Karen Coyle,
The Virtual Union Catalog

from / *aus*:

Union Catalogs at the Crossroad
Edited by
Andrew Lass and Richard E. Quandt
pp. / *S*. 51-66

Erstellt am 31. März 2005

Impressum

Bibliographic information published by Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.ddb.de.

Bibliografische Information Der Deutschen Bibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.

This publication is also openly accessible at the publisher's website. Die Deutsche Bibliothek has archived the electronic publication, which is now permanently available on the archive server of Die Deutsche Bibliothek.

Diese Publikation ist auf der Verlagswebsite ebenfalls open access *verfügbar. Die Deutsche Bibliothek hat die Netzpublikation archiviert. Diese ist dauerhaft auf dem Archivserver Der Deutschen Bibliothek verfügbar.*

Available open access / *open acess* verfügbar:

Hamburg University Press / Hamburg University Press
http://hup.rrz.uni-hamburg.de
Die Deutsche Bibliothek archive server / Archivserver Der Deutschen Bibliothek
http://deposit.ddb.de/

ISBN 3-937816-08-9 (print)

Table of Contents

Part 6 South African Union Catalogs

Chapter 2
The Virtual Union Catalog

Karen Coyle[1]

1 Introduction

Some library consortia have chosen to implement a 'virtual' union catalog through broadcast searching of the catalogs in their consortium. This is generally a less expensive solution than the creation of an actual union catalog database that must receive and store records from each of the library systems. In most cases it is not possible to do an evaluation of the effectiveness of these two solutions, and therefore a cost-benefit analysis is not available to library administrators who are attempting to make a decision about what type of union catalog best serves their users. Because the University of California had both a centralized union catalog (MELVYL®) and a number of contributing systems that were accessible through the Z39.50 search protocol, we were able to do a direct comparison of the retrievals between the union catalog and its 'virtual' equivalent. The study showed that the two union catalogs were far from equivalent, and that broadcast searching across disparate databases produces highly inconsistent results.[2]

The University of California is a system of nine (soon to be ten) campuses that span the state of California from Davis, in the north, to San Diego, at the Mexican border—a distance of 800 kilometers. The campuses

---

[1] California Digital Library, http://www.cdlib.org, http://www.kcoyle.net.

[2] Details of the results of this study were published in D-Lib Magazine in March 2000. See http://www.dlib.org/march00/coyle/03coyle.html.

combined have a student enrollment of 160,000, with 10,000 full faculty members and over 130,000 staff and teaching personnel. The campuses function fairly autonomously for most of their academic activities and their administration, although sharing among the libraries is encouraged and well-supported.

The university was founded at Berkeley in 1873 and the Berkeley library is still the largest in the system, with about nine million volumes. The next largest library is Los Angeles, with 7.5 million volumes. The total number of volumes in the 9-campus system is 31 million. There are at least 200 libraries in the system, although this number does not count the many departmental or faculty libraries. Each library has its own unique characteristics. The library at UC San Diego has made agreements with the University of Beijing to receive full-text copies of millions of volumes of its holdings and to make them available to scholars in the United States. The library at Los Angeles has one of the world's largest archives of films, and now serves as an archival agency for some of the top Hollywood studios. Berkeley's rare books room houses the Mark Twain papers; Santa Cruz has an excellent collection of California poetry; Riverside collects contemporary science fiction.

In the mid-1970s the university was seeking ways to make the library's collections more widely available to students and faculty at the various campuses. It was not unusual for a scholar to travel from one campus to another to take advantage of the library collections. The hard part, though, was knowing what you would find there. There was no central catalog for the libraries, so it was necessary to go to the library and consult the card catalog to determine what materials were available. Clearly a union catalog would greatly facilitate the sharing of collections.

