southern California, as well as the union lists of public libraries, law libraries and medical libraries. For this database there are thirty-seven different input streams representing nearly 600 libraries that are updated anywhere from weekly to yearly.

FUNCTIONS OF THE UNION CATALOG

First and foremost the MELVYL catalog is a document discovery tool for end-users. At a time when most other catalogs were limiting users to left-anchored exact heading matching, MELVYL had keyword searching on titles and subjects as well as a sophisticated personal name algorithm that can retrieve an AACR heading based on a variety of user input.

The catalog also turned out to be an important tool for the libraries themselves and was soon incorporated into inter-library loan, collection development and even cataloging functions. One particular aspect of the catalog that has proven to

have added value beyond our original intentions relates to the unique way that records from different sources are merged and stored.

For each unique title we created a merged record that could contain all the uniquely contributed fields by each cataloging source. This means that our underlying record can have multiple 100 or 245 fields, as well as a variety of other USMARC fields. Naturally, we only show the end-user a single view of this record, but all the variant fields in the record can contribute to record access. This means that if a single library adds a subject heading in their own local catalog, when added to the MELVYL catalog that subject heading provides access to all copies of that title.

It also means (and this is the part that we didn't anticipate) that if one campus contributes a full catalog record and another creates only a minimal record, the latter library gains all the functionality of the full cataloging of the former. This became an important side effect of our merged record when libraries were undergoing retrospective conversion, and again when AACR2 necessitated updating large numbers of name headings. While not a substitute for bringing their own local catalogs up to date, at least union catalog users were benefiting from efforts made by any library in the University of California system, and the libraries themselves may have had greater options in terms of where to put their limited resources during those times of change.

Today, all campuses have integrated library systems and the quality of the records in those systems and input to the MELVYL catalog is quite high. Still, we keep finding new uses for the union catalog. Recently it became the basis for a patron-initiated request system that allows cross-library lending with minimal staff interaction in the ILL departments. The periodicals file is linked in rather clever ways to both locally-mounted and remote abstracting and indexing databases so that users can go from a retrieved citation to a list of libraries that carry the periodical title. The union catalog is also becoming the university's catalog of Internet-accessible resources.

THE VIRTUAL UNION CATALOG CONCEPT

The current MELVYL centralized database model was developed nearly 20 years ago, prior to the widespread availability of networks and distributed databases. It was also developed before the participating libraries had OPACs of their own. It runs on a large mainframe computer using locally-developed software, some of which is

actually over twenty years old. Hardware maintenance costs are high and database update and maintenance functions are labor-intensive. As part of an ongoing process of service evaluation, which includes the evaluation of service efficiency and cost effectiveness, studies were initiated to determine if it is reasonable to seek alternatives to the centrally housed union catalog that could achieve many of the same goals of quality user service, 24x7 availability and excellent response time.

One possible alternative to the central database is a virtual union catalog. Such a catalog would not be maintained in a single location but would be created in real time by searching each local campus or affiliate library's catalog through the Z39.50 protocol. This would eliminate the redundancy of record storage as well as the expense of loading and maintaining access to the central catalog. A distributed catalog makes obvious sense in our current environment where every library has its own database and retrieval interface. The wide-spread use of Z39.50 and its implementation in nearly all modern library systems means that there should not be major technological barriers to a distributed solution. Or, so it seems.

DISTRIBUTED CATALOGS

We are hardly the first to think of, much less implement, a distributed catalog solution. Even just over the last year this technology has moved from the "gee whiz" to the "of course" stage. The ability to send queries to one or more other library catalogs is a regular feature, although the actual implementation details vary. Consortia similar in characteristics to the University of California are actively using this technology. The Committee on International Cooperation (CIC), for example, has published a report of their experience with their virtual catalog implementation.

The CIC report expressed some dissatisfaction with the virtual catalog as a discovery tool and advocated more development on the part of library automation vendors. Some of the problems that they encountered were foreseen, such as inconsistency in results between catalogs that had defined their indexes differently, but there was no way to quantify the dis-ease that the librarians at these institutions experienced. The University of California is unique in that we have a current centralized catalog so we can compare the results between these two catalog technologies.

TEST GOALS

Although some of the advantages and disadvantages of the virtual union catalog approach could be anticipated without a test implementation, there is nothing to compare to actual experience with a new technology. And many of the results presented here would not have been foreseen by study of other systems that have attempted the same design.

There are no absolute measures of OPAC effectiveness that we could use to evaluate the virtual union catalog, but because we do have a centralized union catalog, we are able to make comparisons between the MELVYL catalog, with which we are familiar, and a virtual union catalog. We expected there to be many differences, so the goal was not to rate the virtual union catalog against MELVYL but to describe the differences and determine if the virtual union catalog could provide a reasonable service alternative to the MELVYL catalog.

TEST METHODOLOGY

Participants

Campus main libraries were contacted and given the opportunity to volunteer to participate in the comparison. For this test, no attempt was made to cover all campus input sources (affiliated libraries, special libraries) or non-UC sources. The assumption was that the "main" library was the best target for our purposes.

Six campus libraries chose to participate. These included three different library systems: Innovative Interfaces (four sites), DRA WebCat, and OCLC SiteSearch (one site each).

Catalog Search Capabilities

A preliminary analysis of the search capabilities for participating systems was performed. Not only did we have three different "brands" of library system to connect to, it's also the case that Z39.50 search capabilities are not the same as local OPAC search capabilities. Indexes available via Z39.50 for the six participating libraries are listed in Figure 1. Note that differences occur not only between library

system "brands" but also within different installations of the same vendor system due to configuration choices made by the libraries.

Figure 1 – Fields Available Through Z39.50

FIELD	SiteS	DRA	III-1	III-2	III-3	III-4
Author	X	X	X	X	X	X
Call number	X					X
Computer systems	X					
Conference name	X					
Corporate name	X					
Dewey number		X				
Edition	X					
Genre	X					
Geographical access	X					
ISBN	X	X	X	X	X	X
ISSN	X	X	X		X	X
Keyword		X	X	X	X	X
Language	X					
LCCN		X	X			X
Local number	X		X		X	
Local subjects	X					
Music publisher number	X					
Notes	X	X				
OCLC number			X			X
Other call numbers		X				
Personal name	X					
Place of publication	X	X				

Publication date	X					
Publisher	X					
Record type	X					
Subject		X	X	X	X	X
Subject heading	X					
Title	X	X	X	X	X	X
Title, series	X	X	X			
Title, uniform	X	X	X			
URL	X					

It was actually more difficult finding indexes common to all of the participating systems than we had anticipated. Some of the systems had an overall keyword search that combined keywords from a range of access fields but did not do keyword searching on individual heading types. Yet two of the systems (MELVYL and the SiteSearch implementation) did have index-specific keyword searching (title keyword, subject keyword) but were lacking a general keyword search analogous to the others. When we included author searching, we knew that we were going to see a great variation in how those systems processed queries.

In the end we selected:

- author
- exact title (e.g., title heading, left anchored, with truncation)
- keyword

To test the keyword search, we had to simulate it on the two systems that didn't have that index by searching a combination of title words and subject words. As anticipated, the results were not easily comparable.

So before even beginning our test, we had had to limit ourselves to catalog search functions that would provide only a minimum of searching capabilities for known item and subject access. This, in itself, was an interesting lesson in distributed

searching and we went into the test phase with even less hope that we would be able to show that the virtual union catalog could be a viable public service tool.

The Search Queries

We wanted to see how the virtual union catalog stood up under real user queries. To get these queries we selected a single file, representing about one day's searching, from the MELVYL search logs. This included all commands that were issued to the system during that time span, so although we started with many tens of thousands of log entries, in the end we had a rather small set of viable searches. We selected only those searches that represented the three indexes we wished to test. Of those, we eliminated the searches that received a zero result. This gave us a set of searches that we knew would retrieve some records on the MELVYL catalog. In a later step, we also removed searches that did not get at least one hit among the six libraries that were part of the study, since our actual study group was only a subset of the MELVYL coverage.