Work on a union catalog began in the late 1970s. The first union catalog was a book catalog created from copies of cards contributed by each campus. Before this catalog was completed, a new resource became available: machine-readable records from OCLC, whose card-production service was used by most of the campus libraries. By 1980, the university had produced a microfiche catalog of current cataloging from all nine campuses. But technology was moving forward at a rapid pace, and the key element to delivering machine-readable data directly to the libraries was falling into place: computer networking. The union catalog became a project

of the university libraries that not only created one of the first online catalogs, but also established the first telecommunications network that connected the University of California campuses.

2 The MELVYL Union Catalog

I shall begin by reviewing the situation in 1980, when work began on the University of California's union catalog. There were no online catalogs available commercially for libraries; each of the UC libraries maintained a card catalog with cards obtained through the services of OCLC or RLG. Libraries had begun using these services in the mid- to late 1970s, and thus there were machine-readable records for this period only. The libraries did not receive copies of their machine-readable records because they had no use for them. The MELVYL union catalog would therefore serve a dual purpose: it would be a public access catalog for library users, and it would be the archive of machine-readable cataloging for the libraries. Indeed, when the libraries later developed or purchased new systems, those first systems were often created with records exported from the union catalog.

This dual purpose led to a unique design for the union catalog. Where other systems, such as OCLC, kept a single copy of the bibliographic data and added library holdings to this record for additional contributors, the MELVYL developers were obliged to keep all the bibliographic data from the contributing campuses, not just the holdings. Yet they did not want to show a separate record for each campus, since the repetition would be difficult for catalog users. Instead, the design called for a single bibliographic record, with multiple holdings where libraries held copies of the same item. Using an algorithm to determine when incoming records represented the same work, records were then merged into a single record with multiple holdings, but with no loss of bibliographic data. To do this, a composite record was developed based on the USMARC format, but extending it to allow each field to be stored with a digital flag indicating which campus had contributed it.

Sample record:

```
100 1   Twain, Mark, $d 1835-1910
        <LC,IG,SDG,LAG,DG,BG,SC,SB,HAST>

240 10  How to tell a story. $f 1996 <BG>

245 10  How to tell a story, and other essays / $c Mark Twain
        foreword, Shelley Fisher Fishkin ; introduction, David
        Bradley ; afterword, Pascal Covici, Jr.
        <LC,IG,SDG,LAG,DG,BG,SC>

245 10  How to tell a story, and other essays / $c Mark Twain ;
        foreword, Shelley Fisher Fishkin ; introduction, David
        Bradley, afterword, Pascal Covici, Jr. <SB>

260     New York : $b Oxford University Press, $c 1996.
        <LC,IG,SDG,LAG,DG,BG,SC,SB>

300     lix, 233 p., 29 p. : $b ill. ; $c 23 cm.
        <IG,SDG,LAG,DG,BG>

300     lix, 233, 19 p. : $b ill. ; $c 23 cm. <SC>

300     lix, 233, 29 p. : $b ill. ; $c 23 cm. <LC,SB>

490 1   The Oxford Mark Twain <LC,IG,SDG,LAG,BG,SC,SB>

490 1   Oxford Mark Twain <DG>

500     Facsimile reproduction of the first American ed.,
        published New York, Harper & Brothers Publishers, 1897.
        <SDG,LAG,DG,BG>

500     Originally published: New York : Harper & Brothers
        Publishers, 1897. <LC,SC>

500     Facsimile reproduction of the first American ed.,
        published New York, Harper & Brothers Pub., 1897. <IG>

504     Includes bibliographic references.
        <LC,IG,SDG,LAG,DG,BG,SC>

504     Includes bibliographic references <SB>

505 0   How to tell a story -- In defence of Harriet Shelley –
        Fenimore Cooper's literary offences -- Travelling with a
        reformer – Private history of the "jumping frog" story --
        Mental telegraphy again – What Paul Bourget thinks of us
        -- A little note to M. Paul Bourget. <IG,SDG,LAG,DG,SC>

650 0   Storytelling <LC,IG,SDG,LAG,DG,BG,SC,SB>

700 1   Fishkin, Shelley Fisher <SB>

700 1   Bradley, David <SB>

700 1   Covici, Pascal <SB>

752     United States $b New York $d New York $9 (1996) <BG>

800 1   Twain, Mark, $d 1835-1910 $t Works. $f 1996.
```

```
        <LC,IG,SDG,LAG,DG,BG,SC,SB>
650 0   Storytelling <LC,IG,SDG,LAG,DG,BG,SC,SB>
700 1   Fishkin, Shelley Fisher <SB>
700 1   Bradley, David <SB>
700 1   Covici, Pascal <SB>
752     United States $b New York $d New York $9 (1996) <BG>
800 1   Twain, Mark, $d 1835-1910 $t Works. $f 1996.
        <LC,IG,SDG,LAG,DG,BG,SC,SB>
```