TEST 1 - RECORD "EXPLOSION"

One of the questions we needed to answer on the prospect of moving from a centralized catalog of merged records to distributed catalogs has to do with the total number of records that would be retrieved through the distributed method. The MELVYL catalog has about 10 million titles representing 18 million "copies", but about 2/3 of the catalog is made up of records with only one holding. The other 1/3, or roughly 3 million records, account for the other 12 million holdings.

Because we can limit a MELVYL search to an individual campus, we were able to crudely approximate the effect of searching each catalog separately by running our test queries nine times, each time limiting the results to a single campus. We could then compare the total of these searches against the total retrieved in the merged database. The results varied based on the index used, but were not greatly different from what one would expect from the overall catalog composition:

- Author search: separate campus searches retrieved 1.98 times the merged search
- Exact title search: 2.33 times the merged search
- "Keyword search" (title keyword OR subject keyword): 1.79 times the merged search

Test 2 - Searching Against Campus Z39.50 Servers

After removing any searches that returned zero results in Phase 1, the same searches were run against the six campus catalogs through their Z39.50 server function. This was done using an automated search program, and we were pleased that the results were generally quick. We did have to be careful not to overload the campus servers, because our search engine was going against their public catalog and could potentially send searches fast enough to negatively affect actual users. Fortunately, by now our searches had been reduced to short lists and we didn't have problems.

A sample of results is given in Figure 2. For each search, we knew how many items were retrieved when the search on the MELVYL catalog was limited to that campus' holdings. These were then compared to the numbers retrieved from that campus' online catalog. A zero in any column means that the results were the same; positive means that more records were retrieved using Z39.50 against the campus catalog, and negative means that fewer records were retrieved from the campus system than from MELVYL. Most notable about these results is the lack of any consistency between the MELVYL retrievals and the local system retrievals. Within the same library system, some searches will retrieve many more items than the same search on MELVYL, and some will receive many fewer.

We had expected there to be differences -- explainable differences -- the results of this test exhibited a much wider range of variation than we had anticipated. However, the numbers alone were only an indication that there was something there worth investigating.

Figure 2 – Comparison of Searches, Z39.50 vs. Union Catalog
Single names

	SiteSearch	DRA	III-1	III-2	III-3	III-4
ABBEY	-12	129	-2	-2	-2	4
AURELIUS	307	-155	-211	-213	-197	-313
BARTSCH	-80	7	2	4	4	38
DAVIES	-663	-60	-129	-137	-126	113
HAND	462	33	735	1163	868	1973
KASKEL	-4	-1	-1	-1	-1	-2
MOWAT	-6	0	3	0	2	8

Names with initials

	SiteSearch	DRA	III-1	III-2	III-3	III-4
BRITTEN, J	-4	-11	-1	-2	-1	-2
BRITTEN, JAMES	17	-6	0	-1	-1	-1
GOULD, J	-192	-150	-36	-36	-34	-51
HAIN, L	-11	-5	0	-4	0	-2
J.RUBINSTEIN	-11	-16	-17	-12	-3	-18
RICHARDS, J	-126	-177	-43	-40	-37	-58
ROSENBLUM, L.A.	-10	-10	-10	-10	-11	-11
SEBER, G.A.F.	-9	-8	0	0	0	2
PRAUSNITZ, JM	-9	-8	-8	-6	-3	-8

Order of names

	SiteSearch	DRA	III-1	III-2	III-3	III-4
IMMANUEL KANT	115	-146	-145	-121	-113	-191
LANGSTON HUGHES	19	-91	-64	-64	-86	-103
CHEN YI	-82	-44	35	9	-4	-12

Different forms of the same name

	SiteSearch	DRA	III-1	III-2	III-3	III-4
BEAUVOIR	14	2	1	4	2	17
BEAUVOIR SIMONE	21	-1	2	5	3	15
IMMANUEL KANT	115	-146	-145	-121	-113	-191
KANT	1124	12	263	252	193	357

TEST 3 - QUALITATIVE ANALYSIS OF SEARCH DIFFERENCES

There seemed to be no pattern or consistency to the search results we had received. To understand why, a group of campus librarians (see acknowledgements) undertook to analyze the differences. They did this by manually repeating a selection of the test searches and looking at the resulting retrievals.