One can see from this example that there are multiple versions of many fields with either significant or minor variations (such as the 490 field). There are also fields that were contributed by only one of the libraries, such as the 700 fields contributed only by UC Santa Barbara (<SB>), and the 752 field contributed only by Berkeley (<BG>).

This very complex MARC-like record stayed in the background, and the user of the catalog saw a normal bibliographic display and consolidated holdings:

```
Twain, Mark, 1835-1910.

How to tell a story, and other essays / Mark Twain ;
        foreword,Shelley

Fisher Fishkin ; introduction, David Bradley ; afterword, Pascal
        Covici, Jr. New York : Oxford University Press, 1996.

Series title: Twain, Mark, 1835-1910 Works. 1996.

        HAST 5th Stks PS1322 .H6 1996

        UCB Bancroft PS1322 .H6 1996 Mark Twain Papers *c2 copies

        UCB Main PS1322 .H6 1996

        UCD Shields PS1322.H692 1996

        UCI Main Lib PS1322 .H6 1996

        UCLA EngReadRm PS1322 .H6 1996 Main Reading Room ERRREAD-
        STAX

        UCLA YRL PS1322 .H6 1996 Stacks URLSTAX-STAX

        UCSB Main Lib PS1322 .H6 1996

        UCSC McHenry PS1322 .H65 1996

        UCSD SSH PS1322 .H6 1996
```

Only one contributed record was designated the display record; the other records remained hidden from view. However, these other records did contribute to the indexes for the record group. This meant that if one campus had added a unique field, such as the author fields contributed by UC Santa Barbara in the example, a search on that heading brought up the entire group even though other libraries had not included that heading. The merged record became a kind of 'super-record', combining the bibliographic efforts of the whole UC system.

The 'super-record' also had some additional advantages that we had not considered when we were developing the catalog.

The 1980s and early 1990s were given over in many U.S. libraries to the retrospective conversion of their card catalogs to machine-readable form. Libraries were developing online catalogs but only had records dating from their first use of card services like OCLC's. The entire back file of their card catalog had to be converted to MARC records so they could have a complete catalog online. This retrospective conversion was expensive and time-consuming, and in addition was very prone to error. Libraries sent their card catalogs away to be keyed in factory-like settings, and then had to check and correct the records received. Because full-level cataloging for many titles was not available in machine-readable form, some libraries chose to have only minimum-level records created as a way of saving money. This retrospective conversion effort added tens of millions of titles to the OCLC database, however, and collectively the U.S. libraries created the largest storehouse of full cataloging in machine-readable form.

The University of California libraries undertook retrospective conversion at different rates and using different services. Some created mainly full-level records, others were only able to create minimal records for much of their collection. And this is where we discovered a hidden feature of our system's design: as long as one library in the system contributed a record with full cataloging, all others could do a minimal record that would merge with the full one and gain the advantage of the full record in the union catalog. Eventually, most minimal-level records were upgraded by the libraries because they needed full records in their own integrated library systems, but the creation of minimal-level records allowed the libraries to close their card catalogs in a timely fashion and gave them another decade to complete the work of transforming their catalog. At the time of the study reported

here, retrospective conversion was essentially complete and the union catalog held merged records for about 10 million titles, which represented 18 million contributed library records.

3 The Virtual Union Catalog

By the early 1990s, each library had its own integrated library system (ILS), and therefore its own online catalog. The systems in place represented three different vendors and a variety of versions among those vendors. These local catalogs fed records directly into the union catalog to create a union copy of the cumulative holdings of the campus library systems. Nearing the year 2000, most of these local catalogs had Z39.50 capability which would allow external systems to send queries to their databases and receive search results. The MELVYL system had developed the capability of broadcasting searches to multiple databases simultaneously and bringing back results for users. So it became logical to ask ourselves: could the union catalog be replaced by a virtual union catalog, that is, a broadcast search across the very same local catalogs that were contributing to the union catalog? It seemed logical to assume that the results of a broadcast search would be the same as the results of a search of the same records in the union catalog. And if that was the case, then a virtual union catalog might be able to replace the current centralized MELVYL database, with a potential cost saving to the University.

We organized a test of this theory. We began with a set of real searches from the logs of union catalog activity. We knew that these searches had retrieved items in the union catalog. We then needed to find out how many records these searches retrieved for each contributing library. Our system allows us to limit our searches by library, so we reran the queries for each library to get the number that we would compare to the retrievals using Z39.50 against their own database.

A retrieval in the union catalog resulting in one record that was a composite of contributions for three libraries would, of course, get no records for the other six libraries, so we needed to create a set of searches for each library that got at least one retrieval in the union catalog.

Although we would have liked to include a wide variety of indexes in our study, it was difficult to find even a small number of indexes that were common among the 6 systems that we would be searching. Many systems had a 'keyword' index, but the fields included in this index varied between systems, and MELVYL lacked this field altogether. Some systems allowed only left-to-right searching on certain fields, while others treated those fields as keyword searches. In the end we settled on three indexes:

- Author
- Title (left-to-right search, with truncation)
- Keyword (a combination of title and subject keywords where the system did not have the index).

We then wrote a script that took the searches for each library and sent them as Z39.50 queries to the library's online catalog. The results were logged for further analysis.

Results

We fully expected to find some differences in search results based on the unique qualities of the union catalog, in particular the cumulative effect of the merged campus records with the combined retrieval of their headings. In fact, the resulting differences were much greater than we had anticipated, and only a few of them were related to the merging of campus records in the union catalog. Instead, the differences were related to how indexes were structured in the local systems that would make up the virtual union catalog, and the particulars of how searches were performed in the different systems.

To illustrate the flavor of the degree of difference between the search results, consider Table 1, which has positive numbers where the local system returned more records than the union catalog, and negative numbers where the local system returned fewer. Each column represents a library that was queried (L1, L2, etc.):

Table 1. Author Searches

Search string	L1	L2	L3	L4	L5	L6
ABBEY	-12	129	-2	-2	-2	4
AURELIUS	307	-155	-211	-213	-197	-313
HAND	462	33	735	1163	868	1973
BRITTEN, J	-4	-11	-1	-2	-1	-2
BRITTEN, JAMES	17	-6	0	-1	-1	-1
IMMANUEL KANT	115	-146	-145	-121	-113	-191
LANGSTON HUGHES	19	-91	-64	-64	-86	-103

The searches represent a variety of search types, even though they all are searches on author names. The first three were given just a single name, presumably the family name of the author. The next two are searches in which the family name is given first and is distinguished by the use of the comma; this is followed by a forename or initial. The last two show authors being searched in the form they might appear on a book cover. All of these are legitimate searches on the part of the catalog user.

What caused the differences in search results? After all, these same records are in local catalogs and in the union catalog, so the search results should be nearly identical in both.

Consider the three searches where only a single word was input. In the case of the word ‘aurelius,’ this generally retrieved fewer records in the local catalog than in the union catalog. In the case of the word ‘hand,’ the results were uniformly greater in the local catalog than in the union catalog. Yet both were single word searches against an author index. The explanation for the results in the ‘aurelius’ search is that the union catalog performs a keyword search on author names and therefore ‘aurelius’ retrieves records where that keyword also matches a forename. The local

systems almost uniformly do their searching in a left-to-right manner against a name index that places the family name before forenames:

```
Aurelius, Marcus
```

Thus they would not retrieve a record where the author was

```
Adeodatus, Aurelius
```

which was retrieved by the union catalog.