What they found, as you might have guessed, was that just about every imaginable difference that could occur between library catalogs did indeed manifest itself in our sample.

Author Searching

As expected, author searching turned up numerous reasons for differences in results. The format of author names as input in USMARC records is rigorously standardized. What isn't standard is how our systems index those names, nor how

library system user interfaces deal with the variety of name forms that users will input at the query line.

1. **Order of elements**: Some systems require the author's name to be in last name, first name order. Of those, some require that the comma be present and others do not. Queries that do not follow the prescribed order get zero results. While a friendly user interface can detect queries that are missing the comma, searches through Z39.50 tend to simply reply with no retrievals.

Query:	MARY BEARD	0
	BEARD, MARY	3

2. **Truncation**: Some systems treat the author name as an exact string and truncate, so the entry of SMITH, J will retrieve SMITH, JOHN. Others require the full name to be in the query, in which case SMITH, J only retrieves headings that read Smith, J. However, truncation of the heading produces some unexpected results, such as when a search for KANT retrieves KANTOR, KANTWELL, etc. In these instances, local systems retrieved more items than the MELVYL catalog, which treats author names as complete words.

3. **Single name queries**: Given a single name, some systems assume this to be a truncated left-to-right search while others perform a keyword search. So, a search on AURELIUS will retrieve only those authors whose last name begins with AURELIUS, or it will retrieve all authors with the name AURELIUS anywhere in the author heading.

4. **Initials in author names**: MELVYL uses a complex personal author search that retrieves authors whose forenames have the same initials as those in the search. So, a search on SMITH, J J will retrieve Smith, John James, as well as Smith, J. James or Smith, John J. Systems that simply truncate authors searches or that treat each term as a keyword will produce very different retrievals

5. **Author vs. personal name**: Like other indexes, differences can arise because of library choices made in the configuration of the index in the local system. One

choice that can make a great difference between systems for author searching is whether the index is limited to authors or whether it includes personal names occurring in subject fields.

6. **Corporate authors**: Author indexes can include corporate as well as personal authors, and even conference headings.

Title Searching

Exact title searching should yield fairly consistent results. All of the catalogs are referring the query to a heading index and are searching from left-to-right. But even within this limitation, differences arose.

1. **Title index contents**: Not unexpectedly, there are differences in the fields that contribute to the title indexes in the various systems. In our sample, title indexes varied from containing only the 245 title to ones that had the full range of 2XX fields, $t subfields, series titles and titles from subject headings.

2. **Record types**: One of the changes desired for the MELVYL catalog is to combine the monographs and serials records into a single union catalog. Most library systems today have a single database for all item formats. Some allow users to limit a search by type, such as books or sound recordings, and others do not.

3. **Query truncation**: Most systems truncate exact title searches, and automatic truncation was added to each exact search via Z39.50. It appears, however, that some systems add a blank before the truncation symbol and others do not. So the search "F XT VOICE" could be interpreted as "F XT VOICE #" or "F XT VOICE#". The latter would retrieve any title beginning "Voices" but the former would not.

4. **Key length**: All systems have some limit on the length of their indexes. Where these limits differ, access is affected. In our sample, searches that exceeded the key length of the local system returned zero as their result.

Keyword Searching

Comparison of keyword searching between the MELVYL catalog and the local catalogs via Z39.50 is of limited accuracy because MELVYL does not have a keyword index that combines words from a wide selection of fields. We included this search, however, because we had no other way of testing a subject search; many of the systems did not have a subject index and we felt that it was important to include subject searching in our test.