The 'hand' search is an example of the effects of automatic truncation. In some systems, the author search was automatically truncated so that the search on 'hand' became a search on any name starting with the four characters 'hand.' So this search would retrieve

```
Hand, Jacab
Handen, Max
Handers, May
Handschmidt, Frieda
```

etc. It is not always possible to turn off this truncation in searches and it greatly increases the number of records that any search retrieves.

In the systems that do this truncation it would have also been done for the searches on 'aurelius', yet that produced many fewer 'extra' results. The reason is obvious: fewer words have 'aurelius'as beginning characters than have 'hand.' But the difference in retrieval for these two searches is significant, and we can assume that these nuances are not at all understood by users of our catalogs.

Now let us look at some title searches:

Table 2. Title Searches

Search String	L1	L2	L3	L4	L5	L6
THE PROCESS	-573	75	289	276	177	392
THE SOCIAL ANIMAL	-7	1	2	0	3	1
THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION	-3	-1	0	0	0	1

Search String	L1	L2	L3	L4	L5	L6
VOICE	-566	1262	448	497	461	763
WEBSTER'S DICTIONARY OF SYNONYMS	-2	0	0	0	0	2

These searches were all done on a left-to-right title index in each system. One source of differences in these results was how the system applied truncation. A system can truncate directly after the last character in the query:

```
Voice#
```

Or it can add a space and then truncate, creating a word break:

```
Voice #
```

The first search will retrieve both “Voice of the Master” and “Voices of Our Children.” The second search will retrieve only “Voice of the Master.” Truncating at a word break is often used where truncation is applied automatically by the system after the query is completed by the user. The logic is that few users type in a query that stops in the middle of the last word. What users actually type, of course, has to do with the training they have been given on the system and their experience with results.

Another difference in the title searches resulted from the treatment of articles at the beginning of titles in the indexes, and again at the beginning of queries typed in by users (searches 1–3, above). The MARC21 record considers articles at the beginning of a title to be ‘non-filing’ and these are generally ignored in indexing. So the title “The Magic Mountain” is indexed and filed under “Magic,” not “The.” Users, however, may not always know when to drop these articles in a query. Some of the more clever systems look for the most common of initial articles and remove them from a query if the user includes them. This is imprecise, but it does help some searches which, although they are exact transcriptions of the title, will fail because the user did not know to remove the initial article. In our study, library ‘L1’ clearly was not treating initial articles the way they were treated by the other systems.

We also ran into differences relating to the length of the key that was used for the title index. All systems have some limitation on the length of the title key, but the exact size of the key varies between systems. A longer key means more precision for the user, but it also means more storage for the system. A system with relatively short title keys will retrieve more records for some queries, some of which will be false hits. If the retrieval is not overly large, this merely means that the user must work through some undesired records. If the libraries taking part in the virtual union catalog are large, however, these results could overwhelm the user with unwanted records and make finding the desired records very tedious.

The results of title searching were more consistent with the union catalog, at least in some instances, than the author searching results, and where they differed they tended to retrieve more records in the local library system, whereas the author searches often retrieved fewer. Still, in some instances the differences were significant.

Another source of great differences between systems has to do with what fields have been chosen to populate indexes. Although it may seem that we all know what we mean by 'author' or 'title,' in fact our systems demonstrate that we have taken quite different paths in creating those indexes for our systems.

The MARC21 record that is used in the United States has numerous fields that could be considered titles. There is the title proper, in the MARC 245 field, and there are fields for variations on that title. If a serial, the document may have one or more abbreviated versions of its title. If a monograph, there could be a series title, or two. Any items that have multiple parts, such as a music recording with a variety of pieces, can have titles relating to each of those parts. And other special publications such as conference proceedings have titles for the event as well as the publication. So you can expect that a title index may reflect a wide variety of choices on the part of the librarians who set up that particular system.

The keyword index is equally variable. Some systems index absolutely every possible field in their keyword index, and it is usually understood to be something of a catch-all field, although in America the phrase used is the pejorative 'kitchen sink.' Some system developers may consciously exclude fields that introduce 'noise' but are rarely useful for retrieval, such

as the general notes field. Finding two systems with the same selection of fields in their keyword index would be difficult.

Subject, a search that nearly all systems include, is also problematic. There are the regular subject fields, but there are also fields that have a subject role, at least in the minds of some. The MARC21 record has fields for geographic area covered by the text, for the genre of the item (bibliography, electronic archive, etc.), and additional fields for special collections such as the book binder or the provenance of the item. Are these to be included in the subject index? If not, there may not be another index in which to put them.

Of all fields, it would seem that we share the meaning of the term 'author.' If only that were so. To begin with, there is the question of those authors who are not persons; corporate bodies, institutions and conferences; all have authorship roles that may be recognized by library catalogers, although rarely by the library's users. Users do want to be able to find works using these entry points, although they may not think to search for them in the author index. There are also difficulties defining authors for the less traditional works, such as music or film. Who is the author of a film? Is it everyone whose name is listed in large type in the credits, the producer, the writer, the director? And in the case of a piece of classical music that is performed by a modern orchestra, who or what should be an author-like search point? Composer? Arranger? Conductor? Orchestra? And some systems do not have an author search, but instead a personal name search. This search includes all personal names in the bibliographic record, including those used as subjects. There is a certain logic to this in that a single search retrieves all items by and about a person.

All of these differences contribute to variable results when broadcasting the same search to multiple systems. And this has implications for the creation of a virtual union catalog.

4 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 and Search Accuracy

What our test showed was that the biggest problem in using a virtual union catalog is the inconsistency of results. If all of the systems participating in the union catalog had the same definitions for indexes, did the same normalization of index keys, and performed their searches in the same way, then it would be possible to obtain consistency. This is not the situation in many consortia. It is important, therefore, if you are considering the creation of a virtual union catalog, to study the retrieval capabilities of the library systems that will be included. If you are using Z39.50 to broadcast searches to these systems, you may be able to customize the searches that are sent to each library system to help ensure that the results that you retrieve from the systems are comparable.

This also means that changes to the local systems could affect the union catalog search, so change information must be shared among the library systems.

System Availability

When you create a virtual union catalog, you are dependent on the system availability of each of the systems in the union catalog. It is ideal to have agreement between the systems that they will be available certain days and hours. This catalog solution creates a great interdependency between the libraries that are participating. If a library is taking down its system for maintenance, it may be necessary to inform other libraries in the system that it will not be available.

Capacity Planning for Library Systems and Networking

The development of a virtual union catalog design has important implications for local system search capacity and network load. Each search is broadcast to all of the local library catalogs, with the potential that each catalog will then process as many searches as the cumulative total that the libraries

previously handled individually. Network capacity planning would be required to accommodate the increased bidirectional 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. Although Version 3.0 of the Z39.50 protocol includes a sort function, 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.

5 Conclusion

Do the results of this study mean that a virtual union catalog should not be considered as an option for your library and its partner institutions? Not at all. This study pointed out some of the criteria that must be considered when deciding whether to create a centralized database, as opposed to a virtual union catalog. These can be summarized as:

- The success of virtual union catalogs will increase among libraries with similar local systems and similar cataloging and indexing, and will decrease with differences in those aspects;
- A centralized union catalog may be more costly to create, but it can overcome some of the differences in record quality from different institutions and actually enhance retrieval of minimally cataloged items; and
- A virtual union catalog is highly dependent on the day-to-day functioning of local systems; a centralized catalog needs to receive records from local catalogs but otherwise functions independently.

Each library consortium must decide its goals for a union catalog and weigh this against its budget and technical capabilities. The important thing is to understand the system capabilities and to plan your services around what your system can actually deliver.