1. **Index contents**: The main differences in results in this area come from differences in the design of Keyword indices in the various systems. Participating libraries used different fields when creating this index, which led to very different search results.
2.
3. **Individual fields vs. keyword "pool"**: Although it may seem logical that one could simulate a Keyword index with a Boolean "or" of similar searches of the same fields, this does not produce an equivalent result in many cases. The terms in the keyword indexes are generally a pool of terms from all the fields that are included in the index, so a search on two terms retrieves those terms even if they originated in different fields. The Boolean "or" of keywords from separate fields may not have the same ability to cross field boundaries when more than one term is included. For example, a search on Title Word SOLAR ENERGY or Subject Word SOLAR ENERGY still requires both SOLAR and ENERGY to be from the same heading and will not retrieve a record with SOLAR in one field and ENERGY in the other.

All Searching

Other differences that we found between local systems and MELVYL weren't particular to a specific type of search. Among these were:

1. The MELVYL AT command includes all libraries on a campus. On many campuses, there are multiple online catalogs, such as those at law libraries, that are not included in the main library catalog. This test did not include all of the catalogs at the campuses that participated.

2. At the same time, some campus catalogs have catalog departments whose records are stored separately in the local catalog, creating deliberate duplicates. These are often merged in the MELVYL catalog.

3. The records in a MELVYL merged bibliographic record are all enhanced with the added access that any one record contains. When searching the campus catalogs through Z39.50, only the library that contributed the heading from their own local catalog benefits. Some of the more mysterious differences between MELVYL retrievals and campus catalog retrievals were due to access points added by another cataloging unit.

REQUIREMENTS FOR A VIRTUAL UNION CATALOG

The scope of this project was not sufficient to provide a full test of functional requirements for a virtual union catalog, but some important general areas have been identified which would require further analysis and testing prior to planning for the production use of this architecture.

Database Consistency & Search Accuracy

For a virtual union catalog to be feasible, the participating databases must offer a uniform set of indexes and search functions that retrieve comparable items from each catalog. In the current environment, it is not possible to formulate a search that

yields predictable results from the databases. Evidence of this lack of consistency and its affect on search accuracy and predictability is a significant result of this test. This means that the first step in creating a virtual union catalog is to create compatible local catalogs that are designed to support the distributed environment. It appears that a common use of Z39.50 in libraries today is not a distribution of our catalogs but a kind of harvesting in disparate databases. While this is an obvious statement of fact, we still seem to harbor a somewhat illogical hope that this harvesting will inexplicably yield consistent and accurate results.

System Availability

The MELVYL union catalog serves the entire University of California community as well as the larger research library community. It is essential that the catalog be available as close to 24 x 7 as possible. As part of a virtual union catalog, local system downtime, scheduled or unscheduled, would impact the availability of the catalog as a whole.

Capacity Planning for Campus OPACs and the Network

The development of a virtual union catalog design would have important implications for local system search capacity and network load. Each search that is now directed only to the centralized union catalog would instead be broadcast to all of the campus catalogs and potentially all contributing systems. Local campus systems would each need to be able to respond to an additional 300,000 searches per week, based upon current MELVYL catalog activity. Network capacity planning would be required to accommodate the increased bi-directional traffic between the libraries.

Sorting, Merging and Duplicate Removal

Searches issued against the union catalog retrieve a set of records that have been merged to eliminate duplicate bibliographic records and sorted prior to input into the database. Broadcast searches return a set of records without merging or sorting. Version 3.0 of the Z39.50 protocol includes a sort function but few systems currently support this feature. Even with that sort in place, the union catalog interface would have to merge the retrieved sets as well as remove duplicate bibliographic information while

maintaining individual holdings data. Because searches across our libraries often retrieve large result sets, sorting and merging is expected to be technologically challenging.

KAREN COYLE
California Digital Libary
E-mail: karen.coyle@ucop.edu